Vol. 35

ISBN 978-969-8858-27-8

PROCEEDINGS

18th International Conference on Statistical Sciences:

Using Statistics for Sustainable Development Goals & Human Development

February 18-20, 2021

ISOSS

PISTAR

Editors:

⇒Professor Dr. Muhammad Hanif

⇒Professor Dr. Shahjahan Khan

Venue:

Imperial College of Business Studies

Lahore (Pakistan)

Sponsors:

ISSN 1930-6792 JAPS

ACTIVEKEY

FUTURENOW TECHNOLOGIES

Pakistan Cargo Services

Travel Shop (Pvt) Ltd

Jointly organized by:

Islamic Countries Society of Statistical Sciences

Pak Institute of Statistical Training and Research

44-A, Civic Centre, Sabzazar Scheme, Lahore, Pakistan

Email: secretary@isoss.net

Published by: ISOSS, Lahore, Pakistan.

"All papers published in the

PROCEEDINGS

were accepted after formal peer review

by the experts in the relevant field.

Editors:

- **Professor Dr. Muhammad Hanif**
  National College of Business Administration & Economics, Pakistan.

- **Professor Dr. Shahjahan Khan**
  University of Southern Queensland, Australia.

CONTENTS

FOREWORD

I am very pleased and privileged to share my thoughts on the 18th International Statistics Conference organised by Islamic Countries Society of Statistical Sciences (ISOSS). This is another milestone in the history and achievement record of ISOSS and its highly dedicated professional leadership and team of volunteers.

Organising an international conference is not any easy task, and, unquestionably, it is more challenging during the devastating global pandemic. However, the highest level of commitment, dedication and spirit left behind by most respected Professor Munir Ahmed (may Allah be pleased with him) have once again inspired another generation of statisticians to take the challenges of organising yet another successful international conference.

The conference was a face to face professional gathering but many international participants, like myself, participated in the conference online because of travel restrictions. This added to the further complexities to the organisers as they had to manage both physical and virtual technologies and marry them together to make the event a success.

The conference attracted a large number of papers and participants. However, only 40 papers were accepted to be included in the Proceedings after peer review. This challenging and difficult job was efficiently managed by the Executive Secretary of ISOSS, Mr. Mohammad Iftikhar and his highly skilled team of professionals.

The organisation of regular annual international conferences and publishing Proceedings have made ISOSS one of the few leading statistical bodies in the world serving the international statistics community to spread the knowledge and benefit of statistics to the researchers, academics, policy-makers and common people.

I thank all the organisers, sponsors, participants, presenters and volunteers for their valuable contributions to make the conference a success, and publish this Proceedings so quickly and professionally. May Allah reward everyone of them for their good work.

Professor Shahjahan Khan PhD
President of ISOSS (2005-2011)
University of Southern Queensland
Australia

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 1-6

ISSUES OF STATISTICAL DEVELOPMENT IN PAKISTAN

Muhammad Rashid Salaria
Former DDG, Statistics Division, Government of Pakistan
Islamabad, Pakistan
Email: dr.rashidsalaria@gmail.com

ABSTRACT

Pakistan faces today the worst ever socio-economic crisis of its history. Economy is in a state of jam, governance is weak and flawed and political stability is non-existent. The result is that poverty is increasing, public unrest is of unprecedented scale and there is no immediate hope of any relief to the common man. The latest World Bank report shows Pakistan's debt liability as 94-95% of the GDP which is extremely alarming. The principal reason of this mess is the lack of proper planning and policy formulation for national development and social uplift in a befitting manner. Governments of all shades and ideologies in Pakistan have totally failed in putting in place a sound, scientific and fool proof mechanism to analyze national issues in depth and take timely decisions based on authentic data and statistical information covering all sectors of economy and social life. Although, Pakistan has made strides in establishing a statistical system over time but a lot more is still required to get desired results. The need is to develop an independent and powerful statistical setup with high professional competence and complete authority which may provide data in desired quality and quantity and build a full public confidence in national statistics.

KEYWORDS

Statistical development, national statistics strategy, restructuring of statistical system, statistical data

INTRODUCTION

The diversification of government functions, great challenges of governance and globalization of economic activities has made the need for accurate and reliable statistical information a corner stone for maximizing efficiency and cost effectiveness of the state functions. Governments and civil society need authentic statistics which may serve as a powerful tool to build the political will to initiate genuine socio-economic development activities, prepare short and long term plans and make effective result-oriented policies for the uplift of masses in the desired manner. Without quality statistical data in relevant areas, countries are ill placed to plan and oversee development activities properly. Flawed policies and badly taken decisions in the absence of reliable statistical data just waste scarce resources and adversely affect the poor and rich segments of society equally. As it is, quality statistics are required not only for proper planning, good policy formulation and evidence based decision making in every socio-economic sector, these are essential equally for monitoring progress and success of development projects in relevant areas.

Reliable and timely statistics can only be obtained when the task of data collection, compilation and dissemination is based on a well-planned and properly designed statistical system.

BRIEF HISTORY

At independence in 1947, Pakistan inherited a decentralized statistical system and statistical data was produced as a by-product of administrative functions of the government. The system consisted of statistical cells varying in capacity, professional strength and technical competence, working in different departments of the government. Realizing the shortcomings, the Government of Pakistan established a Central Statistical Office (CSO) in 1950, responsible for collection, compilation and analysis of statistical data to meet the national requirements. Thereafter, the statistical system of Pakistan remained continuously under periodic reviews and critical examination of national and foreign experts with the basic objective to enhance its efficiency and coverage. As a result of these efforts, the status of CSO was raised to a full-fledged Statistics Division in 1972, under the Minister for Finance, Planning and Development, on the recommendations of the World Bank Mission to Pakistan. In 1978, Population Census Organization (PCO) and Agricultural Census Organization (ACO) which were parts of the Ministry of Interior and Ministry of Food & Agriculture respectively, were brought under the umbrella of Statistics Division. As a result of further re-organization in 1981, the technical wing of Statistics Division (former CSO) was converted into Federal Bureau of Statistics as an attached department of Statistics Division. Planning, Evaluation and Research (PE&R) Cell was created in the Statistics Division for providing technical back stopping. Additionally, Pakistan Institute of Statistical Training and Research (PISTAR) was also set up under the Statistics Division in late 1980s to serve as a national centre for statistical training & research in the country.

For providing policy guidelines and for coordination of statistical activities suited to the national requirements, a National Statistical Council (NSC) was constituted under the provisions of General Statistics Act 1975. The Council was headed by the Minister for Finance, Planning and Development and Federal Secretaries of the concerned economic ministries, Chairman Punjab Provincial Planning & Development Board and Add. Chief Secretaries (Dev.) of other Provinces were its members. NSC was mandated to meet periodically to deliberate on statistical policy issues and make recommendations for statistical development in the country. The composition of NSC was further enhanced to include high profile technical experts and professionals, belonging to public and private sectors, in its fold.

This system worked till 1993 when as a restructuring of Government Ministries / Division, the Statistics Division was abolished and merged with the reorganized / enlarged Economic Affairs & Statistics Division. The Planning, Evaluation & Research (PE&R) Cell and administrative Unit of former Statistics Division were made the Statistics Wing of Economic Affairs & Statistics Division. However, the Government soon realized that the abolition of Statistics Division was adversely affecting the statistical work in the country. Thus, for quick decision making and policy formulation for rapid statistical development, the Statistics Division was restored in 1995 with enlarged functions.

RESTRUCTURING OF STATISTICAL SYSTEM

At the beginning of new century, the statistical system of Pakistan was a combination of both centralized and decentralized systems. At federal level, it consisted of Statistics Division, its three attached departments, namely Federal Bureau of Statistics (FBS), Population Census Organization (PCO) and Agricultural Census Organization (ACO) and a few Statistical Cells working in various government Ministries / Divisions / Organizations. At provincial level, Provincial Bureaus of Statistics were functioning for collection, compilation and dissemination of data on provincial subjects and for coordination among Statistical Cells operative in various provincial government departments / organizations.

Although this system had evolved over time with its parameters well defined, it suffered from severe problems of reconciliation and coordination at national level, adversely affecting the quality and reliability of data. Deficient institutional arrangement, particularly at the provincial level, for collection, compilation, processing and dissemination of data was a major hurdle in the way. Due to an ineffective interaction between data users and data producers, the data users were often unaware of reliability and authenticity of data. That was due to lack of their knowledge about the methodology and procedures employed for data collection and processing. Further, in the absence of a proper coordination between the data producers themselves, there were instances of different sources producing different data for the same statistical indicators which added to the confusion of data users. Data gaps at central & provincial level had made the supply line unable to meet the current and future data requirements for socio economic development planning and policy formulation at district and sub-divisional level. Lack of proper leadership, inadequate skill building, unattractive career advancement opportunities and mismatch of infrastructural facilities in the federal and provincial government statistical sets up were the other weak areas which were badly affecting the performance of the system.

To remove the above bottlenecks and to cater for future high data national and international demand, especially to achieve new millennium development goals, the Statistical System of Pakistan needed restructuring and overhauling on modern lines. To give the system a credibility and due respect, statistical activities needed to be made high profile activities in line with international practices. This situation was critically reviewed. In this regard, a high level policy seminar on national statistics strategy was also held by the federal government, attended by key senior policy makers, planners, researchers and data users of the Government. The purpose of the seminar was to deliberate on and explore a well-coordinated national strategic approach to statistical development for the whole of Pakistan. The seminar highlighted the significance of a proper statistical development in the context of policy formulation, planning and evidence based evaluation of development activities in the country. The seminar recommended, among other things, to restructure the national statistical system on modern lines with a focus on quality, transparency, timeliness and subject area coverage of data. Based on a detailed independent study and substantial research work on the restructuring plan, this author suggested to create a single central and powerful statistical organization in the country after merging together various statistical outfits working separately in different areas. Prior to this proposal, special meetings were held with the provincial Bureaus of Statistics to take them also on board. Two models for restructuring of statistical system of Pakistan, with distribution of work well defined, were developed and presented for consideration of the government:

MODEL – I

Statistical set-up at the federal and provincial level may be integrated into one centralized system and the existing statistical outfits may be brought under one umbrella. Under this system, Statistics Division (with its three attached departments), Provincial Bureaus of Statistics, Statistical Cells working in the Federal Ministries / Divisions and Provincial Government Departments may cease to work separately and be part and parcel of the Central System with the new name of Pakistan Bureau / National Office of Statistics. This system working under one authority will cater to the entire data requirements of the federal Government, Provincial Governments, individual Ministries / Divisions / Departments and the general public without any overlapping or duplication of work. This system with a network of statistical offices at district level will work uniformly under set priorities of statistical activities within a specified budget. No other agency may be allowed to undertake any statistical activity in the fields covered by the centralized system to avoid inconsistent and conflicting results. As the statistical offices at all tiers will be vertically and horizontally integrated into one system, there will be an effective interaction and coordination between them and they will have an equal share of career advancement and skill building for their technical staff. Further, the system will have full authority and independence for producing reliable national figures without the problem of data integrity, comparability and accessibility. The statistical system so developed will have an inbuilt mode of effective coordination between the data producers and users. The system will work under the overall supervision of Minister of Finance. However, to give the system a high profile status, Chief Statistician (Head of proposed system) may directly report to the Prime Minister of the country. Constitutional amendments to this effect may be made accordingly. It should be noted that in a number of countries having federal structure of the government, collection of statistics has been designated a federal function irrespective of the constitutional division of subject matter fields.

MODEL – II

The system may be decentralized at the federal and provincial level. At the federal level, Statistics Division, its attached departments and all Statistical Cells working in various Ministries / Divisions may be merged together and renamed as Pakistan Bureau / National Office of Statistics. At the provincial level, the Provincial Bureaus / Offices of Statistics may work as agents of Pakistan Bureau / National Office of Statistics remaining a part and parcel of the provincial governments. In this regard, special institutional mechanism may be put in place, either through financial or through administrative arrangements. The Pakistan Bureau / National Office may undertake all national statistical surveys / censuses with the active participation of Provincial Bureaus who would be custodian of data pertaining to their respective provinces. The Provincial Bureaus may conduct individual statistical enquiries / surveys as and when required by the respective provinces or any agency / department the demand for which may be channeled to the Pakistan Bureau / National Office. The staff of Statistical Cells working in Provincial Departments would be merged in the respective Provincial Bureau / Office and placed on its pay roll. The Provincial Government or the Pakistan Bureau / National Office of Statistics, as the case may be, will be responsible for allocating proper funds to the Provincial Bureaus / Offices and taking care of the training and career advancement of their technical staff. As in the first model, no other agency may be authorized to undertake any statistical activity on their own but may be sending their demands for data to the Pakistan Bureau / National Office of Statistics. Under this model, statistical offices at district level may be operated by the Provincial Bureaus of Statistics.

Although both the above two models could serve as role models for restructuring of statistical system of Pakistan, the first model was considered more appropriate for the country which could be implemented in phases. In the first phase, the new system might be introduced at the federal level without the inclusion in it of the Statistical Cells working in various Federal Ministries / Divisions / Organizations. In the second phase, the job of these individual cells could be taken over and integrated with the mainstream system. In the final phase, the system might be extended to the provincial and district levels. This phasing was necessary in view of inadequate financial and manpower resources available then with the government.

ESTABLISHMENT OF PAKISTAN BUREAU OF STATISTICS

It had been noted that the Federal Bureau of Statistics (FBS), working then as an attached department of the Statistics Division, had the biggest field network for conducting surveys and undertaking statistical activities through its 16 Regional and 19 Field Offices, spread all over the country. These offices were adequately staffed with reasonable infrastructural facilities which handled the work of FBS exclusively with the occasional technical and manpower support to sister organizations (i.e. Population and Agricultural Census Organizations) at the federal level. The Population Census Organization had its offices located at Provincial Headquarters alongwith a few Field / Regional Offices located in major cities. The Agricultural Census Organization had no Field Office and was restricted only to its Headquarters at Lahore. It was estimated that with the merger of three attached departments and Technical Wing of Statistics division and pooling of their existing manpower and infrastructural resources, these 35 Offices of FBS could be further upgraded and adequately equipped with internet facilities with proper sets up at Provincial Headquarters, without any burden on the national exchequer. These offices after up gradation could easily handle the statistical activities of all the existing attached departments of Statistics Division in respective areas with greater cohesion and coordination. The set up proposed as first phase of the restructuring plan could thus be implemented immediately and without further loss of time. Though belatedly, the government did implement it with some modifications in the original plan and launched Pakistan Bureau of Statistics (PBS) as the central agency of Statistics in Pakistan. General Statistics Act 2011 provides legal support to all statistical activities in the country.

CONCLUSION

The Government of Pakistan has established Pakistan Bureau of Statistics (PBS) with fanfare as the prime statistical Agency of Pakistan, responsible for collection, processing, compilation and dissemination of national data relating to various sectors of economy and social life. However, the independence, professional authority and technical command as desired are missing in the organization. This lacking is manifesting itself in many ways. Public distrust of official data is expressed openly and independent experts and analysts put quite frequently question mark on the reliability of official statistics. The recent and classic case is that of 2017 population census whose results have been challenged by and large and government failure to release official figures so far reflects badly on professional stand and integrity of the organization. The delay in release of census data is being taken as normal but one fails to imagine any other issue more serious than this technical debacle in a highly specialized field. There is hardly any precedence in the world of any official statistical agency facing difficulty in defending its own figures and convincing stakeholders on the authenticity and reliability of data. Doubts are also cast on the merit of appointments on key positions in the organization, specially the position of Chief Statistician. These positions require highly qualified, visionary, competent and experienced professionals to stand firmly with principles as fact finders of the nation and not compromise on figures compiled and released by the single authorized agency. There is no doubt that the embarrassment relating to 2017 census could have been avoided had the case been handled with full professional acumen at all stages and presented logically at relevant forums. As for the Governing Council of PBS, it can better perform and deliver if its composition is made broad-based by taking subject area experts and senior bureaucrats of the concerned Ministries/ Divisions / Organizations in its fold to assist the Government to adopt pragmatic approach for the development of statistical system suited to national aspirations. The Council ought to be a fully competent and technically sound forum to determine major statistical policies and objectives, pinpoint data gaps and recommend a set program of priorities to fill them on a short and long term basis. Finally, it is the need of the time to implement the remaining two phases of restructuring of Pakistan Bureau of Statistics as provided in the original restructuring plan of the organization.

REFERENCES

1. Government of Pakistan (1979). *Retrospect, Perspective and Prospect*, Statistics Division, Karachi, Pakistan.
2. United Nations (1993). *East and South Asian Workshop on Strategies for Accelerating the Improvement of Civil Registration and Vital Statistics Systems (Report of the Workshop)*, United Nations Statistical Division, New York.
3. United Nations Statistical Commission (2000). *Report on the Thirty-First Session, (February 29 - March 03, 2000)*, New York.
4. Government of Pakistan (2021). *Pakistan Bureau of Statistics, Publications*, Islamabad, Pakistan.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 7-20

DATA REVOLUTION AND STATISTICS: RESHAPING SOCIETY AND SCIENCE

Shahjahan Khan[1] and **Muhammed Ashraf Memon**[1,2]
[1] School of Sciences, University of Southern Queensland, Toowoomba Australia. Email: Shahjahan.Khan@usq.edu.au
[1&2] South East Queensland Surgery (SEQS) and Sunnybank Obesity Centre Sunnybank, Queensland, Australia. Email: mmemon@yahoo.com

ABSTRACT

Beyond any doubts, the whole world is moving too fast to embrace the artificial intelligence and machine learning technologies to benefit from big data. Consequently, the use of unprecedented volume and intensity of data is becoming more and more an integral part of everyday life of modern science and citizens. The reality is that many of the next generation digital technologies will be data driven, internet dependent and satellite guided. Starting from driver-less vehicles to store-less shopping platforms and delivery of personalised services will be digitised based on data guided evidence. In fact, data revolution is already here, and we are increasingly being exposed to various technologies that are dependent on results from analysis and prediction of data. The role of statistics and computing algorithms in the process are crucial and it will continue to grow. The policymakers in government offices, health services, technology centres and business establishments are moving towards evidence-based decision-making which is predominantly guided by data synthesis and analytics. This paper covers various aspects of data including its dimensions and diversities, sources, owners and methods of analysis, and statistical and computational essence leading to the new paradigm of science, data.

1. INTRODUCTION

Data revolution is already here, and it will continue to grow more and more at an accelerated rate in the days and months and years to come. In the contemporary world, there is no shortage of data. Rather we have a plethora of data almost everywhere. The Covid-19 has added new dimension to already exponentially growing generation of data. Data are useful and potentially beneficial to mankind if the invaluable 'jewel' in the mess of data can be uncovered by using appropriate methods. Increasingly evolving statistical methods are the main vehicle to compile, explore and analyse raw and unformatted data and interpret the results leading to evidence-based decision-making.

The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). The IoT refers to the billions of physical devices around the world that are now connected to the internet, all collecting and sharing data. Connecting all these different objects and adding sensors to them adds a level of digital intelligence to devices that would be otherwise dumb, enabling them to communicate real-time data without involving a human being. It is making the fabric of the world around us smarter and more responsive, merging the digital and physical universes.

In not so far distant past, getting sufficient data on various studies was a real challenge. Also, most of the datasets were so small and in lower dimension, people would analyse it by hand or using manual calculators. The advent of mainframe computer was a big change and made analysis of data must faster and easier. Introduction of high-power computers, laptops and tablets has changed the world of computing forever. Adding that to the access to the fast and high band internet and cloud technology have revolutionised the whole world. Over the last two years alone 90 percent of the data in the world was generated.

Computer and internet revolution have enhanced data revolution by providing means to deal with high dimensional large data, and its easy storage and fast transmission and remote access. We started from very slow manual calculating machine, and then came the scientific hand calculator. That was of no comparison to the mainframe computer of 1970's. The introduction of personal computer (PC) in 1980's, laptop in the later part of twentieth century and later tablet and super computers have revolutionised the computing forever. Couple with the access to the high-power internet superhighway the computing technology has changed the way we think and work. As a more recent development, the introduction of smart phone and other communication technologies have created very interconnected world.

Figure 1: Evolution of Computing Technologies

To emphasise the benefit and importance of big data, in a press release on 29 March 2012 The White House described the goals of the Big Data Initiative of President Barak Hussain Obama, "to advance the state-of-the-art core technologies needed to collect, store, preserve, manage, analyse, and share huge quantities of data; harness these technologies to accelerate the pace of discovery in science and engineering; to strengthen our national security, and transform teaching and learning; and to expand the work force to needed to develop and use Big Data technologies."

On 3 June 2015, the UN General Assembly adopted a Resolution (1) designating 20 October 2015 as the 2nd World Statistics Day, under the general theme “Better data. Better lives”. The then UN Secretary-General pointed out, "statistics are critical for evidence-based decision making across all cultural and historical backgrounds of countries and irrespective of their level of development." (2) highlighted the widespread application of statistics.

Martinez (3) wrote, “Data can now be stored in whatever corner of the world best serves the interests of cost and convenience, and it can be retrieved at the literal speed of light or secreted away in the dark. [The] digital version of mare liberum (‘free seas,’ the guiding legal notion coined by 17th-century Dutch jurist Hugo Grotius) is only now being subject to real legal scrutiny, and it’s a once-in-an-era opportunity to define how we regulate the global trade of information and intellectual property. Unfortunately, bureaucrats are busy writing data localization laws that stem from the age of ink-stained vellum paper rather than our current Dropbox reality. It’s the trade-off between free movement and national sovereignty, as always, that’s tricky…Can a global cloud system peacefully coexist with national sovereignty?”

The Internet World Stats (4) website provides a wide range of internet usage data including big picture. Here is a table on the usage of internet globally.

Table 1
Global Usage of Internet

WORLD INTERNET USAGE AND POPULATION STATISTICS 2020 Year-Q4 Estimates						
World Regions	**Population (2021 Est.)**	**Population % of World**	**Internet Users 31 Dec 2020**	**Penetration Rate (% Pop.)**	**Growth 2000-2020**	**Internet World %**
Africa	**1,357,198,684**	17.3 %	**633,8856,924**	46.7 %	13,941 %	12.8 %
Asia	**4,309,503,789**	55.0 %	**2,563,503,922**	59.5 %	2,143 %	51.8 %
Europe	**835,700,837**	10.7 %	**727,848,547**	87.1 %	593 %	14.7 %
Latin America/ Caribbean	**658,382,700**	8.4 %	**477,824,732**	72.6 %	2,545 %	9.7 %
Middle East	**263,933,993**	3.4 %	**184,856,813**	70.0 %	5,528 %	3.7 %
North America	**370,146,066**	4.7 %	**332,910,868**	89.9 %	208 %	6.7 %
Oceania / Australia	**43,138,089**	0.6 %	**29,066,532**	67.4 %	281 %	0.6 %
WORLD TOTAL	**7,838,004,158**	**100.0 %**	**4,949,868,338**	**63.2 %**	**1,271 %**	**100.0 %**
Notes: (1) Internet Usage and World Population Statistics estimates are for December 31, 2020. (2) CLICK on each world region name for detailed regional usage information. (3) Demographic (Population) numbers are based on data from the United Nations Population Division. (4) Internet usage information comes from data published by Nielsen Online, by the International Telecommunications Union, by GfK, by local ICT Regulators and other reliable sources. (5) For definitions, navigation help and disclaimers, please refer to the Website Surfing Guide. (6) The information from this website may be cited, giving the due credit and placing a link back to www.internetworldstats.com. Copyright © 2021, Miniwatts Marketing Group. All rights reserved worldwide.						

Here are some incredible stats for the volume of communication we send out **every minute**: There are 16 million text messages; 990,000 Tinder swipes; 156 million emails sent; 15,000 GIFs are sent via Facebook messenger; 154,200 calls on Skype; 103,447,520 spam emails; and it was predicted that there will be 2.9 billion email users by 2019.

Vuleta (2020) (5) writes, "Data is the new oil" is perhaps one of the most popular catchphrases highlighting the importance of data. The metaphor is, admittedly, a bit inaccurate, but it paints a picture of our collective online footprint concerning the global economy and our digital lifestyle. But how much data is created every day? Frankly, there is no definitive answer to this basic question. The Google, Amazon and Facebook of the world could not keep count even if they wanted to.

2. THE DATA REVOLUTION

It is difficult to comprehend how much data are being generated every day through the use of internet, social media, commercial transactions, digital images, records of health services, government offices, astronomical tracking, emails, security devices, satellite activities, research laboratories, communications, transport, bank cards, weather indices etc. and the list goes on and on. Imagine how much data is produced and processed by Facebook or Instagram or Twitter or any other social media. What about Google, YouTube, LinkedIn, Research Gate, etc.?

At the beginning of 2020, the digital universe was estimated to consist of 44 zettabytes of data, and it is predicted that by 2025, approximately 463 exabytes data would be created every 24 hours worldwide. As of June 2019, there were more than 4.5 billion people online; and 80% of digital content is unavailable in nine out of every ten languages; and Google processed 3.7 million queries, Facebook saw 1 million logins, and YouTube recorded 4.5 million videos viewed every 60 seconds.

Netflix's content volume in 2019 outnumbered that of the US TV industry in 2005. By 2025, there would be 75 billion Internet-of-Things (IoT) devices in the world; and by 2030, nine out of every ten people aged six and above would be digitally active.

In an online article Marr (2018) (6) noted the following facts: More than 3.7 billion humans use the internet (that's a growth rate of 7.5 percent over 2016). On average, Google now processes more than 40,000 searches every **second** (3.5 billion searches per day)! Worldwide there are 5 billion searches a day.

In the Social Media the data generated every **minute** include Snapchat users share 527,760 photos; more than 120 professionals join LinkedIn; 4,146,600 YouTube videos; 456,000 tweets; and Instagram users post 46,740 photos.

More than a quarter of the world's 7 billion humans over 2 billion are active on Facebook. On this platform alone 1.5 billion people are active daily; five new Facebook profiles created every second; more than 300 million photos get uploaded per day; every minute there are 510,000 comments posted and 293,000 statuses updated. For the Instagram (also owned by Facebook) there are 600 million accounts; 400 million are active every day; each day 95 million photos and videos are shared, and 100 million people use the Instagram "stories" feature daily.

For communications in every minute people send 16 million text messages; 990,000 Tinder swipes; 156 million emails; there are over 2.9 billion email users; 103,447,520 spam emails and 154,200 calls on Skype. No data on zoom meetings during the pandemic.

Due to Covid-19 pandemic, usage of the online meeting platform Zoom has gone up since March 2020. Iqbal (2020) (7) writes, "Zoom truly entered the public consciousness during the coronavirus pandemic of 2020. It was to Zoom that users across the world turned to stay in touch during the lockdown effected to stop the spread of the virus."

Founder of Zoom, Eric Yuan stated in a blog post that over the course of that month (March 2020), Zoom was seeing 200 million daily meeting participants (not DAU). The following month, this figure had risen to 300 million. This compares to 10 million in December 2019. The UK cabinet and 90,000 schools in 20 countries were among new users of the app. In 2020 financial year Zoom revenue was $622.7 million compared to $121.5 million in 2018.

With the growth of big data there have been challenges as well. The security of data and its protection from hackers are some of the main challenges. In recent years data of various sensitive government departments and even prominent universities have been compromised by foreign cyber attackers. Privacy of personal data is increasingly under threat due to organised cyber and internet crimes.

Facts about Big Data

The actual size of the big data, and how fast its volume is increasing is a real surprise to many people. Even with the mainframe computer the volume of data was measured in kilobytes to megabytes. It is only after the introduction of PC and laptop the volume of was counted in gigabytes. But now it is much bigger, and the phenomenal growth of data is so much and so fast that total data volume in the world doubled only in two years.

Table 2
Measure of Volume Ever Growing Data

Data Size	Number of zeros	Short Scale
Kilobytes	10^3	Thousand
Megabytes	10^6	Million
Gigabytes	10^9	Billion
Terabytes	10^12	Trillion
Petabytes	10^15	Quadrillion
Exabytes	10^18	Quintillion
Zettabytes	10^21	Sextillion
Yottabytes	10^24	Septillion

There were approximately 44 zettabytes of data in the world in 2020. Given current rate of data creation, there will likely be 175 zettabytes data added everyday by 2025.

To get real feeling to Big Data the following facts from (8) in the High Scalability website may be helpful:

- Bytes (8 Bits) :100 bytes: A telegram or A punched card
- Kilobyte (10^3): 100 Kilobytes: A low-resolution photograph
- Megabyte (10^6): 100 Megabytes: 1 meter of shelved books or A two-volume encyclopedic book
- Gigabyte (10^9): 100 Gigabytes: A floor of academic journals or A large ID-1 digital tape
- Terabyte (10^12): 10 Terabytes: The printed collection of the US Library of Congress
- Petabyte (10^18): 2 Petabytes: All US academic research libraries
- Exabyte (10^21): 5 Exabytes: All words ever spoken by human beings.
- Zettabyte (10^24): According to International Data Corporation, the total global data to grow to 2.7 zettabytes during 2012. This is 48% up from 2011.
- Human brain's ability to store memories is estimated at about **2.5 petabytes** of binary data.

Who Owns Big Data?

There are lots of data available on the internet but only a handful of large techno companies own balk of these Big Data. The Software Testing Help (9) lists top 13 Best Big Data Companies of 2021. These are Xplenty, IBM, HP, Enterprise, Teradata, Oracle, SAP, EMC, Amazon, Microsoft, Google, VMware, Splunk, Alteryx, and

Cogito. Among these giant owners of Big Data, only IBM provides the following six Big Data Solutions:

- Hadoop System: It is a storage platform that stores structured and unstructured data. It is designed to process a large volume of data to gain business insights.
- Stream Computing: Stream Computing enables organizations to perform in-motion analytics including the Internet of Things, real-time data processing, and analytics
- Federated discovery and Navigation: Federated discovery and navigation software help organizations to analyse and access information across the enterprise. IBM provides below listed Big Data products which will help to capture, analyse, and manage any structured and unstructured data.
- IBM® BigInsights™ for Apache™ Hadoop®: It enables organizations to analyse a huge volume of data quickly and in a simple manner.
- IBM BigInsights on Cloud: It provides Hadoop as a service through the IBM SoftLayer cloud infrastructure.
- IBM Streams: For critical Internet of Things applications, it helps organizations to capture and analyse data in motion.

The above list does not include most of the well-known and popular social media companies. For example, Facebook has 750,000,000 Unique Monthly Visitors (UMV); Twitter – 250,000,000; LinkedIn – 110,000,000; MySpace – 70,500,000; and Google Plus+ – 65,000,000 UMV. Other companies owning Big Data include Flickr, Youtube, iTunes, Pinerest, Commercial loyalty cards, Alibaba, EBay, MarketPlace, Credit card companies, Airline frequent flier cards, health care system, etc.

3. PROBLEMS WITH BIG DATA

Big Data is often defined as a data set that is huge, multidimensional, and complex so that traditional data processing methods, techniques and applications are inadequate to deal with them. There are challenges to managing such a huge volume of data such as capture, store, data analysis, data transfer, data sharing, etc. The common features of Big Data are high "Volume", "Velocity" and "Variety" which are popularly known as 3V.

Before analysing any Big Data there are several big challenges to clean and format the data so that data analytic methods could be applied. First challenge is to understand too many variables (sometime in millions) on the high dimension.

The cleaning of data may include checking and dealing with missing values, gaps, inconsistencies in format, use of different units, geographical and demographic consistencies, appropriateness of any transformation of data etc. Checking assumptions is also not so easy for Big Data.

Essentials to Analyse Big Data

First of all, the Big Data can't be stored and analysed in any ordinary computer because of its size. People are now using internet and cloud technology to store big data. The big data always requires lots of cleaning and formatting before it could be analysed.

Statistical methods are used in analysing many big data. Often interest in big data is to find correlation, classification, clustering, and structure. Artificial Intelligence, Data Mining and Machine Learning algorithms are often used to analyse big data.

Artificial Intelligence (AI) is a constellation of technologies—from machine learning to natural language processing—that allows machines to sense, comprehend, act, and learn. Artificial intelligence will transform the relationship between people and technology, charging our creativity and skills.

The future of AI promises a new era of disruption and productivity, where human ingenuity is enhanced by speed and precision.

Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Data mining is an interdisciplinary subfield of computer science and statistics. Data mining is the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes by using a broad range of computer science and statistical techniques.

In spite of many benefits of Big Data there are concerns about the abuse and misuse of the technologies including **relying on the** algorithms for identifying crime suspects (not necessarily accurately), deciding who gets a credit card, a home loan, and a job. Or even decide on who qualifies for health insurance.

Another consequence of in the post-covid, the “new normal” will be the contactless payments, which become more frequent in the early months of the pandemic. The world is moving towards the cashless transaction era.

Due to the hacking and increase in the cybercrime by organised gangs data security and privacy is becoming a serious issue. As an extra precaution, people are becoming used to two-factor (2FA) authentication, including the use of biometrics and messing passcode via mobile phone message.

4. STATISTICS AND DATA SCIENCE

Statistics is a scientific discipline that uses results from analysis of data to make decision in the face of uncertainty and with minimum risk. It deals with the production, collection, organization, and analysis of data, and interpretation and presentation of results. Statistical methods are used in almost all spheres of life and disciplines, especially in research.

Statistics has made tremendous contribution to the humanity by providing data collection and production tools such as population census, sample survey based on random sampling method, and conducting controlled experiments to determine cause and effect relationship, and randomised clinical Trial (RCT). It has changed the way industry used to control product quality by introducing quality control chart, total quality management (TQM) and sequential sampling method. Statistical models are used for analysis of variance, multiple regression analysis, time series analysis as well and forecasting and prediction. Statistical inference allows to estimate unknown parameters, perform test of hypotheses.

Why Need Statistics?

Three realities of life make statistics inevitable to understand and make them beneficial. We live in a world full of errors, uncertainties, and variables. Only statistical methods could make use of error, uncertainty, and variable to reveal otherwise unavailable useful information from data. *Errors* are everywhere, measurement error, machine error, human error, judgment error etc. But statistical methods are capable of including errors in the model to benefit from it rather than running away from it.

In this world *uncertainty* in everything is uncertain and often unpredictable. We know that weather is uncertain and so are sales quantity, number accidents, expenditure, students’ marks etc. However, statistics uses probability measures to assess uncertainty and use it for modelling data. So, uncertainty is not a problem in statistics, rather it is an ingredient of statistical success stories.

The third feature of nature that makes statistics essential is *variability*. We live in the world of variables. Human height is a variable, and so are income, age, eye colour, number of children, price of houses, length of life and so on. Statistical methods are capable of analysing variables and find their characteristics to make them beneficial to humanity. Variability is not any problem in statistics rather it is an important ingredient in statistical arena. Any data we can think of is collection of values on specific variables. If there was not variability, there would not be any data. Statistical techniques, such as census and sample survey, are used to generate/produce and/or collect data. And then other statistical methods, such as estimation, test and modelling are used to analyse the data to make prediction and/or inference.

Emergence of Data Science

The limitation of conventional statistics to manage and analyse big data has inspired data analysts to venture into data science. Data science is an inter-disciplinary field that uses statistical methods, algorithms, and computing techniques to extract information from structural and unstructured data on the high dimension. Data science is related to data mining, machine learning and big data.

Statisticians must take the challenge to include skills of analysing big data within the curricula of statistics. What is needed is to include techniques required to deal with the analyses of big data along the traditional statistics courses. Since there is a high demand for graduates of data science, if statisticians do not train their students to handle big data someone else will do it. Wegman (2000) (10) suggested that statisticians ought to be data centric, i.e. develop methods which are motivated by the data at hand whether it fits standard models or not. This is a very important matter for the long-term sustainability and usefulness of statistics.

Data science not only require conventional statistical methods, it also needs skills such as statistical signal processing, pattern recognition, data mining, machine learning, bioinformatics, meta-analysis etc. Khan (2020) (11) discussed various statistical models to meta-analyse data from heterogeneous primary studies including the inverse variance heterogeneity (IVhet) model.

Table 3
Comparison of Data Science and Statistics

Characteristics	**Data Sciences**	**Traditional Statistics**
Data Size	Big data sets – terabytes and more	Small to moderate sample sizes
Data Type	Non-homogeneous Data	IID – Stationary Data
Data Dimension	High dimensional	Low dimensional
Computation	Computationally intensive	Manual computation
Tractability	Numerically tractable	Mathematically tractable
Assumptions	Weak or no assumptions	Strong assumptions
Questions	Imprecise questions	Well focused questions
Algorithm	Iterative algorithms	Closed form algorithms
Sample	No sampling	Random samples
Interest	Explore and analytics	Inference
Method	Robustness	Optimality

Unlike the conventional data analyses which is easily managed by personal computers or laptops, the storage and analysis of big data requires cloud technology. Sahinoglu and Cueva-Parra (2010) (12) discussed about cloud computing (Grid or utility computing, computing on-demand) which was the talk of the computing circles at the end of 1990s. Now it has become a much needed data storage and computing technology.

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service typically over the Internet. The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software & computation.

5. DATA RESHAPING SOCIETY

When anyone searches on the internet to purchase an item or find a best airfare, the person will notice that in you next visit to internet you have lots of online advertisement of the items you were searching. The advertising companies use computer algorithm based on the data in your computer to locate you and identify your needs. This had now become an aggressive marketing tool for businesses.

Technology based on internet, satellite and data is reshaping the society. The world is moving to cashless payment option. Currency is very rarely used in purchases in developed countries. Electric cars are in action and more are on its way to replace conventional fossil fuel-based vehicles. Driverless are is already in use and would increase rapidly. Particularly transport and delivery of goods will be done without human involvement. Robotic technology is replacing human jobs in many industries. Amazon has invested billions of dollars to research and develop drones for delivering goods door to door using drone.

Big Data has also been reshaping science. Data science and data analytics are being used by many scientific investigations. Many scientists are using results from data analyses to make evidence-based decisions. Meta-analysis is a powerful statistical tool to combine data from many independent primary studies to pool the summary statistics to estimate common effect size.

Many people now prefer to order online and get no-contact home delivery. This is likely to continue even in the upcoming post-covid new normal. DoorDash, Postmates, and Instacart all offer drop-off delivery options, reportedly borne from customer desires to minimize physical contact. Grubhub and Uber Eats also grew their contactless delivery options and will continue to do so in 2021.

China-based delivery apps like Meituan, which was the first company in China to implement contactless delivery in Wuhan, began using autonomous vehicles to help fulfil grocery orders to customers

In 2020, whether you were a data scientist or not, we all got a glimpse of this growing data curve as scientific researchers, pharmaceutical companies, governments, and healthcare institutes turned every resource toward developing vaccines, novel treatments, and other means to help the world stay healthy during the pandemic. All these efforts required generating and processing vast amounts of data. Whether in healthcare or other applications, the only realistic way to handle all the information we are seeing is to use ingestion and aggregation tools, married to Machine Learning (ML) models that can help make sense of it.

In Sweden, a country rich with technological advancement, thousands have had microchips inserted into their hands. The chips are designed to speed up users' daily routines and make their lives more convenient — accessing their homes, offices and gyms is as easy as swiping their hands against digital readers. They also can be used to store emergency contact details, social media profiles or e-tickets for events and rail journeys within Sweden.

6. DATA – A NEW PARADIGM OF SCIENCE

A paradigm is a distinct set of concepts or thought patterns or way of looking at something. The word paradigm comes up a lot in the academic, scientific, and business worlds. When there is a change in paradigm, there will be a change in how to think about something.

In science and philosophy, a paradigm is a distinct set of concepts or thought patterns, including theories, research methods, postulates, and standards for what constitutes legitimate contributions to a field. A scientific paradigm is a framework containing all the commonly accepted views about a subject, conventions about what direction research should take and how it should be performed.

Traditionally, science has been built on a two-stage paradigm of **Theory** and **Experiment**. Theory guides experiment, experiment suggests refinement of theory. With the advent of powerful personal computers in the early 1980s a third stage was introduced, the **Computation**. Some experiments were too expensive, too dangerous, or not feasible. This made computational experiments as a third component of the paradigm of science: Theory, experiment, computation.

Now we have the Google Experience of acquiring massive data in some sense supplants theory just as computing to some extent has supplanted experiment. As an example: Automatic translation needed the theory of natural languages with morphology, semantics, syntax, and lexicon in two languages. But, if you have enough data in the two languages, as Google does, simply matching multigrams in the two languages not only takes place of theory, but also neatly takes care of idiomatic expressions. Thus **data** has become the fourth paradigm of science after computing.

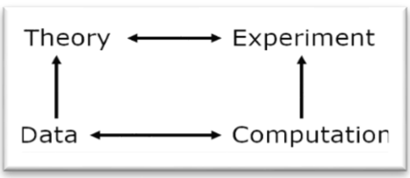


Figure 2: Data - The Fourth Paradigm of Science

Hey et al. (2009) (13) provides a collection of essays expanding the visionary ideas of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized. On the back cover of the book Bill Gates writes, "The impact of Jim Gray's thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."

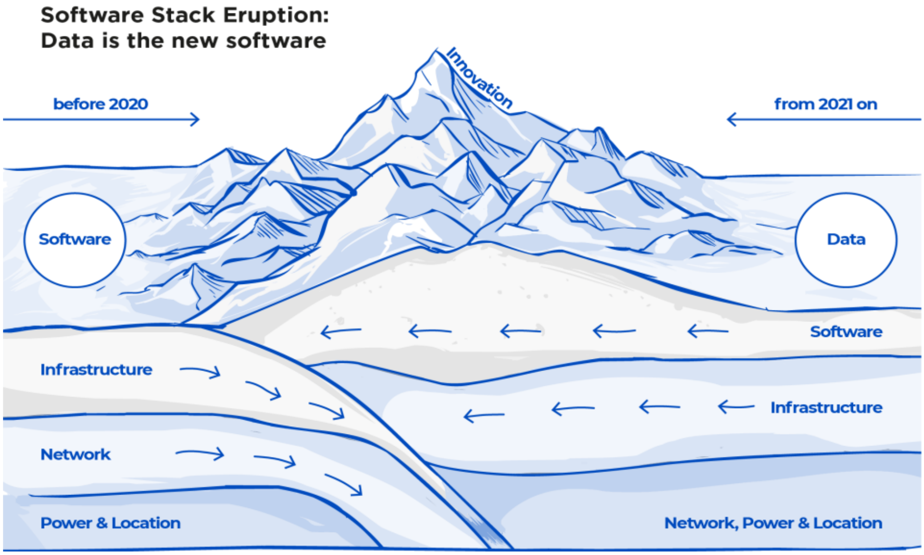


Figure 4: Software Stack Eruption (Source Cloudflight 2020)

7. POWER AND ABUSE OF STATISTICS

Statistics is being increasingly used for decision-making in the face of uncertainty. Statistical methods reveal the hidden 'jewel' in the data and make it beneficial for mankind. At the age of data revolution, statistics is being used in almost all disciplines and in all spheres of life. Statistics is a very powerful tool to create and mobilise public opinion and provide evidence in support of or against any claims or demands. Because of its inherent persuasive power many people in power are free to abuse statistics for personal benefit.

In November 2015 (before running for US Presidency) Mr. Donald Trump tweeted "Whites killed by blacks – 81%", citing "Crime Statistics Bureau – San Francisco". The US fact-checking site *Politifact* says – "Bureau" did not exist, and the true figure is around 15%. When confronted the press, Mr. Trump shrugged and said, "Am I going to check every statistic?" What an outrageous abuse of statistics? There are many more examples of this kind of blunder by other powerful people. Some of the common abuse and misuse of statistics are found in (14).

Then there are unsubstantiated and self-serving claims by peoples who want to avoid transparency and accountability by blaming statistics to be in fault rather than the people who manufacture falsehood. They keep repeating the baseless false statements like the old saying, "There are three kinds of lies: lies, damned lies and statistics." Although attributed to Mark Twain, this is not from him. Research shows that a 1895 article by Leonard H. Courtney is available at http://www.york.ac.uk/depts/maths/histstat/lies.htm suggests to be British politician Sir Charles Dilke (The Bristol Mercury, 19 Oct 1891) stated, "False statements might be arranged according to their degree under three heads, fibs, lies, and statistics." However, the discipline that is well known as the language of science and is now used in all kinds of decision making has nothing to do with being close to fibs or lies. Interestingly it is the politicians who often conveniently use and abuse statistics. Hence one may rephrase the above statement qualifying `False statements' to be fibs, lies, and *politics*. The fact of the matter is 'figures don't lie, but liars can figure' or 'figures fool, when fools figure'.

8. CONCLUSIONS

Data revolution has and will continue to be changing the society and sciences as more and more people engage in data driven activities willingly or unwillingly. It is obvious that machines such as robots empowered by artificial intelligence and machine learning algorithms will make further headway to our everyday activities and needs. Data has already secured its position as a paradigm of science, but it will continue to strengthen and reassure the same over time. Statisticians require to emphasize on computational skills and learn/teach how to use big data related technologies to secure their profession and discipline. Objective and correct statistics should be produced and promoted, and abuse of data and statistics must be protected from corrupt people, especially from those who are in the position of power.

REFERENCES

1. United Nation (2015). *Resolution adopted by the General Assembly on 3 June 2015*. Online available at http://undocs.org/a/res/69/282.
2. Khan, S. (2016). Statistics on spotlight: reflections on World Statistics Day. *Journal of Islamic Countries Society of Statistical Sciences*, 2(2), 10.
3. Martinez, A.G. (2018). *The End of Data Without Borders*. Online available at https://www.wired.com/story/overseas-data-regulation/
4. Group MM. *The Internet World Stats 2020*. Online available at https://www.internetworldstats.com/stats.htm
5. Vuleta, B. (2020). How Much Data Is Created Every Day? *Seed Scientific*, 30.
6. Marr, B. (2018). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Online available at https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=43d8380c60ba.
7. Iqbal, M. (2021). *Zoom Revenue and Usage Statistics 2020*. Online available at https://www.businessofapps.com/data/zoom-statistics/
8. Bunn, J. (2012). *How Big Is A Petabyte, Exabyte, Zettabyte, Or A Yottabyte*? Online available at http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html
9. Software Testing Help (2021). *Top 13 Best Big Data Companies of 2021*. Online available at https://www.softwaretestinghelp.com/big-data-companies/.

10. Wegman, E.J. (2000). On the eve of the 21st century: statistical science at a crossroads. *Computational Statistics & Data Analysis*, 32(3-4), 239-243.
11. Khan, S. (2020). *Meta-Analysis Methods for Health and Experimental Studies*. Singapore: Springer Nature.
12. Sahinoglu, M. and Cueva-Parra, L. (2011). CLOUD computing. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(1), 47-68.
13. Hey, A.J., Tansley, S. and Tolle, K.M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). A. J. Hey (Ed.). Redmond, WA: Microsoft research.
14. Khan, S. and Memon, M.A. (2017). Statistics for evidence-based decisions-use, misuse and abuse. In *Proceedings of the Third International Conference on Science, Engineering and Environment* (SEE-2017) (pp. 22-30). Geomate International Society.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 21-29

WAVELET INTERPRETATION OF MAXIMUM USEABLE FREQUENCY FOR SOLAR - TERRESTRIAL REGION OF PAKISTAN

Muhammad Ayub Khan Yousuf Zai, Bulbul Jan,
Humair Jahangir and **Ali Abbas Naqvi**
Department of Applied Physics and Institute of Space Science and Technology
University of Karachi, Karachi, Pakistan.
Email: ayubzai@yahoo.com

ABSTRACT

In this communication, we have described that the propagation of electromagnetic waves in the ionosphere is considered as complex by the fact that this medium is inhomogeneous, anisotropic, turbulent, non-stationary and multipath. One of the terms used in the Sky wave propagation is Maximum useable frequency (MUF) that refers to the highest possible frequency for high frequency (HF) communications that can be used to transmit over a particular path under given ionospheric conditions that fluctuate continuously due to temporal variations of the ionosphere.

The aim of this research is to study maximum usable frequency (MUF) using Wavelet I-D Haar 1-5 levels for approximation and decomposition over Pakistan region. Wavelet transform can map the power of a particular frequency at different times, giving an expansion of the signals in both time and frequency. For this analysis, a total 151 days of maximum usable frequency data (1st Oct-2015 to 29th Feb-2016) of Pakistan air space were analyzed which have been collected from Space and Upper Atmosphere Research Commission (SUPARCO).

The results obtained in this presentation depict significant information about maximum usable frequency, scale variation of the signals and clearly detect the time localization of the period changes. The constructed models of Maximum usable frequency (MUF) data, the detailed and approximated parts at the lowest resolution several peaks are appeared that can be characterized

This study will encourage the investigators working in the same field to verify the complex nature of ionospheric signals.

1. INTRODUCTION

We know that satellite communication systems necessitate to appraise the monitoring of ionospheric effects via applications of satellite navigation. Sometimes degradations in their performance will be observed. Weakly ionized plasma of solar atmosphere that ranges from 60km to beyond 1000km altitude above the earth atmosphere and electrically neutral to high degree of approximation It has been mentioned that the positive and negative charged particles are always formed and destroyed together in such a situation The term plasma is defined as "*quasi-neutral gas of charged and neutral particles that exhibits in collective behavior*"

Ionospheric regulatory has been characterized as a major factor to present both radio-communications and navigation systems and Space–Earth coupling, climatological global change, and anthropogenic activities impact issues. Therefore, establishing a statistical description of Ionosphere variations and relating them to possible driving sources such as solar and geomagnetic behaviors. The Ionosphere irregularities are described in terms of long-term trends and short-term fluctuations

The solar atmosphere is essential for Sky-wave radio propagation and illustrates the basis for almost all HF communications beyond line of sight. Here one of the variables Maximum useable frequency (MUF) that poses a significant ionospheric figure for satellite navigation and satellite communication systems and a detailed study of its behavior and especially of its diurnal variability. The term Maximum useable frequency (MUF) defined to the highest possible frequency for high frequency (HF) sky-wave transmission that can be used to broadcast over the given path under given Ionospheric conditions that fluctuate continuously due to temporal variations in the ionosphere

2. MANIFESTATION OF MUF

The sky wave Radio propagation users have great interest to know the daily behavior of the maximum usable frequency (MUF) for its applications. The MUF is obviously dependent on the ionospheric F-layer critical frequency (foF_2) and its corresponding propagation factor M (3000) of F_2 layer. As we know that the maximum usable frequency (MUF) manifests its importance for radio users in order to achieve better frequency execution. Mathematically, the MUF is the product of F_2-layer critical frequency foF2 and propagation factor M (3000) F_2 that is defined by the equation,

$$\text{MUF (3000) } F_2 = f_oF_2 * \text{M (3000) } F_2 \tag{1}$$

where, MUF (3000) is Maximum useable frequency received at 3000km when reflected by ionosphere and described the critical frequency or highest frequency reflected at vertical incidence. From a given layer and $M(3000)F_2$ the propagation factor of F_2 layer of which its influence is significantly less than foF_2.

It also follows a secant law if the angle of incidence or in other words launching angle as indicted between the incident and the normal and considered as θ, as it is given by

$$\text{MUF} = f_c \sec \theta \tag{2}$$

If we intend to use this frequency a flat surface of the earth or flat reflecting layer can be utilized It can be determined by the distance between the points that are to be met by a link known as Sky-wave. MUF can be explained on the two such points rather than the angle of incidence, while the highest frequency that can be utilized for Sky wave communication link of the earth. It has different values. Normally MUF varies from 8-35MHz. Due to solar activity it may take as high as 50 MHz.

3. WAVELET DESCRIPTION FOR MUF

The aim of the wavelet approach in space weather application is the time-frequency decomposition (Domingues et al., 2004). Wavelet analysis is a particular time-or space-scale illustration of signals that has been found in atmospheric application i.e.

maximum useable frequency (MUF) fluctuations due to ionosphere and the radio wave interactions. It is newly developed based signal processing approaches allows on several time scales of the local properties of complex signals that can present non stationary areas (Zai and Mian, 2012). The wavelet approach used to find out the mode of variation and also to explore how it fluctuate with time by decomposition time series in to frequency domain and powerful multi resolution with respect to one dimensional Haar level. Wavelet analysis and synthesis are consider highly important in the field of signal processing which decomposes actual signal in to different signals to be analyzed in to principle and residual part (Mian and Zai, 2013).

4. DATA AND METHOD OF ANALYSIS

The MUF data of five months (1^{st} Oct-2015 to 29^{th} Feb-2016) of Pakistan air space region taken from Space and Upper Atmosphere Research Commission (SUPARCO) which updates daily space weather data.

There are several techniques which were used to study the Ionospheric irregularities based on solar activities and ionospheric properties. For this purpose, wavelet technique was applied to investigate the nonlinear behavior of Ionosphere.

It is known that the coefficient of wavelet series realized the properties like strong transient, discontinuity and unforeseen of the function or distribution precisely thus wavelet analysis are considered as a mathematical microscope that detects the aspects of functions at different resolution (Domingues et al., 2004). Mathematically, wavelet analysis defined as

$$S = Aj + \sigma\, Dj \tag{3}$$

$$S = Aj + Dj \tag{4}$$

where, S, Aj and Dj and are defined by principle j level and residual part j level. Further, the relation between principle and residual part j level expressed as

$$Aj\text{-}1 = Aj + Dj \tag{5}$$

In 1910, A. Haar developed wavelet, considered the most suitable function to the introduction and of understanding of wavelet analysis. This study discussed Haar of level 3, 4 and 5 with dyadic scale a=2 and considered levels 3, 4 and 5 the resolution is given by 1/a or 2^{-j} .Further, the simple orthogonal mother wavelet called Haar wavelet presented as equation # (6)

$$\psi(t) = \begin{cases} 1, & 0 \le t < \frac{1}{2} -, \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{cases}$$

This detects signal abrupt variations i.e. one localized feature in the physical space (Domingues et al., 2004).

5. RESULTS AND DISCUSSIONS

The wavelet technique have been widely used in the atmospheric sciences in last decades. Now, the wavelet transforms had turned out to be a very valuable approach in atmospheric signal analysis, creating an encouraging new perspective to the research activities. By comparison, wavelet transform is better than Fourier transforms due to high-resolution features, good localization both in time and scale domains and its capacity of analyzing signals at multi time scales.

Figure 1 shows the variation of different resolution at level 1-3 of Haar wavelet form in the detailed and approximated part and the cyclic variation

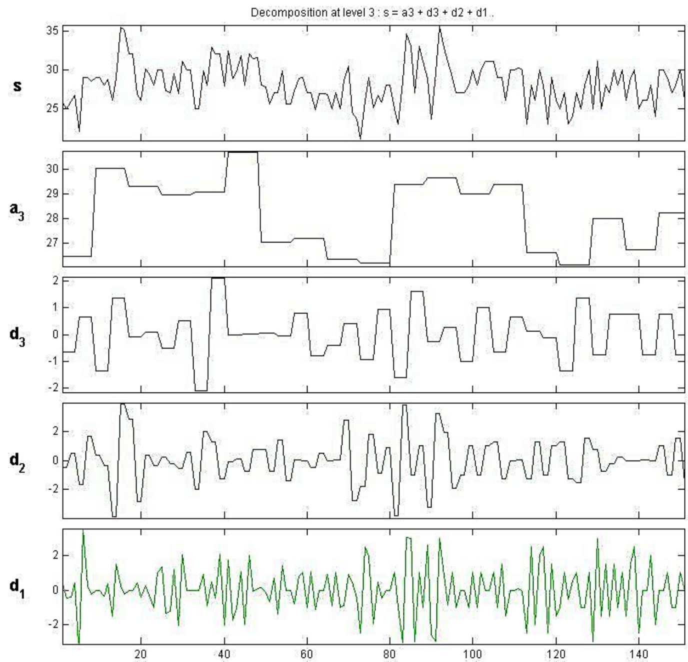


Figure 1

6. WAVELET TRANSFORM

Decomposition at level 3, s = $a_3 + d_3 + d_2 + d_1$ of Ionosphere of F2 layer MUF at Pakistan air space. For more convenient discussion about the behavior of wavelet, we compared this illustration to level 4 and level 5 of Haar wavelet transform as revealed in Figure 2.

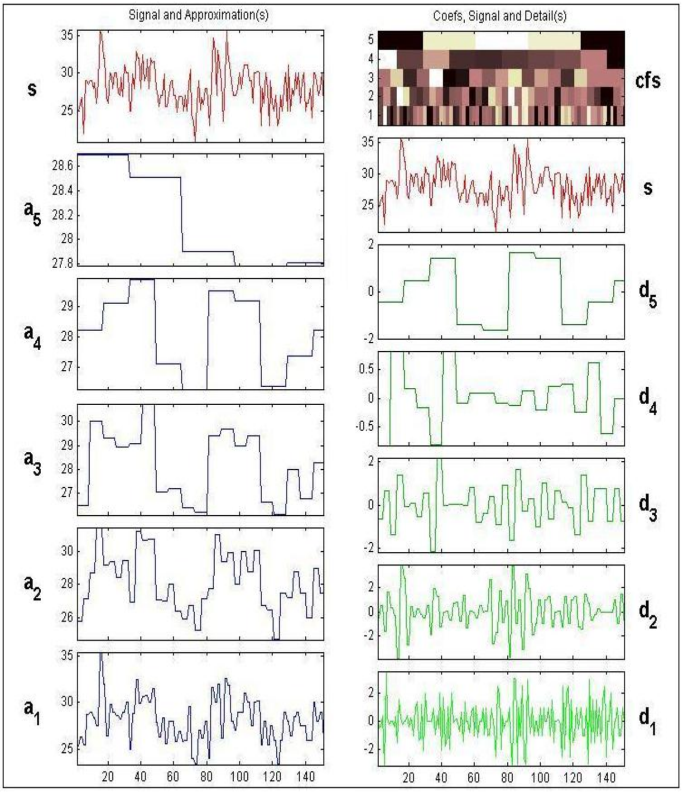
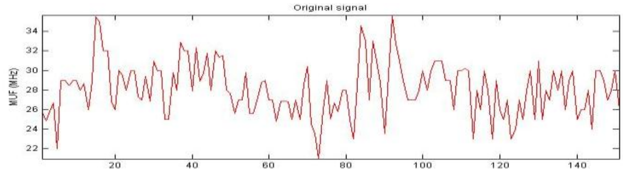


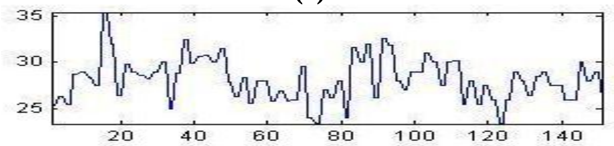
Figure 2

7. WAVELET COEFFICIENTS INDICATION

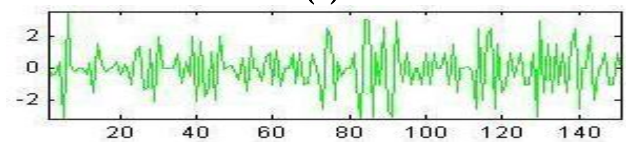
The wavelet irregularity can be used to find out the exact values of the distinctive frequencies. This can be only possible in Haar wavelet transform which has the best localization in frequency domain and also to observe the high values of the modulus of the wavelet coefficients indicate a transition region among unusual nature of movement (Domingues et al., 2004). Figure 3(a-c) indicate original signal, the upper panel approximation level 1 and lower panel details level 1 of MUF of F2.



(a)



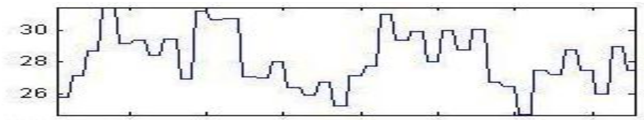
(b)



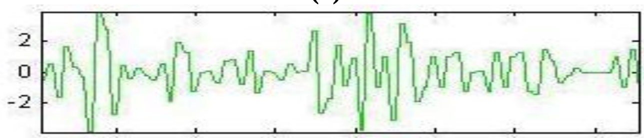
(c)

Figure 3: Exhibition of Different Signals with Original Signal

The Figure 4 (a-b) illustrated the upper panel approximation level 2 and lower panel detail level 2 of MUF of F2.



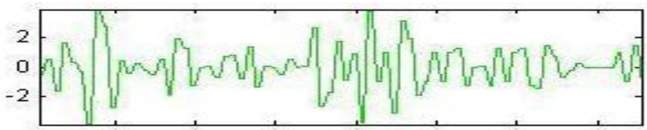
(a)



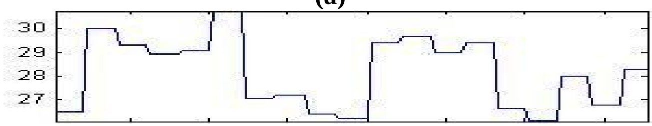
(b)

Figure 4

Figure 5(a-b) Plot of upper panel approximation level 3 and lower panel detail level 3 of MUF Ionosphere of F2 layer at Pakistan air space



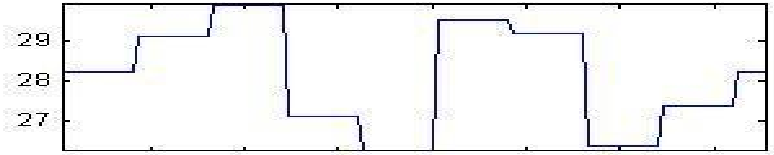
(a)



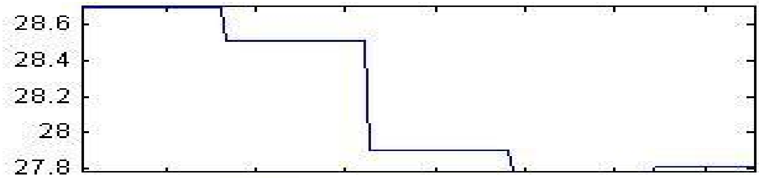
(b)

Figure 5

Figure 6(a-b) Plot of upper panel approximation level 4 and lower panel detail level 4 of MUF Ionosphere of F2 layer at Pakistan air space



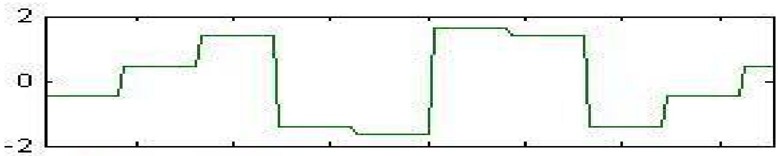
(a)



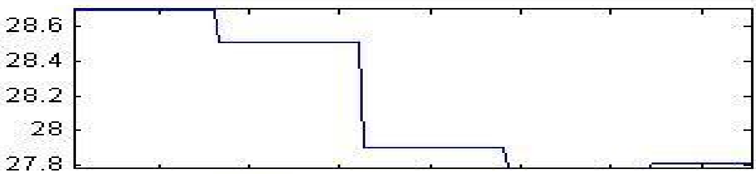
(b)

Figure 6

Figure 7(a-b) Plot of upper panel approximation level 5 and lower panel detail level 5 of MUF Ionosphere of F2 layer at Pakistan air space



(a)



(b)

Figure 7

8. COMMENTS AND CONCLUSION

This study has observed the variations of MUF for radio wave transmission of F2 layer using wavelet approach at Pakistan air space region. The meaningful data of maximum useable frequency was collected from Space and Upper Atmosphere Research Commission (SUPARCO) Pakistan which updates on daily basis. Wavelet, I-D Haar 1-5 levels for approximation and decomposition details are achieved. To observe the behavior of MUF, we have compared illustration of level 3, 4 and level 5 of Haar wavelet transform.

To declare a comprehensive sympathetic and to ensure the consistency of the obtained results in the investigation of space weather phenomena using this technique, a good knowledge of the wavelet implementation is required.

The constructed models of Maximum usable frequency (MUF) data, the detailed and approximated parts at the lowest resolution several peaks are appeared that can be characterized.

This study will support new horizon to the research activities going on in the field of Ionospheric signals Processing to verify their complex nature using wavelet interpretation of the complex signals

9. ACKNOWLEDGMENT

We are thankful to the technical staff of Space and Upper Atmosphere Research Commission (SUPARCO) for updating daily Ionospheric data like critical frequency, maximum usable frequency (MUF) and other variables on their respective website from where the space weather data can be freely downloaded on daily basis. Appreciation is due to the organizers of ISOSS conference for providing me Local hospitality and opportunity to present this piece of information among the eminent gathering

REFERENCES

1. Oyekola, O.S. (2015). Variation in the ionospheric propagating factor M (3000) F2 at Ouagadougou, Burkina Faso. *Advances in Space Research*, *46*(1), 74-80.
2. Malik, R.A., Abdullah, M., Abdullah, S. and Homam, M.J. (2016). Comparison of maximum usable frequency (MUF) variability over Peninsular Malaysia with IRI model during the rise of solar cycle 24. *Journal of Atmospheric and Solar-Terrestrial Physics*, *138*, 87-92.
3. Hadi, K.A. and Aziz, A.Z. (2014). Studying the Impact of the Solar Activity on the Maximum Usable Frequency Parameter Over Iraq Teritory. *IOSR Journal of Computer Engineering (IOSRJCE)*, *5*(3), 35-39.
4. Kouris, S.S. and Fotiadis, D.N. (2012). Ionospheric variability: a comparative statistical study. *Advances in Space Research*, *29*(6), 977-985.
5. Fotiadis, D.N., Baziakos, G.M., & Kouris, S.S. (2014). On the global behaviour of the day-to-day MUF variation. *Advances in Space Research*, *33*(6), 893-901.
6. Wen, Z. and Xun-Jie, Z. (2009). Predictions of HF communication MUF in the region of the South China Sea. *Antennas and Propagation Magazine, IEEE*, *41*(4), 35-38.
7. Malik, R.A., Khamis, K., Zakariya, A.S., Tompang, A., Mohamed, H.T., Kamal, M.R.A., Rajab, A.R., Lay, L.S., Kasmoni, H., Ramli, I. and Zahari, N. (2010). Determination of maximum usable frequency (MUF) for high frequency (HF) communications. *Def. Sci. Technol. Tech. Bull*., 3(1), 1-13.
8. Chen, C., Wu, Z.S., Sun, S.J., Ding, Z.H. and Zhao, Z.W. (2018). MUF variability at Haikou, China. In *Antennas, Propagation and EM Theory, 2008. ISAPE 2008. 8th International Symposium on* (pp. 400-403). IEEE.

9. Ijayyil, M. (2016). *Investigation of the variation of maximum usable frequency (MUF) to ensure a good communications link between two locations*, Thesis Master of Science, College of Science, University of Baghdad.
10. Roux, S.G., Knížová, P.K., Mošna, Z. and Abry, P. (2018). Ionosphere fluctuations and global indices: A scale dependent wavelet-based cross-correlation analysis. *Journal of Atmospheric and Solar-Terrestrial Physics*, *90*, 186-197.
11. Gilliland, T.R., Kirby, S.S., Smith, N. and Reymer, S.E. (1998). Maximum usable frequencies for radio sky-wave transmission, 1933 to 1937. *Radio Engineers, Proceedings of the Institute of*, *26*(11), 1347-1359.
12. Mudzingwa C., Nechibvute A. and Chawanda A. (2013). Maximum Useable Frequency Prediction Using Vertical Incidence Data. *International Journal of Engineering Research & Technology (IJERT)*, 2(8), 2050-2056.
13. Barclay, L.W. (2003). *Propagation of radio waves*, The Institution of Engineering and Technology, UK.
14. Chen, F.F. and Trivelpiece, A.W. (1986). Introduction to plasma physics. *Physics Today*, 29, 54.
15. Zai, M.A.K.Y. & Mian, K. (2012). Wavelet Characterization of Turbulence for Ionospheric Region of the Pakistan Air Space, *Proceedings of the Pakistan Academy of Sciences*, 49(4), 295-300.
16. Domingues, M.O., Mendes Jr, O. and Da Costa, A.M. (2004). On wavelet techniques in atmospheric sciences. *First Latin American Advanced School on Space Environment*. Elsevier Science.
17. Mian, K. and Zai, M.A.K.Y. (2013). Comparison of Wavelet Characterization of Computed Perturbed and Un-perturbed Parcel Velocities of F. *Proceedings of the Pakistan Academy of Sciences*, 50(1), 71-76.
18. Singh, R.C. and Bhatla, R. (2012). Wavelets in meteorology. In *American Institute of Physics (AIP) Conference Proceedings*, 1463(1), 176-193.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 31-41

BAYESIAN ANALYSIS OF FACTORS INFLUENCING FOOD SECURITY IN PAKISTAN

Syed Adil Hussain[1,3], Nasir Abbas[2] and **Gulraiz Ghaffar[1]**
[1] Department of Statistics, University of Gujrat, Gujrat, Pakistan
Email: syed.adil@uog.edu.pk
[2] Department of Statistics, Govt. Postgraduate College Jhang, Pakistan. Email: nabbasgcj@yahoo.com
[3] Department of Mathematics & Statistics, International Islamic University, Islamabad, Pakistan

ABSTRACT

The objective of this research is to identify the important influencing factors of food security and rank over them. This research considers the Bayesian inference for the factors influencing the food security. A sample survey of employees is carried out from various NGO's dealing with food security issues, situated in Lahore and Islamabad for employing the method of paired comparisons. Worth estimates are computed which reflects the overall worth probabilities for each factor of food security. The ranking is done and posterior predictive probabilities are computed for each of the possible fifteen pairs of six factors of food security. Results for analysis are computed in C language using programs coding for six parameters' inference.

KEY WORDS

Factors of Food Security, Worth probabilities, Bayesian Inference, Ranking, Paired Comparisons

INTRODUCTION

Pakistan is a lower middle-income country and the sixth most populous in the world. While undernourishment has slightly declined over the last two decades, it remains "moderately high". Strictly speaking, Pakistan is a food surplus country and a series of good harvests has meant that food availability is high. Despite that, 60 per cent of the country faces food insecurity, with many of the country's poorest people unable to afford an adequate diet. Pakistan is facing a major challenge of meeting food security and nutrition targets of SDG 2 by 2030. FAO estimates show that the Prevalence of Undernourishment (PoU) in Pakistan is 20.3 per cent, and the marginal reduction has been erased by population growth such that 40.0 million Pakistanis are not having enough food (According to Pakistan Overview of Food Security and Nutrition-2019).

In this study, six factors of food security are identified from the past literature (i.e., population growth, climate change, water resources, availability of arable lands, food availability and accessibility, and food loss), and discussed below.

The world population is the complete number of persons now existing. According to World Population Clock, 2020, it is estimated that the world population has reached 7.79 billion. The world population would be range from 8.0 to 10.4 billion in 2050. Using the information from Demographics of Pakistan 2020, there seems an increase of 1.05 % in population by comparing the population of the preceding year. Islam et al., (2017) provided deep understanding of the complex relationship between climate change and food security by critically examining both systems. Fahim, M. (2017) developed a series of models that captures the impact of water scarcity on the components of food security at Micro level. Aiqi Chen et al., (2019) considered the relationship between arable land demand to grain demand and production capacity and analysed the changes in national population, grain production, and consumption. Abbade (2017) identified the association between the levels of food utilization, food availability, and economic and physical accesses to food in developing countries, underlying the concept of food security. Sufficient food production and trade are essential to confirm suitable food availability. Irani et al., (2018) developed and applied a Fuzzy Cognitive Map (FCM) approaches to elucidate dynamic interrelationships between food distribution (organisational) and consumption (societal) factors.

Some other researchers also investigated on factors affecting food security. For example; Utuk and Daniel (2015) investigated land degradation as a threat to Food Security on a global assessment. Hall et al. (2017) investigated the impact of population growth and climate change on food security in Africa. Burke et al. (2010) found that the climate change, availability and accessibility of food as factors of food security in Portugal, lead to the food insecurity.

This study considers Bayesian inference for ranking the identified six factors of food security. The analysis would also produce estimates for posterior predictive probabilities and worth estimates.

METHODS & MATERIAL

This study considers method of paired comparisons and the Bradley-Terry (1952) model for the analysis purpose, given below.

2.1 The Bradley Terry Model for Paired Comparisons

For paired comparisons, Bradley-Terry (1952) established a basic model after Zermelo (1929) consideration which states that, "The objects or treatments have merit η_i and η_j when judge on some typical characteristics and may be represented by the continuous random variable with the following limit".

$wi, (-\infty < wi < +\infty),$"

$$H\left(\eta_i - \eta_j\right) = \frac{1}{4} \int_{-\left(\ln\theta_i - \ln\theta_j\right)}^{\infty} \sec h^2 (w/2) dw$$

where i $\neq j$ and $I = 1, 2 \ldots, m$

$$\Omega_{ij} = H\left(\eta_i - \eta_j\right) = \frac{1}{4} \int_{-\left(\ln\theta_i - \ln\theta_j\right)}^{\infty} \sec h^2(w/2)dw\,.$$

This model suggests that the difference amongst the two underlying variables $\left(\zeta_i - \zeta_j\right)$ has a logistic density with parameter $(\ln \omega_i - \ln \omega_{\mathrm{j}})$ and expressed as

$$\Omega_{ij} = \frac{\mu_i}{\mu_i + \mu_j}$$

where Ω_{ij} denotes the preference probability for object i when ω_i and ω_j are to be compared.

RESULTS AND DISCUSSION

In this section, we'll investigate and explore the collected data by using the method of Bayesian Inference. To find the posterior means we used uniform distribution. We calculated the posterior predictive probabilities (PPD) for the comparisons of all factors in each pair of single factors. Which allow us the assessment and comparison between the factors in each pair of a single factor. All results are given below in the tables with explanation and the clear interpretation.

Integration of six factors is not a simple task, as it is very complex to solve. For this difficult and more complex integration 'C' language programming is used and given below in the appendix. Data used for the paired comparison is also given below in the appendix.

3.1 Data (Appendix: Table-A1)

Table-A1 shows all possible pairs of factors and data according to preferences. 'PG' denotes the Population Growth; 'AL' denotes the Arable Lands, 'WR' denotes the Water Resources, 'CC' denotes the Climate Change, 'FAA' denotes the Food Availability and Accessibility, 'FL' denotes the Food Loss. Here, N_{ij} denotes the number of preferences on first factor, and the N_{ji} denotes the number of preferences on the second factor, for each pair of factors correspondingly.

3.2 Bayesian Inference

Bayesian analysis is performed on the composed data in Table-A1 by using C-language software with the help of most complex integrations programming and then also using the uniform prior distribution. Programming codes for this analysis are designed or planned for six factors and given in the Appendices table.

3.2.1 Bradley Terry Model

Bradley and Terry offered a model in mid of twentieth century (1952), which is called Bradley Terry model. It is a probability or likelihood model, to predict the outcomes of a comparison we used this model. Probability of observed outcomes in S^{th} recurrence of pair of parameters is given as follow:

$$\Omega_{i\,j} = \frac{\mu_i}{\mu_i + \mu_j}.$$

3.2.2 Notations for the Model

The following notations are used in the analysis of model.

$y_{i.ij}$ = Number of times the i^{th} object is preferred or ideal on the j^{th} object.

$y_{j.ij}$ = Number of times the j^{th} object is preferred or ideal on the i^{th} object.

α_{ij} = $y_{i.ij}$ + $y_{j.ij}$ = total number of comparisons in the i^{th} & j^{th} objects.

π_i = Total number of times the i^{th} object is favored or preferred on any other objects.

3.2.3 Likelihood Function of the Model

The likelihood function of the model for the six factors, influencing food security (an energy crisis) is given below:

$$l(y;\mu_1,\mu_2,\mu_3,\mu_4,\mu_5) \propto \frac{\prod_{i=1}^{6}(\mu i)^{\pi_i}}{\prod_{i<j}^{6}(\mu i+\mu j)^{\alpha_{i\,j}}}$$

Where $\pi_i = \sum_{j\neq i}^{m} y_{i.\,i\,j}$ and $\alpha_{i\,j} = y_{i.\,i\,j} + y_{j.\,i\,j}$ be the total number of comparisons between these factors or objects.

3.2.4 Joint Uniform Prior

For the paired comparison models, Bayesian analysis is not as simple as others. It is very difficult because no prior distribution is considered to be the conjugate prior of the data from the paired comparison models. Non- informative uniform prior is considered for the analysis. The joint uniform prior for the parameter space μ is given below:

$$P(\mu) \propto 1$$

Here, $\mu = \mu = (\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6)$

and, $0 < \mu < 1$

3.2.5 Posterior Distribution for Model of Factors Influencing Food Security

Joint posterior distribution with the uniform prior density holds the following form.

$$\text{P}(\underline{\mu}|y) = \frac{\prod_{i\,=\,1}^{6}(\mu i)^{\pi_{\,i}}}{M\left(\prod_{i<j}^{6}(\mu i+\mu j)^{\alpha_{\,i\,j}}\right)}$$

Here 'M' denotes the normalizing constant of the distribution and the condition is $\sum_{i=1}^{m} \mu_i = 1$.

3.2.6 Posterior Means for Factors Influencing Food Security

General integral formula for μ_i is given below.

$$E(\mu_i) = \int_{\mu_1=0}^{1} \int_{\mu_2=0}^{1-\mu_1} \int_{\mu_3=0}^{1-\mu_1-\mu_2} \int_{\mu_4=0}^{1-\mu_1-\mu_2-\mu_3} \int_{\mu_5=0}^{1-\mu_1-\mu_2-\mu_3-\mu_4} \frac{(\mu_i).\prod_{i=1}^{6}(\mu_i)^{\pi_i}}{M\left(\prod_{i<j}^{6}(\mu_i+\mu_j)^{\alpha_{ij}}\right)} d\mu_5 d\mu_4 d\mu_3 d\mu_2 d\mu_1$$

Here, $\mu_6 = 1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5$ is the constraint or restriction on the integration of numerical. The outcomes of posterior means are calculated and organized in Table A3 for posterior means as below;

3.3 Worth Probabilities of Factors Influencing Food Security (Appendix: Table–A2):

This table expresses the preferences of various employees of NGOs from different cities on different factors influencing the food security. The factor which has high or maximum probability value will be the most ideal or preferred factor for food security with respect to the respondent's point of views or the employees' point of views. And the factor which has minimum or low probability value will be the less important factor for the food security with respect to the respondent's point of views or the employees' point of views.

In this table, the probability value for the factor Food Availability and Accessibility is maximum and high among all the factors, that is 0.26493 and it is 26.5% out of 100%. So, these results show that the Food Availability and Accessibility factor is the most important factor among all other factors which are Influencing the food security and it got the 26.5% preference or favor among all the other factors. We can also interpret it as the food security will be high and more secure if food is available and accessible to all people easily.

Similarly, probability or likelihood value for the Water Resources is 0.25395 and it is almost 25%. We can say that water resources are also very important amongst all the factors which are Influencing food security.

The factor known as Food Loss has the probability or likelihood value 0.24879 and it is near to 25%. This shows that food loss is also play very important role amongst all the factors which are Influencing food security.

The factor Climate Change has the minimum probability value which is 0.06000 and it is just 6% out of 100%. This shows that the least important or minimum factor is climate change amongst all factors which are Influencing food security.

From these results we finally conclude that if food is available and people of that region also has access to that food, water resources are maximum and food loss and its wastage is at its minimum level, then we say that this region is almost 75% to 80% food secure. There is enough food for almost 75% to 80% people of specific region.

These results also show that food security in any region does not mostly depends on the climate changes and population growth rather food security mostly depends on the water resources, food loss and its availability and accessibility. We can also generalize these outcomes to the all-homogeneous regions.

3.4 Ranking of Factors Influencing Food Security (Appendix: Table–A3):

Table A3 shows the ranking of various factors, all factors are ranked according to their probability or likelihood values. The Food Availability and Accessibility factor has the maximum probability value that is 0.26493 amongst all those factors which affects the food security, so we'll give rank 1 to this factor amongst all the other factors. Similarly, we'll give the rank 2 to the Water Resources amongst all the factors which affects the food security, because its probability value is second highest 0.25395.

After these two above mentioned factors, Food Loss has the highest probability value that is 0.24879, so we'll give rank 3 to this factor. And similarly, then we'll give rank to Arable lands, Population growth and climate change as 4, 5 and 6 respectively according to their probability values.

3.5 Posterior Predictive Probabilities of Factor of Food Security (Appendix: Table–A4):

Posterior predictive probabilities (PPD) for six factors Influencing food security are calculated for each pair of factors by using a program designed for six factors in C^{++} the posterior predictive probabilities for i^{th} and j^{th} factors is given as follow for, i < j as:

$$P(ij) = \frac{1}{M} \int_{\mu_1=0}^{1} \int_{\mu_2=0}^{1-\mu_1} \int_{\mu_3=0}^{1-\mu_1-\mu_2} \int_{\mu_4=0}^{1-\mu_1-\mu_2-\mu_3} \int_{\mu_5=0}^{1-\mu_1-\mu_2-\mu_3-\mu_4} P(\underline{\mu} \mid \underline{y}).\Omega_{ij} d\mu_5 d\mu_4 d\mu_3 d\mu_2 d\mu_1$$

Here

$$P(\underline{\mu} \mid y) = \frac{\prod_{i=1}^{6} (\mu i / \mu i + \mu j)^{\pi_i}}{N\left(\prod_{i<j}^{6} (\mu i / \mu i + \mu j + \mu i / \mu i + \mu j)^{\alpha_{ij}} \right)}$$

Be the posterior distribution.

$\Omega_{ij} = \dfrac{\mu_i}{\mu_i + \mu_j}$ will be the preference probabilities of 'i' factor over 'j'.

Here $\mu_6 = 1 - \mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$ is the constraint on the numerical integration.

Table No. A4 presents the preference probabilities of i^{th} object or factor on j^{th} object or factor. Each value in the table shows the preference probabilities of one factor on the second factor in a single pair of factors which are Influencing food security. For example, the first pair of factors which are population growth and arable lands. In this pair of factors, the population growth factor has the preference probability on the other factor arable land is 0.36860 and it is almost 37% out of 100%. And the arable land factor has preference probability on the population growth is 0.6314 and it is 63% almost. This shows that the factor arable land is ideal and more preferred as compared to other factor that is population growth, because arable land's preference probability is 26% larger as compared to population growth.

Similarly, in the second pair of objects and factors, this pair is consisting of population growth and water resources. In this pair population growth factor has the preference probability on the water resources is 0.193114 and it is 19% out of 100% almost. And the water resources factor has preference probability on the population growth is 0.806866 and it is 81% almost. This shows that the factor water resources are ideal and more preferred as compared to other factor that is population growth, because water resources' preference probability is 62% larger as compared to population growth.

In the third pair of object or factors, this is consisting of population growth and climate change. In this pair population growth has the preference probability on the climate change is 0.541105 and it is 54% out of 100% almost. And the factor climate change has the preference probability on the population growth factor is 0.458895 and it is 46% almost out of 100%. This shows that the factor population growth is ideal and more preferred as compared to other factor that is climate change, because population growth preference probability is 8% larger as compared to climate change.

In the fourth pair of object or factors, this is consisting of population growth and food availability and accessibility. In this pair population growth has the preference probability on the food availability and accessibility is 0.186405 and it is 19% out of 100% almost. And the factor food availability and accessibility have the preference probability on the population growth factor is 0.813595 and it is 81% almost out of 100%. This shows that the factor food availability and accessibility is ideal and more preferred as compared to other factor that is population growth, because food availability and accessibility preference probability is 62% larger as compared to population growth.

In the fifth pair of object or factors, this is consisting of population growth and food loss. In this pair population growth has the preference probability on the food loss is 0.196597 and it is 20% out of 100% almost. And the factor food loss has the preference probability on the population growth factor is 0.803403 and it is 80% almost out of 100%. This shows that the factor food loss is ideal and more preferred as compared to other factor that is population growth, because food loss preference probability is 60% larger as compared to population growth.

In the last pair of object or factors, this is consisting of food availability and accessibility and food loss. In this pair food availability and accessibility has the preference probability on the food loss is 0.516812 and it is 52% out of 100% almost. And the factor food loss has the preference probability on the food availability and accessibility factor is 0.483188 and it is 48% almost out of 100%. This shows that the factor food availability and accessibility is ideal and more preferred as compared to other factor that is food loss, because food availability and accessibility preference probability is 4% larger as compared to population growth.

Similarly, we can discuss each and every pair in this similar manner one by one.

COMMENTS & CONCLUSION

The employees of various NGO's highly preferred food availability and accessibility factor which is influencing food security amongst all the six factors. They also highly preferred water resources and food loss which has great influenced on food security. They less preferred population growth and climate change factor for food security. Results showed that it has minimum influenced on food security. The combined effect of food availability and accessibility, food loss and water scarcity has undeniably impacted the production of global food and security. The factor food availability and accessibility are ranked 1 amongst all the factors playing roll in Influencing food security.

Since many of the things in different regions or countries are directly related to good governance, there is a need to delink the political interest to serve mankind with minimum basic needs such as food. In this globalized era of the 21st century, many determinants of food security are trans-boundary and require multilateral agreements and actions for an effective solution. Food insecurity and hunger all aviation on a global scale are within reach provided that technological innovations are accepted and implemented at all levels.

- In this study we used paired comparison method. In future we can use multiple comparison methodology.
- Ranking of factor effecting food security of employees of NGOs from Lahore and Islamabad is done but in future we do this study at larger scale in all over the Pakistan.

ACKNOWLEDGMENT

The authors are thankful for valuable comments in conference session.

REFERENCES

1. FAO, IFAD, UNICEF, WFP and WHO. (2020). *The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets*. Rome, FAO. World Population Clock: 7.6 Billion People (2017) - Worldometers.
2. Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324-345.
3. Islam, M.S. and Wong, A.T. (2017). Climate change and food in/security: A critical nexus. *Environments*, 4(2), 38.
4. Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1), 436-460.
5. Fahim, A. (2017). Water Scarcity: Impacts on Food security at Macro, Meso and Micro levels in Pakistan. *MPRA Paper* No. 78295

6. Chen, A., He, H., Wang, J., Li, M., Guan, Q. and Hao, J. (2019). A study on the arable land demand for food security in China. *Sustainability*, 11(17), 4769.
7. Abbade, E.B. (2017). Availability, access and utilization Identifying the main fragilities for promoting food security in developing countries. *World Journal of Science Technology and Sustainable Development*, 14(4), 322-335.
8. Irani, Z., Sharif, A.M., Lee, H., Aktas, E., Topaloğlu, Z., van't Wout, T. and Huda, S. (2018). Managing food security through food waste and loss: Small data to big data. *Computers & Operations Research*, 98, 367-383.
9. Utuk, I.O. and Daniel, E.E. (2015). Land degradation: a threat to food security: A global assessment. *Journal of Environment and Earth Science*, 5(8), 13-21.
10. Hall, C., Dawson, T.P., Macdiarmid, J.I., Matthews, R.B. and Smith, P. (2017). The impact of population growth and climate change on food security in Africa: looking ahead to 2050. *International Journal of Agricultural Sustainability*, 15(2), 124-135.

APPENDIX

Table A1
Preference Data Collected on All Factors

Factors	$N_{ij} = 0$	$N_{ji} = 1$
PG, AL	09	27
PG, WR	07	29
PG, CC	24	12
PG, FAA	07	29
PG, FL	11	25
AL, WR	11	25
AL, CC	25	11
AL, FAA	09	27
AL, FL	07	29
WR, CC	29	07
WR, FAA	07	29
WR, FL	29	07
CC, FAA	07	29
CC, FL	08	28
FAA, FL	07	29

Table A2
Worth Probabilities of Factors Influencing Food Security

Population Growth	Arable Lands	Water Resources	Climate Change	Food Availability and Accessibility	Food Loss
0.06119	0.11114	0.25395	0.06000	0.26493	0.24879

Table A3
Ranking of Factors Influencing Food Security

Factors Influencing Food Security	Expected Probabilities (θ_i)	Rank
Food Availability and Accessibility	0.26493	(1)
Water Resources	0.25395	(2)
Food Loss	0.24879	(3)
Arable Lands	0.11114	(4)
Population Growth	0.06119	(5)
Climate Change	0.06000	(6)

Table A4
Preference Probabilities of Factors Influencing Food Security

P(ij)	Estimate	P(ji)=1 - P(ij)	Estimate
P(12)	0.368600	P(21)	0.631400
P(13)	0.193134	P(31)	0.806866
P(14)	0.541105	P(41)	0.458895
P(15)	0.186405	P(51)	0.813595
P(16)	0.196597	P(61)	0.803403
P(23)	0.286526	P(32)	0.713474
P(24)	0.672124	P(42)	0.327876
P(25)	0.277397	P(52)	0.722603
P(26)	0.291177	P(62)	0.708823
P(34)	0.836525	P(43)	0.163475
P(35)	0.488800	P(53)	0.511200
P(36)	0.505616	P(63)	0.494384
P(45)	0.157361	P(54)	0.842639
P(46)	0.166617	P(64)	0.833383
P(56)	0.516812	P(65)	0.483188

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 43-50

MEDIALS AND SEMIGROUPS SATISFYING LEFT AND RIGHT DOUBLE DISPLACEMENT LAW

Syed Aleem Shah§, Nisar Ahmad, Abdus Saboor
Shahid Shehzad, Nasim Ullah and **Wahid Ullah**
Institute of Numerical Sciences, Kohat University of Science and Technology,
Kohat, Pakistan
Email: 1984aleem@gmail.com
nisarnn22@yahoo.com
saboorhangu@gmail.com
shaniktk992@gmail.com
nasimkhattal64@gmail.com
wahidktk16@gmail.com

ABSTRACT

In this paper we have solved open problem in [1] and developed the idea that on what conditions a non-commutative medial, semigroup and paramedial becomes left double displacement semigroup (LDD-Semigroup) and right double displacement semigroup (RDD-Semigroup). We also elaborated that on what conditions medial, paramedial, semigroup, LDD-Semigroup and RDD-Semigroup become commutative. We proved that the relation of left almost semigroup (LA-Semigroup) with LDD-Semigroup, right almost semigroup (RA-Semigroup) with RDD-Semigroup and semigroup with paramedial is only commutative property and example discussed in [12] by Nares and Chaiwat is wrong example constructed on the relation of semigroup with paramedial.

KEY WORDS

Medial; Semigroup; LDD-semigroup; RDD-semigroup; Paramedial.

1. INTRODUCTION

The term semigroup was first used in mathematical literature by French Mathematician Seguier (1904) which is an algebraic structure S that holds associative law i.e. $(ab)c = a(bc)$. Howie (1976) wrote comprehensive book on semigroup theory and elaborated the concept of E-semigroups, regular semigroups and also discussed about left ideals, right ideals and ideals. Kazim and Naseeruddin (1972) introduced the concept of LA-semigroup, RA-Semigroup and Almost semigroup and elaborated that a groupoid S is LA-semigroup that satisfies the condition $(ab)c = (cb)a$ and by the same way a groupoid S is RA-semigroup that satisfies the condition $a(bc) = c(ba)$. Mushtaq and Yousaf (1978) developed the idea of locally associative LA-semigroups that hold condition $(aa)a = a(aa)$. Further Mushtaq (1983) elaborated that on what conditions LA-semigroups becomes commutative group. This is proved in (1972, 1978) that "If S is LA-Semigroup or RA-Semigroup then S holds medial law i.e. S holds the condition $(ab)(cd) = (ac)(bd)$ but the converse may not be true". LA-semigroup is also known as

Abel Grassmann's Groupoid (AG-Groupoid) in (2008) and left invertive groupoid in (1993). Jazek and Kepka (1984) developed results on almost semigroup which is an algebraic structure S satisfying both left and right invertive law and used right modular groupoid for LA-semigroup, left modular groupoid for RA-semigroup and bimodular groupoid for Almost Semigroup. Cho et al. (1999) elaborated the concept of paramedial groupoid S that satisfies condition $(ab)(cd) = (db)(ca)$ and discussed its relation with commutative groupoid and medial. Madad (2008) and then Madad et al. (2015) elaborated about idempotent LA-Semigroup known as LA-Band or AG-Band and proved that LA-Semigroup with left identity holds paramedial law and by this way the relation of medial, paramedial and LA-Semigroup was developed. Nisar et al. (2018) introduced concept of LDD-Semigroup which is an algebraic structure S that holds left double displacement law (LDD-law) i.e. $(ab)(cd) = (cb)(ad)$ and explained R^+ is LDD-semigroup w.r.t. binary operations $a.b = e^b, a\eta b = b^b$ and $a * b = ln(b)$. Nares and Chaiwat (2019) discussed that S is semigroup and paramedial in *"Example 2.2"*; we proved in this paper that *"Example 2.2* discussed in (2019) is semigroup with left identity but not paramedial. We used concepts elaborated in (1904 – 2019) and solved open problem in (2018) with following examples:

Open Problem in [12]: Here we only study about LDD-semigroup where RDD-semigroup is left as an open problem for researchers.

Non Commutative Medials which are RDD-Semigroup:

i) If we define binary operation on R^+ by $ab = e^a$ then R^+ is neither commutative nor associative groupoid that satisfies medial law as well as RDD-law. This is medial and RDD-semigroup without idempotent.
ii) If we define binary operation on non-negative real numbers $(R - R^-)$ by $ab = a^k$ where $k \neq 0$ and $k \neq 1$ then R^+ is medial and RDD-semigroup.
iii) If we define binary operation on R^+ by $ab = ln(a + 2)$ then R^+ is medial and RDD-semigroup without idempotent.
iv) If M is set of square matrices of order $n \times n$ and binary operation on M is defined by $AB = A^T$ where A^T means transpose of matrix A then M is medial and RDD-semigroup.

Non Commutative Semigroups which are either LDD or RDD Semigroup:

i) The set of all constant functions say $F(S)$ defined from non-empty set S to S is semigroup, medial and LDD-semigroup w.r.t binary operation composition of mapping.
ii) If we define binary operation on $P(S)$ by $AB = (A \cup B) - B^C$ where $P(S)$ is power set of a non-empty set S then $P(S)$ is semigroup, medial and LDD-Semigroup.
iii) If we define binary operation on set of all collinear vectors say V by $uw = (\hat{u}.w).\hat{w} \,\forall\, u$ and $w \in V$ where $\hat{u}$ and $\hat{w}$ are unit vectors of vectors u and w respectively then V is semigroup, medial and LDD-Semigroup.

iv) The set of all constant functions say F(S) defined from non-empty set S to S is semigroup, medial and RDD-semigroup w.r.t binary operation composition of function.

v) If we define binary operation on $P(S)$ by $AB = (A \cup B) - A^C$ where $P(S)$ is power set of a non-empty set S then $P(S)$ is semigroup, medial and RDD-Semigroup.

vi) If we define binary operation on set of all collinear vectors say V by $uw = (\hat{w}.u).\hat{u}\ \forall\ u$ and $w \in V$ where û and ŵ are unit vectors of vectors u and w respectively then V is semigroup, medial and RDD-Semigroup.

2. OUR RESULTS

2.1: Non commutative medial T is LDD-semigroup if $\forall a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ba)(cd)$
(ii) $(ab)(cd) = (ca)(bd)$
(iii) $(ab)(cd) = (bc)(ad)$.

2.2: Non commutative LDD-semigroup T is medial if ∀ a, b, c, d ∈ T, T holds any condition from:

(i) $(ab)(cd) = (ba)(cd)$
(ii) $(ab)(cd) = (ca)(bd)$
(iii) $(ab)(cd) = (bc)(ad)$.

2.3: Non commutative medial T is RDD-semigroup if $\forall a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ab)(dc)$
(ii) $(ab)(cd) = (ac)(db)$
(iii) $(ab)(cd) = (ad)(bc)$.

2.4: Non commutative RDD-Semigroup T is medial if $\forall\ a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ab)(dc)$
(ii) $(ab)(cd) = (ac)(db)$
(iii) $(ab)(cd) = (ad)(bc)$.

2.5: If T is medial and commutative then T is LDD-semigroup as well as RDD-semigroup.

Note: The only connection between LDD-Semigorup and RDD-Semigroup is commutative semigroup.

2.6: Non commutative semigroup T is LDD-Semigroup if T holds condition $a(bc) = b(ac)$ or $(ab)c = (ba)c$.

2.7: Non commutative semigroup T is RDD-Semigroup if T holds condition and holds the condition $(ab)c = (ac)b$.

Note in 2.6 and 2.7: In both cases semigroup T is medial. If any semigroup T satisfies both conditions then T is commutative semigroup.

2.8: Non commutative paramedial T is LDD-Semigroup if $\forall\ a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ab)(dc)$
(ii) $(ab)(cd) = (cb)(da)$
(iii) $(ab)(cd) = (db)(ac)$.

2.9: Non commutative LDD-Semigroup T is paramedial if $\forall\ a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ab)(dc)$
(ii) $(ab)(cd) = (cb)(da)$
(iii) $(ab)(cd) = (db)(ac)$.

2.10: Non commutative paramedial T is RDD-Semigroup if $\forall\ a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ba)(cd)$
(ii) $(ab)(cd) = (ca)(bd)$
(iii) $(ab)(cd) = (bc)(ad)$.

2.11: Non commutative RDD-Semigroup T is paramedial if $\forall\ a, b, c, d \in T, T$ holds any condition from:

(i) $(ab)(cd) = (ba)(cd)$
(ii) $(ab)(cd) = (ca)(bd)$
(iii) $(ab)(cd) = (bc)(ad)$.

Next we show on what conditions groupoids, medials, semigroup, LDD-Semigroup, RDD-Semigroup and paramedial become commutative by the following results:

2.12: If groupoid S is medial, paramedial and LDD-semigroup then T is commutative.

2.13: If groupoid S is medial, paramedial and RDD-semigroup then T is commutative.

2.14: Semigroup S is commutative if S holds condition $(ab)c = (bc)a$ or $(ab)c = (ca)b$.

2.15: LDD-semigroup S is commutative if S satisfies any condition from the following:

(i) $(ab)(cd) = (ba)(dc)$
(ii) $(ab)(cd) = (dc)(ba)$
(iii) $(ab)(cd) = (ad)(bc)$
(iv) $(ab)(cd) = (ac)(db)$

2.16: RDD-semigroup S is commutative if $\forall\ a, b, c, d \in S, S$ satisfies any condition from the following:

(i) $(ab)(cd) = (ba)(dc)$
(ii) $(ab)(cd) = (dc)(ba)$
(iii) $(ab)(cd) = (bc)(ad)$
(iv) $(ab)(cd) = (ca)(bd)$

2.17: If each element $x \in$ groupoid S is idempotent element then S is commutative if S satisfies any condition from:
(i) $(ab)(cd) = (ba)(dc)$
(ii) $(ab)(cd) = (dc)(ba)$.

Next we show that on what conditions LDD-Semigroup becomes semigroup with left identity and RDD-Semigroup becomes semigroup with right identity by the following results:

2.18: If groupoid S satisfies LDD-law, medial law and contains left identity then S is semigroup with left identity.

Proof: From the given conditions we do the following steps:

$$\begin{aligned}(\mathbf{ab})\mathbf{c} &= (\mathbf{ab})(\mathbf{ec}) = (\mathbf{eb})(\mathbf{ac})\\ &= \mathbf{b}(\mathbf{ac}) = (\mathbf{eb})(\mathbf{ac})\\ &= (\mathbf{ea})(\mathbf{bc}) = \mathbf{a}(\mathbf{bc})\end{aligned}$$

Thus S holds associative law so S is semigroup with left identity.

Exp: If we define binary operation on *R*, *Q* or *Z* by $ab = |a|.b$ where means product then *R*, *Q* and *Z* are semigroups with left identity that satisfy medial law as well as LDD-law.

2.19: If groupoid S satisfies RDD-law, medial law and contains right identity then S is semigroup with left identity.

Proof: From the given conditions we do the following steps:

$$(\mathbf{ab})\mathbf{c} = (\mathbf{ab})(\mathbf{ce}) = (\mathbf{ac})(\mathbf{be}) = (\mathbf{ac})\mathbf{b}$$
$$\mathbf{a}(\mathbf{bc}) = (\mathbf{ae})(\mathbf{bc}) = (\mathbf{ac})(\mathbf{be}) = (\mathbf{ac})\mathbf{b}$$

We can also do the following steps:

$$\mathbf{a}(\mathbf{bc}) = (\mathbf{ac})\mathbf{b} = (\mathbf{ac})(\mathbf{be}) = (\mathbf{ab})(\mathbf{ce}) = (\mathbf{ab})\mathbf{c}$$

Thus S holds associative law so S is semigroup with right identity.

Exp: If we define binary operation on *R*, *Q* or *Z* by $ab = a.|b|$ where means product then *R*, *Q* and *Z* are semigroups with left identity that satisfy medial law as well as LDD-law.

Next we discuss the relation of LA-Semigroup with LDD-Semigroup and RA-Semigorup with RDD-Semigroup by the following theorems:

2.20: If *T* is RDD-semigroup with left identity then *T* is commutative monoid

2.21: Non commutative medial *T* is LDD-Semigroup if *T* holds condition $(ab)c = (ba)c\ \forall\ a, b, c \in T$ and vice versa.

2.22: If *T* is LA-Semigroup and LDD-semigroup then T is commutative.

Proof: Every LA-Semigroup holds medial law proved in [3] and also using *2.21* that if medial holds the condition $(ab)c = (ba)c$ then T is LDD-Semigroup.

So $\forall\, a, b, c \in T$ the conditions $(ab)c = (cb)a = (bc)a = (ac)b = (ca)b$ are satisfied. So using the conditions we do the following steps:

$$\begin{aligned}(\mathbf{ab})(\mathbf{cd}) &= \big((\mathbf{cd})\mathbf{b}\big)\mathbf{a} = \big((\mathbf{cb})\mathbf{d}\big)\mathbf{a} = ((\mathbf{cb})\mathbf{a})\mathbf{d} \\ &= \big((\mathbf{ab})\mathbf{c}\big)\mathbf{d} = \big((\mathbf{ab})\mathbf{d}\big)\mathbf{c} = (\mathbf{cd})(\mathbf{ab})\end{aligned}$$

2.23: Non commutative medial T is RDD-Semigroup if T holds condition $a(bc) = a(cb)\ \forall\, a, b, c \in T$ and vice versa.

2.24: If T is RA-semigroup and RDD-semigroup then T is commutative.

Proof: Every RA-semigroup holds medial law proved in [3] and also using *2.21*, so $\forall\, a, b, c \in T$ the conditions $a(bc) = a(cb) = b(ca) = b(ac) = c(ab)$ are satisfied. So sing the given conditions we do the following steps:

$$\begin{aligned}(\mathbf{ab})(\mathbf{cd}) &= \mathbf{d}(\mathbf{c}(\mathbf{ab})) = \mathbf{d}(\mathbf{c}(\mathbf{ba})) \\ &= \mathbf{c}(\mathbf{d}(\mathbf{ba})) = \mathbf{c}(\mathbf{b}(\mathbf{da})) \\ &= \mathbf{b}(\mathbf{c}(\mathbf{da})) = \mathbf{b}(\mathbf{c}(\mathbf{ad})) \\ &= \boldsymbol{b}(\boldsymbol{a}(\boldsymbol{cd})) = (\boldsymbol{cd})(\boldsymbol{ab})\end{aligned}$$

2.25: Paramedial S with right identity is RA-Monoid and vice versa.

2.26: If S is medial and satisfies the property $(ab)(cd) = (dc)(ba)$ then S satisfies paramedial law and vice versa.

Indicating Error in Example 2.2 in [13]:

Nares and Chaiwat constructed an example on a set A and discussed that this semigroup holds paramedial law by the following table:

~	α	β	ř
α	α	α	α
β	α	β	ř
ř	α	α	α

Clearly this is semigroup with left identity β. This is set of three square matrices of order 2 × 2 say

$$O = \begin{Bmatrix} 0 & 0 \\ 0 & 0 \end{Bmatrix},\ L = \begin{Bmatrix} 1 & 0 \\ 0 & 0 \end{Bmatrix},\ J = \begin{Bmatrix} 0 & 1 \\ 0 & 0 \end{Bmatrix},$$

Clearly L is left identity because $LJ = J$ but L is not right identity because $JL = O$. A is not paramedial if we see $(ř\,\beta)(\beta\beta) = \alpha\beta = \alpha$ but $(\beta\beta)(\beta ř) = \beta ř = ř$.

2.27: If S is semigroup as well as paramedial then S is commutative.

Proof: There is no direct way to prove that S is commutative so we use different conditions i.e. $(ab)(cd) = a(b(cd))$ and $(db)(ca) = d(b(ca))$. So we use these conditions and take six elements a, b, c, d, f and g and do the following steps:

$$\begin{aligned}(\mathbf{ab})\big((\mathbf{cd})(\mathbf{fg})\big) &= (\mathbf{ab})\big((\mathbf{gd})(\mathbf{fc})\big) = (\mathbf{ab})(\mathbf{g}(\mathbf{d}(\mathbf{fc})))\\ &= \big((\mathbf{ab})\mathbf{g}\big)\big(\mathbf{d}(\mathbf{fc})\big) = \big((\mathbf{fc})\mathbf{g}\big)\big(\mathbf{d}(\mathbf{ab})\big)\\ &= (\mathbf{f}(\mathbf{cg}))((\mathbf{da})\mathbf{b}) = (\mathbf{b}(\mathbf{cg}))((\mathbf{da})\mathbf{f})\\ &= ((\mathbf{bc})\mathbf{g})\big((\mathbf{da})\mathbf{f}\big) = (\mathbf{fg})((\mathbf{da})(\mathbf{bc}))\\ &= \big((\mathbf{fg})\mathbf{d}\big)\big(\mathbf{a}(\mathbf{bc})\big) = ((\mathbf{fg})\mathbf{d})\big((\mathbf{ab})\mathbf{c}\big)\\ &= (\mathbf{cd})((\mathbf{ab})(\mathbf{fg}))\end{aligned}$$

This shows that $\forall x, y$ and $z \in S$ the condition $(xy)z = x(yz) = y(xz) = (yx)z$ is satisfied. Binary operation is well defined so if a and $b \in S$ then $ab \in T$ also and also by the same way cd as well as $fg \in S$ also. So S is commutative by using the *Results 2.6 and 2.12*.

2.28: If S is commutative groupoid then S is semigroup if S satisfies any condition from:

(i) $a(bc) = b(ac)$
(ii) $a(bc) = c(ba)$
(iii) $a(bc) = b(ca)$ or $a(bc) = c(ab)$
(iv) $(ab)c = (ac)b$
(v) $(ab)c = (cb)a$
(vi) $(ab)c = (bc)a$ or $(ab)c = (ca)b$

3. CONCLUSIONS AND COMMENTS

(i) Non commutative semigroups can satisfy medial law, LDD-law and RDD-law.

(ii) Paramedial with left identity is medial and LA-Monoid but this is not necessary that medial with left identity holds left invertive law e.g. if we define binary operation $*$ on R by $a * b = |a|.a^K.b$ where k is some finite positive integer. If k is some finite positive even number then R contains two left identities -1 and 1. By the same way paramedial with right identity is medial and RA-Monoid but this is not necessary that medial with right identity holds right invertive law e.g. if we define binary operation $*$ on R we define by $a * b = a.|b|.b^k$ where k is some finite positive integer. If k is some finite positive even number then R contains two right identities -1 and 1.

(iii) Medial, semigroup and LDD-Semigroup may contain more than one left identity and by the same way medial, semigroup and RDD-Semigroup may contain more than one right identity.

(iv) If $\forall\, a, b, c, d \in T$ the condition $(ab)(cd) = (bd)(ca)$ is satisfied then $(ab)(cd) = (da)(cb)$ also holds and vice versa and by the same way if the condition $(ab)(cd) = (db)(ac)$ is satisfied then $(ab(cd) = (cb)(da)$ also holds and vice versa.

(v) If $\forall\, a, b, c, d \in T$ the condition $(ab)(cd) = (ca)(bd)$ is satisfied then $(ab)(cd) = (bc)(ad)$ also holds and vice versa and by the same way if $(ab)(cd) = (ac)(db)$ is satisfied then $(ab)(cd) = (ad)(bc)$ also holds and vice versa.

(vi) RDD-semigroup T satisfying the property $(ab)(cd) = (ab)(dc)$ holds medial law but RDD-Semigroup T satisfying the condition $(ab)(cd) = (ba)(cd)$ holds paramedial law and by the same way LDD-semigroup T satisfying the property $(ab)(cd) = (ba)(cd)$ holds medial law but LDD-Semigroup T satisfying the condition $(ab)(cd) = (ab)(dc)$ holds paramedial law.

(vii) We investigated that property $(ab)(cd) = (dc)(ba)$ in a groupoid T holds when T is any one of the:
(a) LA-Monoid (b) RA-Monoid
(c) Commutative (d) Both Medial and Paramedial.

(viii) We investigated that if $\forall\, a, b, c, d \in$ groupoid T the condition $(ab)(cd) = (bd)(ac)$ holds then $(ab)(cd) = (ca)(db)$ and $(ab)(cd) = (dc)(ba)$ are also satisfied but found no example.

(ix) We investigated that if $\forall\, a, b, c, d \in$ groupoid T the condition $(ab)(cd) = (cd)(ba)$ holds then $(ab)(cd) = (ba)(dc)$ and $(ab)(cd) = (dc)(ab)$ are also satisfied but found no example.

REFERENCES

1. Ahmad, N., Ali, M., Ali, F. and Khattak, A.M. (2018). Left Double Displacement Semigroup: A First Result. *Matriks Sains Matematik (MSMK)*, 2, 37-39.
2. Cho, J.R., Ježek, J. and Kepka, T. (1999). Paramedial groupoids. *Czechoslovak Mathematical Journal*, 49(2), 277-290.
3. Holgate, P. (1992). Groupoids satisfying a simple invertive law. *Math. Stud*, 61(1-4), 101-106.
4. Howie, J.M. (1976). *An introduction to semigroup theory* (Vol. 7). Academic Press.
5. Jazek, J. and Kepka, T. (1984). Modular Groupoids, *Czechoslovak Mathematical Journal*, 34(3), 477-487.
6. Kazim M. and Naseeruddin M. (1972). On Almost Semigroups. *Alig. Bull. Math*, 2, 1-7.
7. Khan, M. (2008). *Some studies in AG*-groupoids*, Doctoral dissertation, Quaid-i-Azam University, Islamabad.
8. Khan, M., Smarandache, F. and Anis, S. (2015). *Theory of Abel Grassmann's Groupoids. Infinite Study*. Educational Publishers Columbus.
9. Mushtaq, Q. and Yusuf, S.M. (1978). On LA-semigroups. *Aligarh Bull. Math*, 8, 65-70.
10. Mushtaq, Q. (1983). Abelian groups defined by LA-semigroups. *Studia Scient. Math. Hungar*, 18, 427-428.
11. Mushtaq, Q. and Kamran, M.S. (2001). Finite AG-groupoid with left identity and left zero. *International Journal of Mathematics and Mathematical Sciences*, 27(6), 387-389.
12. Sawatraksa, N., Namnak, C. and Namnak, C. (2020). Some Remarks on Paramedial Semigroups and Medial Semigroups. *Thai Journal of Mathematics, Special Issue: Annual Meeting in Mathematics*, 167-176.
13. Seguier S. (1904). *Elements De La Theorie Des Groupes Abstraits*, (Elements of the Theory of Abstract Groups).

Proc. 18th international Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 51-59

RESULTS ON IDEMPOTENT AND COMMUTATIVE GROUPOIDS

Syed Aleem Shah§, Nisar Ahmad, Abdus Saboor, Wali Khan Mashwani, Muhammad Sagheer, Shahid Shehzad, Nasim Ullah and **Wahid Ullah**
Institute of Numerical Sciences, Kohat University of Science and Technology, Kohat, Pakistan
Email: 1984aleem@gmail.com
nisarnn22@yahoo.com
saboorhangu@gmail.com
mashwanigr8@gmail.com
sagheer.maths@gmail.com
shaniktk992@gmail.com
wahidktk16@gmail.com
nasimkhattal64@gmail.com

ABSTRACT

In this paper we developed the concept that on what patterns and binary operations finite idempotent commutative as well as non-commutative groupoids can be constructed. We elaborated that how commutative groupoids can be constructed that contain zero semigroup, semilattice and cyclic group. Further we extended some results on finite medials. We discussed a Matlab code by which we can easily find the number of idempotents of any mod n.

KEYWORDS

Idempotent; Commutative; Medial; Semigroup.

1. INTRODUCTION

The term semigroup was first used by French Mathematician Seguier (1904) which is an algebraic structure that holds associative law i.e. $(ab)c\ a(bc)$. Kimura (1957) and then Kimura and Yamada (1958) explained the idea of idempotent semigroups known as bands that satisfy property $a^2 = a$. Hall (1969) presented idea of such regular semigroups whose idempotents also form regular semigroup and also discussed the concept of commutativity, left normality, right normality and normality of regular semigroup. Howie (1976) wrote comprehensive books on semigroup theory and elaborated the idea of E-Semigroups, regular semigroups, orthodox semigroups and left ideals, right ideals, ideals, zero minimal ideals of semigroups and also defined that a commutative groupoid is such algebraic structure that satisfies the condition $ab = ba$.

The concept of semilattice i.e. commutative band and zero semigroup are also discussed by Howie (1976), Howie elaborated that contains zero semigroup is such algebraic that contains one zero element a and $\forall\, b, c \in S$ the condition $ab = ba = ac = ca = bb = cc = bc = cb = a$ is satisfied. Kazim and Naseeruddin (1972) introduced the concept of LA-Semigroup, RA-Semigroup and Almost semigroup and discussed that a groupoid S is

LA-Semigroup that satisfies left invertive law $(a * b) * c = (c * b) * a$ and if $a * (b * c) = c * (b * a)$ then *S* is RA-Semigroup and also elaborated that if *S* fulfils both the conditions then *S* is Almost semigroup. Concept of LA-Semigroup, RA-Semigroup and Almost semigroup and discussed that a groupoid *S* is LA-Semigroup that satisfies left invertive law $(a * b) * c = (c * b) * a$ and if $a * (b * c) = c * (b * a)$ then S is RA-Semigroup and also elaborated that if *S* fulfils both the conditions then *S* is Almost semigroup. Mushtaq and Yousaf (1978) developed results on locally associative LA-Semigroup S that satisfy property $(aa)a = a(aa)$. Mushtaq (1983) introduced the concept that on what conditions LA-Semigroup becomes commutative monoid and commutative group. This is proved in (1972) and (1978) that if *S* is either LA-Semigroup or RA-Semigroup then *S* hold medial or bisymmetry law i.e. $(ab)(cd) = (ac)(bd)$ but the converse may not be true and the connection of LA-Semigroup and RA-Semigroup with semigroup is only commutative property. Every commutative semigroup always holds medial law and in some specific cases non commutative semigroups also holds medial law elaborated by Howie (1976). Madad (2008) and then Madad et al. (2015) constructed examples on idempotent LA-Semigroups known as LA-Band or AG-Band. Jazek and Kepka (1984) developed results on Almost semigroup. Cho et al. (1999) discussed idea of paramedial groupoid or simply paramedial *S* that satisfies condition $(ab)(cd) = (db)(ca)$ and elaborated that on what conditions paramedial *S* becomes commutative, and also proved that if S is paramedial then S is medial if S satisfies one of the conditions: (a) Commutative (b) Left(Right) modular (c) Unipotent and left(right) cancellative or (d) Idempotent. Nisar et al. (2018)introduced concept of LDD-Semigroup S that holds condition $(ab)(cd) = (cb)(ad)$ and explained that on set ofpositive real numbers (R^+) if binary operation is defined by $ab = ln(b)$ and $ab = e^b$ then R^+ is medial as well as LDD-Semigroup w.r.t these binary operations. Further they (2018) defined RDD-Semigroup *S* an algebraic structure that satisfies the condition $(ab)(cd) = (ad)(cb)$ and put the discussion of RDD-Semigroup as an open problem.

2. METHODOLOGY

We analyzed concepts discussed in $[1 - 13]$ and by using some new patterns and binary operations we developed the following *Matlab Code for Finding Idempotents in any* Z_n*:*

```
function [] = Idempotent:
clc;
syms BO a b
Set = sym(input( 0 Set = 0 ));
Modulo = sym(input( 0 Modulo = 0 ));
Binary Operation = input( 0 Binary Operation (when a & b are any elements) = 0);

Table = subs(Binary Operation, a, Set);
Table = subs(Table, b, Set 0 );
Table = mod(Table, Modulo);
Diagonal = diag(Table) 0 ;
Idemp = Set(Diagonal == Set);
Display
Table
disp( 0 Cayley Table 0 )
```

disp([[BO; Set 0],[Set; Table]])
Idempotent element
disp([0 Idempotent = [0 num2str(double(Idemp)) 0] 0])
Number of idempotent
disp([0 Number of Idempotent = 0 num2str(length(Idemp))])
end.

3. RESULTS AND DISCUSSION

3.1: If n is finite positive odd integer with the condition $n \geq 3$ then Zn is finite commutative idempotent as well as medial if we define binary operation $*$ on Z_n by:

$$\boldsymbol{a * b = \left[\left(\frac{n+1}{2}\right)(a + b)\right]\{mod\ n\}} \qquad \text{(i)}$$

3.2: If n is finite positive even integer with the condition $n \geq 4$ then Zn is finite non commutative idempotent groupoid if we define binary operation * on Zn by:

$$\boldsymbol{a * b = \left[\left(\frac{e}{2}\right)a + \left(\left(\frac{e}{2}\right) + 1\right)b\right]\{mod\ n\}} \qquad \text{(ii)}$$

3.3: If n is any finite positive integer with the condition $n \geq 3$ then Zn is finite non commutative idempotent groupoid if we define binary operation * on Zn by:

$$\boldsymbol{a * b = [ga + hb]\{mod\ n\}} \qquad \text{(iii)}$$

Here g and h must satisfy the condition $g + h \equiv 1$ *(mod n)* and both g and h must be non zero.

3.4: If we define binary operation on order m of groupoid S where m is finite positive integer with the condition $m \geq 3$ in such way that an element $s_i \in S$ is idempotent and in the table in the column (row) entries succeeding pattern of g and then h elements is followed and then succeeding pattern of h and g elements is followed, then S is commutative idempotent groupoid but neither holds associative nor medial law.

Here g + h ≡ 1 (mod n).

3.5: Zn is finite commutative medial if we define binary operation * on Zn by:

$$\boldsymbol{ab = [f(a) + f(b)]\{mod\ n\}} \qquad \text{(iv)}$$

Note in 3.5: Zn is commutative w.r.t following conditions:

a) If $f(a) = 0$ then surely Zn is zero semigroup and holds medial law.
b) If $f(a) = k$ where k is some non zero constant then Z is commutative medial but not associative.
c) If $f(a) = a$ i.e. identity function then surely Z is cyclic group because this is trivial mod addition case.
d) If $f(a) = -a$ then Zn is commutative medial and if n is some finite even number with the condition $n \geq 4$ then Zn contains cyclic group of two elements $\{0, n/2\}$.
e) Zn is not medial if $f(a) = a^k$ where $k \neq 0$.

3.6: If order of finite groupoid V is even where binary operation $*$ is defined on V in such way that v_i is idempotent element and the row entries as well as column entries follow pattern of succeeding k elements where k is odd number with the condition $k \geq 3$ then groupoid V contains cyclic group of order h where either $\boldsymbol{k = j.h + 1}$ $\boldsymbol{or\ h = j.k + 1}$ and j is positive integer. Also in **3.6** we have achieved following results:

a) In all cases order of cyclic group divides the order of groupoid V i.e. $\mathbf{n(V)/h = g}$ where g is some finite positive integer.
b) When $\boldsymbol{k = 3}$ then there are exactly two elements in commutative groupoid V say $v\ i$ and $v\ j$ that forms cyclic group and $\mathbf{|i - j| = |j - i| = n(V)/2}$.
c) When $\mathbf{k = \left(\frac{n}{2}\right) - 1}$then there are exactly two elements in commutative groupoid V say $v\ i$ and $v\ j$ that forms cyclic group and $\mathbf{|i - j| = |j - i| = n(V)/2.}$
d) When $\mathbf{k = \left(\frac{n}{2}\right) + 1}$ then there are $(\boldsymbol{n/2})$ elements that forms cyclic groupand if $n/2$ is prime numbers then subset of V say V_C has behavior same like Zp w.r.t mod p addition.

3.7: If we define binary operation on any Zn by:

$$\mathbf{ab\ =\ (-a - b - K)\{mod\ n\}\ v - (a)}$$

So Zn is also commutative groupoid and if n is finite positive integer with the condition $n \geq 4$ then Zn contains cyclic group of order **2** and also in this case the subset of Zn say Z_C containing elements i and j forms cyclic group and following condition holds:

$$\mathbf{|i\ -\ h| =\ |h\ -\ i| = \frac{m}{2}}\ \boldsymbol{v - (b)}$$

Table 1
Table for Result 3.7 w.r.t mod 14

K	Cyclic Group	K	Cyclic Group
0	{0, 7}	**1**	{9, 2}
2	{4, 11}	**3**	{13, 6}
4	{8, 1}	**5**	{3, 10}
6	{12, 5}	**7**	{7, 0}
8	{2, 9}	**9**	{11, 4}
10	{6, 13}	**11**	{1, 8}
12	{10, 3}	**13**	{5, 12}

Note: We can also define different binary operations on any **mod n** e.g. if we define following binary operations:

$$\mathbf{ab\ =\ -(a + b - k)\{mod\ n\}}\ \boldsymbol{v - (c)}$$

$$\mathbf{ab\ =\ (a + b - k)\{mod\ n\}}\ \boldsymbol{v - (d)}$$

$$\mathbf{ab\ =\ (a + b + k)\{mod\ n\}}\ \boldsymbol{v - (e)}$$

3.8: Let V be groupoid of even order m with the condition $\boldsymbol{m = LK}$ with the condition either $\boldsymbol{L = 2.K - 1}$ or $\boldsymbol{K = 2.L - 1}$ and binary operation $*$ is defined on groupoid V in such way that v_i is idempotent of V and succeeding pattern of $\boldsymbol{K}$ element difference is used then following results hold:

a) L subsets of V which are zero semigroup of order k

b) A subset V_{CIG} of order L is commutative idempotent that is neither semigroup nor medial.

Note in 3.8(b): The general proof of this is not easy so to grasp this idea in easy way we give the short details of this result in *"Table for 3.8(b)"* in which we use these acronyms which are m (order of groupoid), $n(ZSG)$ means number of zero semigroups, $n(CIM)$ means number of commutative idempotent groupoid, $OEZSG$ means order of each zero semigroup, $OCIM$ means order of commutative idempotent groupoid.

Table 2
For Result 3.8(b)

$\boldsymbol{M}$	$\boldsymbol{K}$	$\boldsymbol{L}$	$\boldsymbol{n(ZSG)}$	$\boldsymbol{n(CIM)}$	$\boldsymbol{OEZSG}$	$\boldsymbol{OCIM}$
15	3	5	5	1	3	5
15	5	3	3	1	5	3
28	4	7	7	1	4	7
28	7	4	4	1	7	4
45	5	9	9	1	5	9
45	9	5	5	1	9	5
66	6	11	11	1	6	11
66	11	6	6	1	11	6
91	7	13	13	1	7	13
91	13	7	7	1	13	7
120	8	15	15	1	8	15
120	15	8	8	1	15	8

Note: This is quite easy to construct finite as well as infinite following semilattices:

a) If N_m is finite subset of non-negative numbers integers then we can construct finite semilattices if $\boldsymbol{ab = min(a, b)}$ then N_m is semilattice with zero element 0 and m is identity element and by the same way if we define binary operation by $\boldsymbol{ab = max(a, b)}$ then N_m is semilattice with zero element m and 0 is identity element.

b) The same idea can be extended to infinite sets e.g. if we define binary operation on Z^+ by $\mathbf{ab = min(a, b)}$ then 1 is zero element in Z^+ but Z^+ has no identity element and by the same way if we define binary operation on Z^- by $\mathbf{ab = max(a, b)}$ then Z^- has zero element $\mathbf{-1}$ and Z^- contains no identity element.

c) If we define binary operation on Z by $\boldsymbol{ab = min(a, b)}$ or $\boldsymbol{ab = max(a, b)}$ then Z is semilattice but this does not contain a zero element but if we do some change in the definition of binary operation by $\boldsymbol{ab = min(|a|, |b|)}$ then Z is semilattice with a zero element *0*. The same result can be achieved on uncountable set if we define the same binary operation on set of real numbers R.

3.9: Commutative Groupoids Formed by Some Experimental Data:

Let we have an experiment that we roll two dices and A be the event such that sum of two dices is even number then possible even sum of dices are 2, 4, 6, 8, 10, 12 and the outcomes for each even value are:

n(sum = 2) = 1 n(sum = 4) = 3 n(sum = 6) = 5

n(sum = 8) = 5 n(sum = 10) = 3 n(sum = 12) = 1

Also if we see the differences values then we have values *2, 2, 0, -2, -2*. So we can construct two groupoids of order **7** and **6** which are following:

Table 3

Table for No. of Outcomes and Differences

Table For Z_7 w.r.t No. of Sum Outcome							
~	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	**0**	1	4	2	**0**	3	4
1	1	2	5	3	1	4	5
2	4	5	1	6	4	0	1
3	2	3	6	4	2	5	6
4	**0**	1	4	2	**0**	3	4
5	3	4	0	5	3	6	0
6	4	5	1	6	4	0	1

Table for Z6 for Difference of Outcomes						
∘	**0**	**1**	**2**	**3**	**4**	**5**
0	***0***	2	4	4	2	***0***
1	2	4	0	0	4	2
2	4	0	**2**	2	0	4
3	4	0	2	2	0	4
4	2	4	0	0	**4**	2
5	***0***	2	4	4	2	***0***

$\boldsymbol{Z_7}$ is commutative groupoid which is neither associative nor medial but contains zero semigroup {0, 4} while Z_6 is commutative groupoid that contains idempotent {0, 2, 4}.

3.10: We also constructed finite medials and groupoids that contain more than two left or right identities e.g. if n ≥ 6 and we define following binary operations:

$$\mathbf{ab} = [\left(\mathbf{b} + \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\mathbf{a}\right] \{\mathbf{mod\ n}\}\ \mathbf{vi} - (\mathbf{a})$$

$$\mathbf{ab} = [\left(\mathbf{a} + \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\mathbf{b}\right] \{\mathbf{mod\ n}\}\ \mathbf{vi} - (\mathbf{b})$$

$$\mathbf{ab} = [\left(\mathbf{b} + \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\mathbf{a} \pm \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\right] \{\mathbf{mod\ n}\}\ \mathbf{vi} - (\mathbf{c})$$

$$\mathbf{ab} = [\left(\mathbf{b} + \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\mathbf{a} \pm \left(\frac{\mathbf{n}}{\mathbf{2}}\right)\right] \{\mathbf{mod\ n}\}\ \mathbf{vi} - (\mathbf{d})$$

Zn contains **(n/2)** idempotents and these idempotents form semigroup and LDD-Semigroup w.r.t **vi-(a)** and **vi-(c)** and Zn also contains **(n/2)** idempotents that form semigroup and RDD-Semigroup **vi-(b)** and **vi-(d)**.

Note in 3.10:

a) **vi-(a)** to **vi-(d)** represent the cases of neither commutative nor associative medial groupoids.

b) Medials and groupoids of finite odd and composite order can contain more than two left or right identities e.g. if we define binary operations on **mod 9** by:

$$\boldsymbol{ab} = [(\boldsymbol{b} + \boldsymbol{3a}]\{\boldsymbol{mod\ 9}\}\boldsymbol{vii} - (\boldsymbol{a})$$

$$\boldsymbol{ab} = [(\boldsymbol{a} + \boldsymbol{3b}]\{\boldsymbol{mod\ 9}\}\boldsymbol{vii} - (\boldsymbol{b})$$

4. APPLICATIONS

(i) Applications in Managerial Sciences:

If n is some finite positive odd number with the condition $n \geq 3$ and we define binary operation on Z_n by $ab = [(\boldsymbol{n} + \boldsymbol{1})/\boldsymbol{2}].(\boldsymbol{a} + \boldsymbol{b})]_{\{\boldsymbol{mod\ n}\}}$ then Z_n is commutative idempotent medial groupoid which is not semigroup, however this idea can be applied in any firm, organization and department in which person to person, office to office and even sub department to sub department, work relation and co-ordination of workers and offices can be defined.

By the same way if n is any positive integer with the condition $n \geq 3$ and we define binary operation on Zn by $\boldsymbol{ab}\ [\boldsymbol{ga} + \boldsymbol{hb}]_{\{\boldsymbol{mod\ n}\}}$ then this is non commutative groupoid however this is idempotent groupoid and any firm, organization or association can define the work relation of workers and offices can be defined.

(ii) Applications in Cryptography and Securing Information:

The idea to convert plain text into cipher text is a technique that is used to hide the information from hackers and can also makes the document confidential so only one or very few authorized officials or persons can see the information. The famous RSA (Rivest Shamir Adleman) algorithm [14-15] is used by multinational companies, industries and even by high govt. officials to encrypt and then to decrypt messages to keep information secret.

The benefit of using this technique is that every person has a public key which can be obtained by any one through authorized officials by using proper channel but the private key of each member is known by that member or person. If by some human or clerical a sender sends any message to many persons or members by using the public key of some member I (PB-I) then there is no need to worry because the public key of member I can only be opened by private key of member I (PR-I).

We can convert each word into cipher text e.g. if we use **mod n** concept and from A to Z we use numbers 0 to 25 and let we suppose a sentence “I have a question” then first we do coding and **I** appears as 8, **have** appears as 7-0-21-4, **a** appears as 0 and **question** appears as 16-20-4-19-20-8-14-13. Now its upto us that whether we do addition, multiplication of all these numbers under the **mod 26** or we take square, cube, nth power of all these numbers where n is some finite positive integer and then we add, subtract or multiply or use other binary operations these numbers under the **mod 26** and can also makes the document confidential so only one or very few authorized officials or persons can see the information.

The famous RSA (Rivest Shamir Adleman) algorithm [14-15] is used by multinational companies, industries and even by high govt. officials to encrypt and then to decrypt messages to keep information secret.

The benefit of using this technique is that every person has a public key which can be obtained by any one through authorized officials by using proper channel but the private key of each member is known by that member or person. If by some human or clerical a sender sends any message to many persons or members by using the public key of some member I (PB-I) then there is no need to worry because the public key of member I can only be opened by private key of member I (PR-I).

We can convert each word into cipher text e.g. if we use **mod n** concept and from A to Z we use numbers 0 to 25 and let we suppose a sentence "I have a question" then first we do coding and **I** appears as 8, **have** appears as 7-0-21-4, **a** appears as 0 and **question** appears as 16-20-4-18-19-8-14-13. Now its upto us that whether we do addition, multiplication of all these numbers under the **mod 26** or we take square, cube, nth power of all these numbers where n is some finite positive integer and then we add, subtract or multiply or use other binary operations these numbers under the **mod 26**.

5. COMMENTS AND CONCLUSION

a) Finite semigroups always contains idempotent but finite medials, medial and RDD-Semigroup and medial and LDD-Semigroup and LA-Semigroups may not contain idempotent element e.g. if we define binary operation on Zn by $ab = (b+k)_{\{mod\ n\}}$ or $ab = (a+k)_{\{mod\ n\}}$ where $k \neq 0$.
b) If S is medial with left identity then $\forall a, b, c \in S$ then S satisfies condition $a(bc) = b(ac)$ but S may not satisfy left invertive law e.g. if we define binary operation on $R\ or\ Q$ by $ab = |a|.a^k.b$ and by the same way if S is medial with right identity then $\forall a, b, c \in S, S$ satisfies condition $(ab)c = (ac)b$ but S may not satisfy right invertive law e.g. if we define binary operation on $R\ or\ Q$ by $ab = a.|b|.b^k$.
c) Identity element is always unique in groupoids, semigroup and groups but left identity and right identity may not be unique e.g. if we define binary operation on $R,\ Q\ or\ Z$ by $ab = |a|.b$or $ab = a.|b|$ respectively. Similarly we also discussed this concept in the *Result-3.10.*
d) If each element of paramedial groupoid S is idempotent then S is commutative proved in [12] but if each element of medial, LDD-Semigroup, RDD-Semigroup or semigroup is idempotent then this is not necessary that medial, LDD-Semigroup, RDD-Semigroup or semigroup is commutative.
e) The smallest neither commutative nor associative groupoid that is either LA-Semigroup and RA-Semigroup can be constructed of order 3 but neither associative nor
f) commutative medial and either LDD-Semigroup or RDD-Semigroup of order 2 can be constructed e.g. if we define following binary operations on **mod 2**:

$$\boldsymbol{ab = (b+1)\{mod\ 2)\ or\ ab = (a+1)\{mod\ 2\}}.$$

REFERENCES

1. Ahmad, N., Ali, M., Ali, F. and Khattak, A.M. (2018). Left Double Displacement Semigroup: A First Result. *Matriks Sains Matematik (MSMK)*, 2, 37-39.
2. Cho, J.R., Jazek, J., Kepka, T. (1999). Paramedial Groupoids. *Czechoslovak Mathematical Journal*, 49(124), 277-290.
3. Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C. (2001). *Introduction to Algorithms* (2nd ed), MIT Press and McGraw-Hill, 881-887.
4. Hall, T.E. (1969). On Regular Semigroups Whose Idempotents Forms A Subsemigroup, *Bull. Austraila. MATH. SOC*, 1, 195-208.
5. Howie, J.M. (1976). *An introduction to semigroup theory* (Vol. 7). Academic Press.
6. Jazek, J. and Kepka, T. (1984). Modular Groupoids, *Czechoslovak Mathematical Journal*, 34(3), 477-487.
7. Kazim, M. and Naseeruddin, M. (1972). On Almost Semigroups. *Alig. Bull. Math.*, 2, 1-7.
8. Khan, M. (2008). *Some studies in AG*-groupoids*, Doctoral dissertation, Quaid-i-Azam University, Islamabad.
9. Khan, M., Smarandache, F. and Anis, S. (2015). *Theory of Abel Grassmann's Groupoids. Infinite Study*. Educational Publishers Columbus.
10. Kimura, N. (1957). Note on idempotent semigroups I. *Proceedings of the Japan Academy*, 33(10), 642-645.
11. Kimura, N. and Yamada, M. (1958). Note on Idempotent Semigroups II. *Proceedings of the Japan Academy*, 34, 110-112.
12. Menezes, A., Oorschot, P.C.V. and Vanstone, S.A. (1996). *Handbook of Applied Cryptography*, CRC Press.
13. Mushtaq, Q. and Yusuf, S.M. (1978). On LA-semigroups. Aligarh Bull. Math, 8, 65-70.
14. Mushtaq, Q. (1983). Abelian groups defined by LA-semigroups. *Studia Scient. Math. Hungar*, 18, 427-428.
15. Seguier, S. (1904). *Elements De La Theorie Des Groupes Abstraits*, (Elements of the Theory of Abstract Groups).

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 61-67

APPLICATION OF ARIMA ON INFLATION RATE IN PAKISTAN

Zahid Iqbal and **Abdul Rehman Malik**
Department of Statistics, Allama Iqbal Open University
Islamabad, Pakistan.
Email: zahid.iqbal@aiou.edu.pk

ABSTRACT

Major purposes of present study were to apply ARIMA model for forecasting inflation rate in Pakistan. Annual data from 1970 to 2017 has been analyzed. Data were taken from World Development Indicator (WDI). In current study, OLS method was applied for obtaining the purposes of estimation. For forecasting inflation rate in Pakistan, different diagnostic and selection criteria were used for preferring optimal model. Based on lowest AIC, SC and HQC, the study found ARIMA (2, 0, 2) model was most suitable model. High R-square and Log likelihood in value also showed that the model is best fitted.

1. INTRODUCTION

Inflation is a monetary phenomenon. The high inflation rate has become a serious alarm in the industrial and rising market economic all over the world. Generally, inflation is an extensive measure i.e. the overall rising in prices or the rising in the cost of living in a state.

Webster (2000) defined, "inflation is the persistent increase in the level of consumer prices or a persistent decline in the purchasing power of money". McMahon (2007) defined, "an increase in the price a person pays for goods".

Various ways are used to analyze time series data. The methodology of Box and Jenkins is repeatedly used on the basis of ARIMA model. Though the effectiveness of ARIMA model to study time series, it implies only to one variable and it does not describe some turning points in the data and also could not express relations among variables well in the system. The ARIMA model divides time series pattern into three parts: AR part, differencing part, and MA part. This time series technique uses past value to examine its own trend and forecast upcoming cycle. Hence the requirement for a statistical model that will capture exterior disturbances that impact on the model. Major purpose of present study is to forecast inflation rate in Pakistan via ARIMA.

2. LITERATURE REVIEW

Various statistical techniques have been used for the analysis of inflation dynamics.

Islam (2017) studied about forecasting Bangladesh's inflation through econometric models and ARIMA model was observed. The outcome proved that ARIMA(1,0,0) is the most perfect model to forecast inflation for next five years having predicted an inflation rate of 4.40 per cent in 2016 with a slightly raised price in the next consecutive years.

Osuolale et al. (2017) studied about Time Series Analysis to Model and Forecast Inflation Rate in Nigeria. Annual data was used for analysis from 2006 to 2015. They discovered that ARIMA (0, 1, 1) appears appropriate model for predicting inflation rates after three years which declared a similar movement from January, 2016 to December, 2018.

Abdulrahman, Ahmad and Abdellah (2018) studied about inflation rates in Sudan by utilizing ARIMA model. They were analyzed annually (1970-2016). The outcomes indicated that, there is convergence between forecasted value and original values for the duration of 1970 to 2016. Therefore, the inflation rates in Sudan will up in the upcoming years from 2017 to 2026.

Adubisi et al. (2018) studied about a predictive autoregressive integrated moving average (ARIMA) model for predicting inflation rate. ARIMA(1, 2, 1) model was found to be the best one on the basis of lowest AIC, AICc and BIC criteria. The forecasted value divulged that there is a forecasted value increase in the inflation rate in Nigeria, following the 95 per cent confidence interval.

Delima and Lumintac (2019) studied an appropriate ARIMA model in forecasting Philippines' inflation rate for the years 2018 to 2022 using data 1960-2017. In selecting the best model, ACF and PACF plotted and unit-root test for identification were observed and selection of the model according lowest AIC. After observing the lowest AIC, the finest model is ARIMA(1, 0, 0). So, error techniques were applied for forecasting that may be RMSE, MAE, and MAPE.

Mia, Nabeen and Akthar (2019) studied about CPI in Bangladesh. Annual data from 1986 to 2018 were used. Different ARIMA models were executed in this study. To find the best ARIMA model AIC, AICc and BIC were used. ARIMA(2, 2, 0) was the parsimonious model on the basis of lowest AIC, AICc and BIC. Forecasted CPI was based on the selected model ARIMA(2, 2, 0). The result showed that CPI is to continue an upward trend with respect to time.

Nyoni (2019) studied on ARIMA (Autoregressive Integrated Moving Average) modeling and forecasting of inflation in Egypt from 1960 to 2017. Annual data was used on inflation in Egypt. The diagnostic tests more imply that the presented best ARIMA (0, 1, 1) model is established and believable for forecasting inflation in Egypt. The study planned to predict inflation in Egypt for further ten years from 2018 to 2027 and the finest model was carefully choose based upon the least AIC value. The ARIMA (0, 1, 1) model is most suitable model to predict inflation in Egypt for future ten years.

Nyoni (2019) studied about predicting inflation in Sri Lanka using ARMA model. Annual data were used from 1960 to 2017. The purpose of present framework was to select best ARMA model for modeling and predicting Sri Lankan's inflation. Finest ARMA(1, 0, 0) model was selected on the basis of lowest AIC. The outcomes of the study showed that inflation of Sri Lanka will be approximately 8.17 per cent by 2020.

By covering inflation rate, a lot of statistical tools are used. In current study Auto Regressive Integrated Moving Average (ARIMA) model are used.

3. MATERIALS AND METHODS

Annual data were taken from World Development Indicator (WDI) period from 1970 to 2017. ARIMA model was applied for forecasting by Box-Jenkins methodology that approach mingles two different procedures into an equation: the first is an AR, and the second is an MA. ARIMA modeling shows that there is a relation in between a time series data and its own lagged. Commonly, the model is moved to ARIMA (p, d, q) model, where p, d and q ≥ 0 and submit to the order of autoregressive integrated and moving average features. To generate an ARIMA model, we started by an econometric equation with no independent variables $P_t = \alpha_o + \varepsilon_t$ and both AR and MA procedure was added.

$$P_t = \alpha_o + \alpha_1 P_{t-1} + \cdots + \alpha_p P_{t-p} - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q} + \varepsilon_t$$

where α's and β's are the coefficients of the autoregressive (AR) and moving average procedures, respectively.

Methodology of the Box-Jenkins involves five steps for identification, selection and assessing conditional means models

4. COMMENTS AND CONCLUSION

Basic outcomes of inflation show unequal mean and median of inflation. Skewness is far from zero. Kurtosis is not close to 3. In addition to significant JB probability suggests non-normal data. So, log transformation is applied to convert non normal series into normal which shows both mean and median are approximately equal. Skewness is close to zero and kurtosis is near to 3. Moreover, insignificant JB probability shows normality.

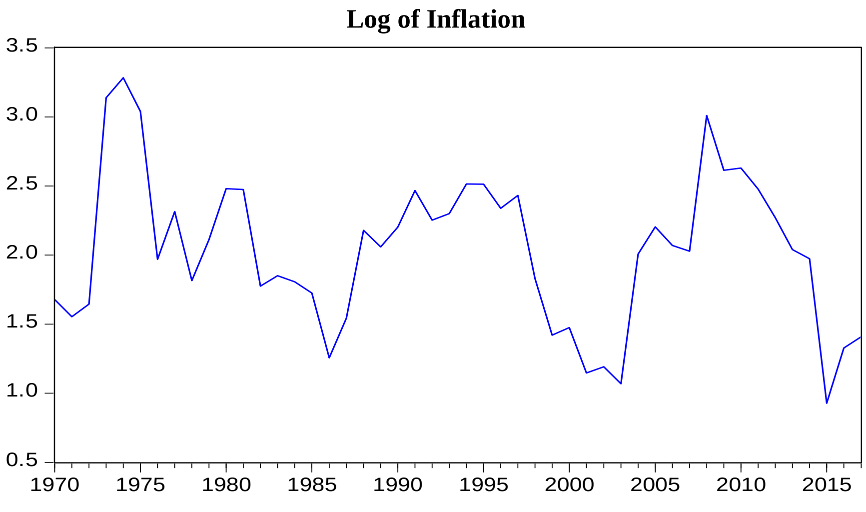


Figure 4.1 Time Plot of Inflation with log at Level

Figure 4.1 indicates a random pattern which means the series are stationary. Now, unit root test was employed to ensure that series is stationary or not.

4.1 Unit Root Test

A stationary test is required for specifying and estimating the best fitted model. Augmented Dickey Fuller (ADF) test is used to ensure the stationary problem.

Table 4.1
Outcomes of Augmented Dickey Fuller Test

Variable	Level	
	Intercept	Intercept & Trend
LOGINF	-2.978706*	-3.126517

* and ** show significant at 1% and 5% respectively

Table 4.1 shows that log of inflation (LOGINF) became stationary at 1%.

4.2 ARIMA Model Identification

For model identification different models are executed as under:

Table 4.2
Outcomes of ARIMA Models for Forecasting Inflation

Model	R^2	$\bar{R}^2$	LL	AIC	SC	HQC
ARIMA (2, 0, 0)	0.4348	0.4086	-24.489	1.1952	1.3145	1.2399
ARIMA (2, 0, 1)	0.5677	0.5368	-18.325	0.9707	1.1297	1.0303
ARIMA (2, 0, 2)	**0.6172**	**0.5798**	**-15.531**	**0.8927**	**1.0914**	**0.9671**
ARIMA (1, 0, 2)	0.5402	0.5081	-20.087	1.025	1.1825	1.0843
ARIMA (0, 0, 2)	0.4504	0.426	-24.513	1.1464	1.2634	1.1906
ARIMA (3, 0, 0)	0.445	0.4044	-23.757	1.2337	1.3943	1.2936
ARIMA (3, 0, 1)	0.5431	0.4974	-19.384	1.0838	1.2845	1.1586
ARIMA (3, 0, 2)	0.571	0.516	-17.966	1.0652	1.3061	1.155
ARIMA (3, 0, 3)	0.5719	0.5042	-17.921	1.1076	1.3886	1.2124

In choosing optimal ARIMA model, above outcomes confirmed that ARIMA (2, 0, 2) model is preferred because its AIC, SC and HQC are least and R-square, adjusted R-square and Log likelihood is most.

4.3 ARIMA(2, 0, 2) Estimation and Interpretation

After identification of ARIMA model, further was to find the coefficient of parameters. The parameters of the model were estimated as under:

Table 4.3
Outcomes of Best fitted ARIMA (2, 0, 2) Model.

Variable	Coefficient	Std. Error	t-Statistic
C	2.0035	0.0142	142.1272
AR(1)	1.7390	0.0814	21.3571
AR(2)	-0.8876	0.0780	-11.3741
MA(1)	-1.4782	0.1735	-8.5176
MA(2)	0.4782	0.1687	2.8352
R-squared	0.6172	LL	-15.5316
Adjusted R-squared	0.5798	AIC	0.8927
SC	1.0914	HQC	0.9671

Table 4.3 shows the coefficients of AR (1 to 2) and MA (1 to 2) that were extremely significant at 1%. The AIC, SC and HQC were lesser in values from all other evaluated models. So, forecasting of ARIMA (2, 0, 2) model is preferred. The value of R-squared was 0.6172 i.e. 62% of the variation in inflation in Pakistan is explained by past values and the past errors. So, equation of the model can be written as under:

$$(loginf)_t = (2.0035) + (1.7390)(loginf)_{t-1} + (-0.8876)(loginf)_{t-2} \\ + (-1.4782)e_{t-1} + (0.4782)e_{t-2}$$

4.4 ARIMA (2, 0, 2) Diagnostic Tests

After estimating the parameters, it was essential to observe the model adequacy by testing. The ARIMA(2, 0, 2) was tested for serial correlation, heteroskedasticity and normality.

Table 4.4
Diagnostic Tests of Residuals of ARIMA (2, 0, 2) Model

Test	Statistic Value	p-value
Serial Correlation LM Test Breusch-Godfry: (Obs. R-square)	1.4503	0.4842
Heteroskedasticity Test Breusch-Pagan-Godfrey: (Obs. R-square)	7.7840	0.1685
Normality Test (Jarque-Bera)	4.8270	0.0895

Table 4.4 suggests that the model was not serially correlated and there is no hetroskedastic problem in residuals. The Jarque-Bera test shows normality.

Table 4.5
In-sample and Out-sample forecast of Pakistan's inflation 2011-2030

Years	Actual Value	Predicted Value	95% Confidence Lower Bound	95% Confidence Upper Bound
2011	11.92	12.22	5.26	24.52
2012	9.68	10.13	4.36	20.33
2013	7.69	8.28	3.57	16.61
2014	7.19	6.89	2.97	13.81
2015	2.53	6.43	2.77	12.9
2016	3.77	3.89	1.68	7.8
2017	4.09	5.00	2.15	10.02
2018		6.06	2.61	12.15
2019		8.16	3.19	17.49
2020		10.32	3.98	22.32
2021		12.09	4.64	26.19
2022		13.08	4.85	28.97
2023		13.13	4.55	30.38
2024		12.39	3.97	30.07
2025		11.19	3.37	28.18
2026		9.89	2.87	25.42
2027		8.75	2.5	22.67
2028		7.91	2.26	20.52
2029		7.41	2.12	19.23
2030		7.22	2.05	18.82

In present study, an ARIMA (2, 0, 2) model has been selected for forecasting inflation rate in Pakistan which consists of 48 values from 1970 to 2017. The time plot of inflation and ADF test connoted stationarity at level with log. The estimated parameters of ARIMA (2, 0, 2) found significant which ensure that model is best fitted on the other hand serially uncorrelated, free from heteroskedastic problem and residuals are normal. This model is most suitable to forecast inflation rate in Pakistan.

REFERENCES

1. Abdulrahman, B.M.A., Ahmed, A.Y.A. and Abdellah, A.E.Y. (2018). Forecasting of Sudan Inflation Rates using ARIMA Model. *International Journal of Economics and Financial Issues, 8*(3), 17-22.
2. Adubisi, O.D., David, I.J., James, F.E., Awa, U.E. and Terna, A.J. (2018). A predictive Autoregressive Integrated Moving Average (ARIMA) Model for Forecasting Inflation Rates. *Research Journal of Business and Economic Management*, 1(1), 1-8.
3. Box, G.E.P. and Jenkins, G.M. (1976). *Time series analysis: forecasting and control*. San Francisco: Holden-Day.
4. Delima, A.J.P. and Lumintac, M.T.Q. (2019). Application of time series analysis for Philippines' inflation prediction. *International Journal Recent Technology and Engineering*, 8(1), 1761-1765.
5. Islam, N. (2017). Forecasting Bangladesh Inflation through Econometric Models, *American Journal of Economics and Bussiness Administration*, 9(3), 56-60.
6. McMahon, G. (2007). Measuring short-run inflation for Central Bankers. *Federal Reserve Bank of St. Louis Review*, 79, 3-8.
7. Mia, M.S., Nabeen, A.H.M.M.R. and Akhter, M.M. (2019). Modelling and Forecasting the Consumer Price Index in Bangladesh through Econometric Models. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS).* 59(1), 118-127.
8. Nyoni, T. (2019). ARIMA modeling and forecasting of inflation in Egypt (1960-2017). *Munich Personal RePEc Archive (MPRA), Paper No. 92446*, 1-9.
9. Nyoni, T. (2019). Predicting inflation in Sri Lanka using ARMA model. *Munich Personal RePEc Archive (MPRA), Paper No. 92432*, 1-9.
10. Osuolale, P.P., Ayanniyi, W.A., Adesina, A.R. and Matthew T.O. (2017). Time series analysis to model and forecast inflation rate in Nigeria. *Anale. Seria Informatica*, 15(1), 174-178.
11. Webster, D. (2000). *New Universal Unabridged Dictionary*, Barnes and Noble Books.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 69-82

'OPEN DATA' AND ITS IMPACT ON ACHIEVING SUSTAINABILITY DEVELOPMENT GOALS IN PAKISTAN

Taha Shabbir and **Umair Ansari**
University of Karachi, Karachi, Pakistan
Email: tahashabbir51@gmail.com

ABSTRACT

Any country of the world needs to undergo developmental projects to prosper greatly. They must ensure that their people are living in an excellent socio-political environment that could grant them safety. However, it varies from region to region and from place to place. Pakistan as a country is a low economy that cannot give its people the type of security they need. An open data system is one of these crucial developments initiated within government setup, which provides enormous access to databases. It improves learning and knowledge and allows them to think 'out of the box.' It also brings them the ability to polish their learning skills and take the concession to take things out of their resources and utilize them fully. Although there is some compulsion to use private data sources, we must know that not all data are confidential. Hence, this type of open data is integral for the people to get input from and use it accordingly. The research looks into the pros and cons of using available data systems to measure its impacts on Pakistan's sustainability development goals. The use of an open data system helps people perform their daily life tasks; it is important to evade corruption. In other words, these available systems foster great accountability of national institutions. So, there is a feeble chance of witnessing the element of corruption in the society that could cause chaos. Open data systems are essential for creating rapid growth of innovation and creativity for developing countries like Pakistan. With the proper access to 'Open Data,' Pakistan can quickly achieve SDGs within the given timeframe.

KEYWORDS

Open Data, Pakistan, SDG's, COVID-19, Data Analysis, Development.

1. INTRODUCTION

The progress of every society is highly contingent on the execution of construction programs. Developmental programs are critical for empowering communities and maintaining a good social approach that supports justice and protection. Pakistan, as a developing nation, is working hard to provide stability to its people. Various techniques assist the government in achieving this aim, one of which is the usage of the open data system (ODS), which was launched by the Pakistani government by enabling access to databases. This study explores the benefits and drawbacks of using ODS to affect Pakistan's social growth. The usage of ODS has many advantages, including making everyday life activities simpler for the general public in Pakistan, as well as ensuring the eradication of corruption by promoting greater transparency within national institutions. People (students) may use open data structures to find the information they need about study topics. Everyone has quick access to the details, which allows it more efficient to use the results. Various regimes around the world promote the usage of transparent data to gain accountability and improved administration of affairs. Companies who follow an accessible data approach would have more visibility to the media and will be willing to discuss their success goals with others who are involved (Attard, 2016). The value and advantages of utilizing this system have long been accepted by developing countries. In the United States, President Barack Obama effectively launched this drive through an openness and accountability campaign. The primary goal of using this policy was to guarantee that public institutions operated efficiently. After that, the government of the United Kingdom introduced it, encouraging people to engage in the country's social growth (Alcock, 2014). The primary motivation for adopting the ODS was to make clear the obligation of civil departments toward their constituents and enable them to participate in the country's social progress. Apart from supporting social growth, it also provided citizens with social empowerment. Governments mostly used it to ensure timely delivery of citizens' rights and to keep them informed regarding the usage of their money and the programs to which their funds were allocated. The government of the United Kingdom adopted this policy with the same aim in mind (Judge, 2005). In view of the foregoing instances, it is argued that the implementation of such systems could provide strength to Pakistan's governments, enabling them to function more effectively and efficiently with greater accountability, as well as build a positive reputation in the minds of Pakistanis.

Research Questions

Following research questions has been devised in order to find the impact of ‘Open Data’ on acheiving SDG`s in Pakistan:

1. How will Open Data System benefit Pakistan's social development points discussed in SDG`s?
2. How can ODS aid in the formulation of people-friendly policies?
3. What part will open data systems play in ensuring transparency and good governance in Pakistan?

Research Objectives

It is preferable to analyze the following goals in order to understand how the researcher would perform this study:

- To identify explanations, evaluate the meaning and need for the words "Open Data."
- Assess the value of applying a case study to the social sector.
- Examine the consequences of ‘Open Data’ on SDG`s.
- Establish a plan that illustrates the effect of ODS on the country's social growth with aspet to SDG`s.
- An ‘Open Data' exposure can be a crucial asset for every nation to utilize in achieving its development goals.

Research Hypothesis

H1: Using accessible data, there is a constructive partnership between social and political growth in Pakistan with aspect to SDG`s.

H2: Using an accessible data scheme, a nation's (Pakistan's) economic growth can be enhanced.

2. LITERATURE REVIEW

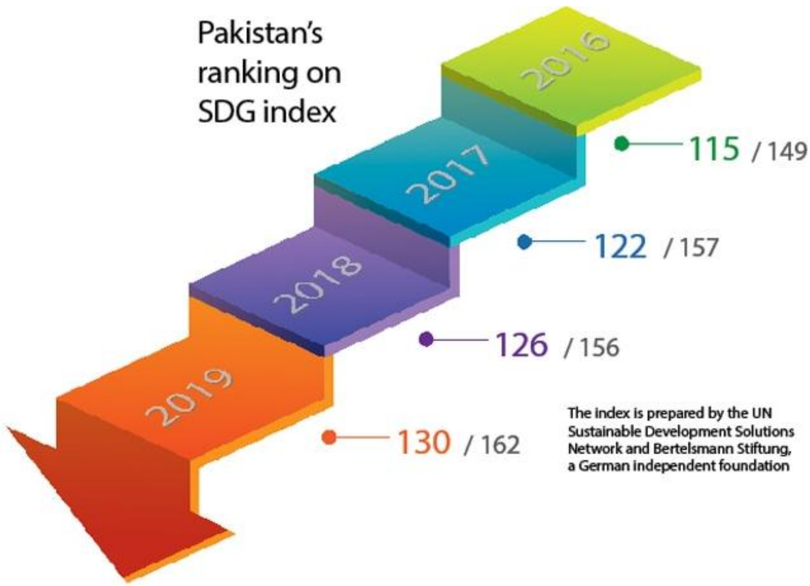
According to the literature, the definition of open data may be described as data that is freely accessible for public usage, subject to certain limitations such as copyright, patents, or other control mechanisms. The word "accessible data" relates to the notion that data should be publicly usable for everyone to use and republish if required and wherever they choose, with no constraints. And there is currently little literature accessible on each country's "Open Data" policies. As a consequence of this void, mechanisms for improving a country's socioeconomic growth are being created. It is important to remember that open data must consider exchanging knowledge relevant to companies that is required by clients of organizations and government departments. In addition, transparent data programs must provide information that is regularly accessed, such as regional details, forecast information, development reports, information on the amount of tourists each day, and information on the organization's functions. In March 2011, the Finnish government emphasized the importance of developing a database through the use of accessible data by making it a large part of the government's program. In 2012, government agencies begin exchanging their details through open data, inspired by the government's campaign and with priorities such as accountability and greater outreach in mind. The National Land Survey of Finland, for example, began publishing advanced topographic data in May 2012. In a similar way, the Finnish Meteorological Institute began exchanging climate data publicly a year later. The key aim is to render all significant knowledge gathered and processed by the transparent organisation available and functional through the internet and computers by 2020, at no out-of-pocket charges, and the details can be exchanged under specific terms of usage. The United Kingdom was the first nation in Europe to expose its data to the public, and it has reaped a slew of benefits in terms of social and economic growth.

3. METHODOLOGY

The study was conducted in an inductive reasoning manner. The association between 'Open Data' and the country's socio-political growth with aspect to all 17 points of SDG`s is studied. The hypotheses were tested for affirmation or dismissal using the correlation approach to quantify variables. Since all the variables could quantify the same definition and therefore be highly correlated, construct validity and convergent validity were essential for this reason. We used factor analyses to evaluate this association. The relevance of the instrument is checked here by ensuring that the questionnaire includes questions that are applicable to the concept being tested. In quantitative methods, the validity of an instrument is more concerned with the analytical nature of its function. The goals are simple, and the build validity and checking methodology can be used to substantiate the instrument's validity as well. Data was obtained using both qualitative and quantitative approaches. Data was gathered from four groups of respondents that were in favour of utilizing open data structures, including companies, media people, the general public, and non-governmental organizations, through interviews and survey polling. Furthermore, predictive research was carried out using a variety of methods, including correlation, regression analysis, and factor analysis. The outcomes findings revealed a connection between the usage of ODS and positive outcomes in Pakistan's social sectors. To create a connection between the independent and dependent variables, Pearson's correlation was used. In addition, regression analysis was used to investigate the impact of independent variables on dependent variables. The descriptive statistics assisted in the retrieval of valuable conclusions from the analysis. The preparation and maturity of Open Data are two main metrics that are often used to assess its maturity. The national Open Data platform would include these characteristics. The "readiness" of an accessible data repository guarantees the presence of open data policy for data consumers, as well as the fiscal, social, and political implications of open data. Whereas "maturity" refers to the use of a web-based portal and its features, such as the total use of data and the spread of data.

4. RESULTS

According to the findings, Pakistan's Social Sector has completed just 44% of the path to fully embracing Open Data structures. Adoption of full ODS will necessitate the development of a transparent Open Data Policy that resolved licensing concerns. It therefore necessitates the development of national standards to assist local councils in learning from and adjusting to current structures. Full implementation of open data structures will necessitate a clearer comprehension of the expected effect of open data. Aside from that, it necessitates effective traffic control towards open data structures. The site can have features that cover a variety of data topics and include a data download option.



DOI: https://www.dawn.com/news/1506867

5. DISCUSSION

The term "Open Data" does not begin with the word "data." Instead, it begins with a challenge. Applying data science to the tradition of ‘Open Data' necessitates more than simply presenting meaning and seeking clarification in large volumes of unstructured data: it can even necessitate media organizations rethinking how they function and how they revere. Understanding the influence of ODS in Pakistan necessitates an investigation of the past of such programs. Initiation of such programs occurred in the years prior to 2010, and prompted the nation to create numerous datasets as part of its structured open data framework. Many difficulties were encountered along the way to implementing open systems, but they were eventually implemented due to their undeniable benefits in managing government affairs. Western countries led the way in implementing certain schemes to improve government efficiency. With the passing of time and the usefulness that these programs offer, eastern countries decided to follow them as well. This is because policymakers encounter almost identical issues when communicating with the general population and consumer organizations that need accurate and reliable input from government agencies. Because of their increased importance, countries are increasingly opting for such systems so that governments can be more transparent about their operations and responsibilities to their citizens. This aids governments in reducing corruption and allowing more direct and honest facts to reach the general public. As a result, it is often advised that these frameworks be used to ensure that systems are critical for a country's success and to improve its social and political setup. Furthermore, by reducing wasting of capital such as financial and other resources, GDP and economic development may be improved. Sharing government data would almost certainly lead to the creation of new government structures.

The usage of transparent systems is critical for citizens, government agencies, and general institutions working under the supervision of the government to stay linked and openly sharing knowledge. This research is noteworthy because it emphasizes the importance of open data to Pakistan's economy. To gain productivity and effectiveness whilst meeting high targets, the country must improve system accountability and openness, which can only be accomplished by implementing and efficiently utilizing open data programs. Pakistan will combat corruption by incorporating accessible data and involving citizens in social and cultural growth.

‘Open Data’ and SDG`s in Pakistan.

The 2030 Plan for Sustainable Development was adopted by UN Member States on September 25, 2015, which contains a package of 17 Sustainable Development Goals (SDGs) to eliminate hunger, battle inequalities and discrimination, and address climate change by 2030.

The Sustainable Development Goals (SDGs), commonly known as the General Goals, are a global call to action to eradicate hunger, preserve the climate, and ensure stability and prosperity for all citizens. These 17 Goals draw on the Millennium Development Goals (MDGs), while still addressing new issues including climate change, income disparity, innovation, healthy consumption, stability, and fairness. The objectives are intertwined, and achieving one frequently necessitates addressing problems that are more generally identified with another.

The SDGs act in a spirit of collaboration and pragmatism to make the best decisions now in order to better life for future generations in a sustainable manner. They laid out specific recommendations and goals for all countries to meet, focused on their own objectives and the global environmental issues. The SDGs are a non-discriminatory agenda. They address the root causes of injustice and get us together to create a real difference between citizens and the environment. The determination to leaving no one behind is at the forefront of the 2030 Agenda, as is poverty eradication.

The SDGs are a bold promise to finish what we began to address some of the world's most urgent issues. All 17 goals are interconnected, because progress with one has an effect on the others. Dealing with the challenge of climate change has an effect on how we handle our finite environmental capital, promoting gender equity or greater wellbeing aids in poverty eradication, and encouraging unity and egalitarian communities reduces inequality and aids economic growth. At a nutshell, this is the biggest way we have to better the lives of future people.

Source: UNO Website

Pakistan Sustainable Development Goals

Pakistan has demonstrated exemplary dedication to the 2030 Plan for Sustainable Growth, having been one of the first countries to sign up to it in 2015. The Sustainable Development Goals (SDGs) were overwhelmingly adopted as the national development plan by Parliament on February 16, 2016. It has founded the Parliamentary SDGs Secretariat, which is one of only a few countries to do so. This legislative process was the first and most important phase in mainstreaming and localizing the SDGs. In Pakistan, localizing the SDGs has entailed more than just committing to structured agreements with government partners. It's also been a collective mechanism that focuses on evidence-based strategies to transform abstract SDG ambitions into clear, meaningful priorities for all stakeholders, including local actors. Several public duties, including the distribution of social care, were devolved from the federal government to the local governments as part of the 18th constitutional amendment. Pakistan's government held consultations with all stakeholders regarding the post-Millennium Development Goals (MDGs) in order to organize and enhance initiatives at the federal and provincial levels to meet Pakistan's long-term development and poverty reduction goals. The need for national categorization of SDGs, increased data collection, and implementation of reporting systems was highlighted during the consultation period. The SDGs was fully compatible with Vision-2025's seven pillars, culminating in a detailed long-term plan for promoting equitable growth and sustainable development.

Open Data and Improved Transparency

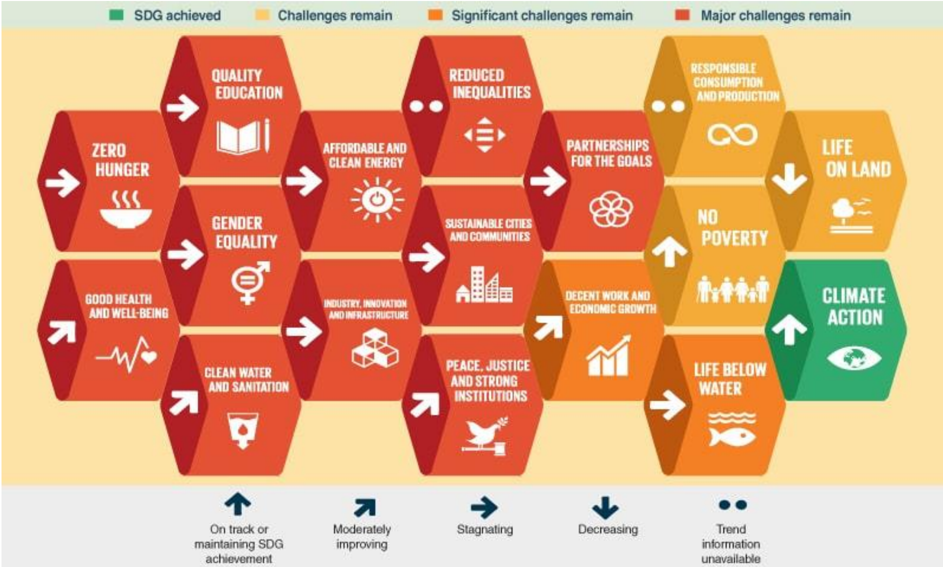
Having knowledge accessible in real time and that all the period will help to build a consistent structure. As the British government has achieved by linking taxpayers and collectors, this could be the focal point for improving government services. This offers the government an advantage of knowing of any of the taxes raised, as there is no risk of tax revenues being lost. Tax revenue is paid immediately into the government's banks, where it can be put to good use for infrastructure programs. This structure provides more openness, and the community is thereby referred to as an accessible society, in which knowledge is openly exchanged. As we will see in Europe, this helps governments become more transparent, and potential risks to wrongdoing can be prevented. The study by Ahonen-Rainio et al. (2014) revealed a number of open data framework observations, but it overlooked the contributions to the country in Euro terms. According to Koski (2011), the implementation of spatial knowledge was aligned with profoundly proving to be the quicker means of growth in fifteen countries. Koski's (2012) investigation finds that the age of modern information-driven management has arrived in Finnish organisations, based on responses from 531 organizations. The gains greatly outweigh the expenses, and administrative changes should be anticipated in a number of organisations and regions. According to Kiuru et al. (2012), small companies will profit easily from newly implemented accessible knowledge systems by different government administrations.

https://www.dawn.com/news/1506867

Open Data Impacting Lives of People

According to Worthy (2018), accessible data programs have three main impacts on people's lives:

1. The government is more complicit than it has ever been due to the use of such systems. It improves the government's organization and performance. It can become competent as a result of providing high-quality public services (Stefaan & Andrew, n.a.).
2. These systems encourage people by actively impacting peoples' lives.
3. Policymakers should provide advanced economic opportunities to citizens in cities and/or rural regions, allowing new businesses and sectors to emerge.



DOI: https://www.dawn.com/news/1580721

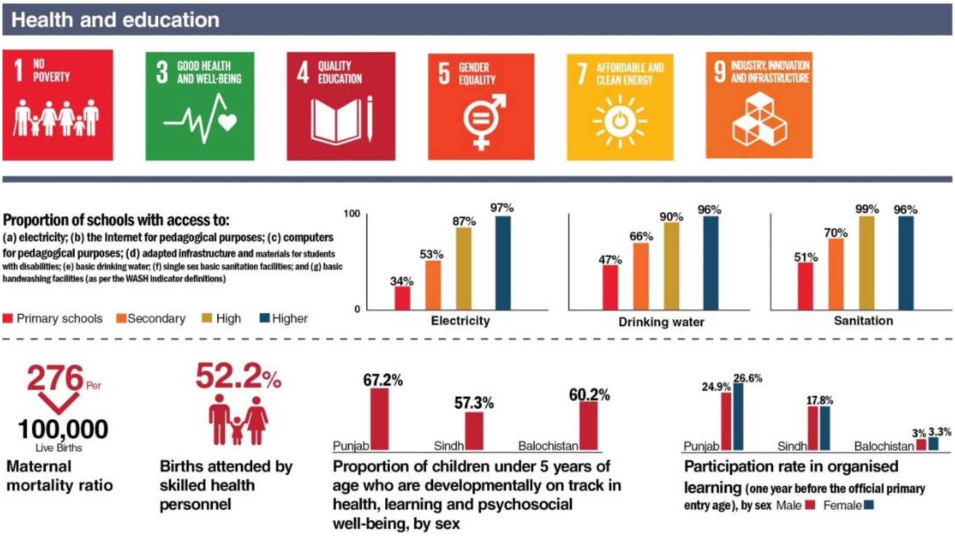
What influence Occurred in Pakistani Society?

Pakistan, as a developing nation, faces numerous social issues. Gender disparity, women's empowerment, child (especially girl) schooling, malnutrition, inflation, corruption, economic unrest, law and order problems below the belt, and border insecurity are only a few instances. Pakistan, like the United Kingdom, will easily address many of these societal problems and challenges by implementing an "Open Data" scheme. For Pakistan's growth, the model of the UK's "Open Data" framework is highly successful.

1. **Corruption:** The most important and real benefit of ‘Open Data' is that it aids in the prevention of corruption. The method utilizes the capacity of real-time transparency to keep authorities responsible by tracing any corruption using public records. With the passage to the Pakistani parliament's right of information act law in 2016, the government has restricted the availability of all public records. Today, corruption is without a doubt the greatest challenge to the country's economic stability. The National Oversight Bureau (NAB) and the Federal Investigation Agency (FIA) both have a cyber-branch.

2. **Terrorism:** The worst of all ills in Pakistan, along with injustice, is the threat of terrorism, which needs new horizons to tackle. Open data aids in the monitoring of terrorist acts and allows officials to respond quickly to any signs of them. Since 9/11, when Pakistan became an anti-NATO partner in the war on terror, the nation has become the victim of several militant attacks, with millions of lives lost. The country's thin tank has devised a range of counter-terrorism techniques, one of which is the usage of the "Free Data" framework.

3. **Health and Education:** When it comes to health and education, Pakistan has still lagged behind in terms of figures and quality. Accessible data can aid provide real-time statistics on the number of schools and students, as well as the location and efficiency of physicians at every hospital. According to 2017 statistics on gender disparities in education, the gap between boys and girls in education from 5th to 12th grade is around 10%. (Nine years to 17 years old respectively). It is critical to notice that the percentage of boys and girls who drop out of government public schools is high, especially at the primary level. Inadequate schooling services accessible at the post-primary and high school levels are one of the main causes of dropout. In comparison to neighboring countries, the literacy rate is also poor. Bangladesh, which gained independence from Pakistan in 1971, has a higher literacy rate. The usage of open data will assist in the battle against illiteracy in Pakistan.



DOI: https://www.dawn.com/news/1360165

4. **Cyber Security & Internet Access:** As social media's dominance rises, a slew of new threats arise. The new hybrid war between India and Pakistan, also known as the fifth generation war, has increased the influence and constructive usage of social media. Cybercrime, such as using the Internet to stalk or intimidate others, or stealing someone's data without their permission, may be conveniently handled with the aid of "Open Data." The effect of the United Kingdom demonstrates that this can be accomplished in Pakistan as well. The internet has helped the general public to a large degree by encouraging them to exchange information through computers, digital devices, and networks. The cyber age is the name given to this period. Around the same period, cyber-crime has escalated leading to the hacking of personal and corporate data all over the planet. The crimes are not restricted to records; they have spanned the spectrum from information to stealing millions of rupees from online banks, among other things. Spreading viruses, cracking documents, and publishing defamatory comments are all examples of cybercrime.

Social engineering is a popular tactic used by cyber criminals to gain access to confidential details. The attackers use deception and persuasion to persuade users to divulge sensitive details. Another method of accessing computer-based knowledge is data theft. An individual's privacy is breached, and his identifying knowledge is collected. It is becoming more prevalent by the day, even large corporations are experiencing the same issues. Individuals and organisations are harmed by a lengthy list of data theft practices. The first is identity theft, which is a felony in which identification is collected with intent and then used for frauds or tricks by utilizing a false or fabricated identity. The primary aim of such events is to make money. Another example occurs where cybercriminals acquire access to an individual's e-mail to steal sensitive details in order to collect purchases. Such offences can be combated by using ‘Open Data' schemes, as the UK has effectively done.

6. CONCLUSION

Limitations and Future Direction

What Are Pakistan's 'Free Data' Barriers? The biggest barrier to utilizing ‘Open Data' in Pakistan is a lack of knowledge of what it is. There's also the government's casual approach toward some constructive use of technology. Today, all developing countries are using innovative technology technologies to tackle social ills. In terms of utilizing technology, Pakistan also lags behind other developed countries. Without a question, Pakistan has many brilliant minds who excel in technology, but there is little incentive or funding for them at the government level. Nonetheless, more positive improvements will be achieved once investigative journalists understand how to utilize "Free Data." The government, for example, is affected by accessible data structures to ensure that knowledge is usable in a consistent way for social and political changes. The government should use the following options:

1. Increasing public understanding about new technology.
2. Disseminating knowledge to improve civic involvement.
3. Improving ODS in the public and government industries.
4. A general system for valuation should be established in order to value people's recommendations.
5. Doing experimental analysis to determine whether or not a system is innovative.

REFERENCES

1. Kolucki, B. and Lemish, D. (2011). *Communicating with Children*, UNICEF, New York, ISBN: 978-0-578-09512-7.
2. Lugavere L. and Grewal P. (2018). *Genius Foods*, HarperCollins Publishers, Sydney.
3. Fabi, B., Lacoursière, R. and Raymond, L. (2015). Impact of high-performance work systems on job satisfaction, organizational commitment, and intention to quit in Canadian organizations. *International Journal of Manpower*, 36(5), 772-790,
4. Luna-Arocas, R. and Camps, J. (2007). A model of high performance work practices and turnover intentions. *Personnel Review*, 37(1), 26-46,

5. Chen, S.Y., Uen, J.F. and Chen, C.C. (2016). Implementing high performance HR practices in Asia: HR practice consistency, employee roles, and performance. *Asia Pacific Journal of Management*, 33(4), 937-958.
6. Chadwick, C., Super, J.F. and Kwon, K. (2015). Resource orchestration in practice: CEO emphasis on SHRM, commitment-based HR systems, and firm performance. *Strategic Management Journal*, 36(3), 360-376.
7. Maharjan, M.P. and Sekiguchi, T. (2016). The influence and effectiveness of US-style and Japanese-style HR practices on Indian firms: A conceptual perspective. *South Asian Journal of Human Resources Management*, 3(1), 58-74.
8. Kim, D. and Choi, Y. (2014). Social exchange model between human resource management practices and innovation in software engineering. *Seoul Journal of Business*, 20(2), 46-69.
9. Bamber, G.J., Bartram, T. and Stanton, P. (2017). HRM and workplace innovations: formulating research questions. *Personnel Review*, 46(7), 1216-1227.
10. Anitha J. (2014). Determinants of employee engagement and their impact on employee performance. *International Journal of Productivity and Performance Management*, 63(3), 308-323.
11. Pawirosumarto, S., Sarjana, P.K. and Muchtar, M. (2017). Factors affecting employee performance of PT. Kiyokuni Indonesia. *International Journal of Law and Management*, 59(4), 602-614.
12. Rivai, V. (2004). *Manajemen Sumber Daya Manusia untuk Perusahaan Dari Teori ke Praktek*, Grafindo Persada, Jakarta.
13. Pandian, S.S. (2018). Impact of Fourth industrial revolution in human resource management. *International Journal for Research Trends and Innovation*, 3(2), 59-61.
14. Wilkesmann, M. and Wilkesmann, U. (2018). Industry 4.0 – organizing routines or innovations? *VINE Journal of Information and Knowledge Management Systems*, 48(2), 238-254.
15. Hanley, T., Ersahin, Z., Hebron, J., Prescott, J. and Sefi, A. (2019). House of Commons Science and Technology Committee Report: Impact of social media and screen-use on young people’s health, House of Commons, UK.
16. Ireland F.C. (2018), *Good Health or the Physiology of Dietetics and Massage*, Forgotten Books, USA, ISBN-10: 0267627165.
17. Jordan A.B. (2008). *Children and Electronic Media*, Princeton University, New Jersey, ISSN: 1550-1558, ISBN: 978-0-9814705-0-4.
18. Kidron, B. and Rudkin, A. (2017). *Digital Childhood: Addressing childhood development milestones in the digital environment*. 10.13140/RG.2.2.23405.08167.
19. Boselie, P., Dietz, G. and Boon, C. (2005). Commonalities and contradictions in HRM and performance research. *Human Resource Management Journal*, 15(3), 67-94.
20. Takeuchi, R., Lepak, D.P., Wang, H. and Takeuchi, K. (2007). An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology*, 92(4), 1069-1083.
21. Kim, H.J., Tavitiyaman, P. and Kim, W.G. (2009). The effect of management commitment to service on employee service behaviors: the mediating role of job satisfaction. *Journal of Hospitality and Tourism Research*, 33(3), 369-390.

22. Aladwan, K., Bhanugopan, R. and D'Netto, B. (2015). The effects of human resource management practices on employees' organisational commitment. *International Journal of Organizational Analysis*, 23(3), 472-492.
23. Sattar, T., Ahmad, K. and Hassan, S.M. (2015). Role of human resource practices in employee performance and job satisfaction with mediating effect of employee engagement. *Pakistan Economic and Social Review*, 53(1), 81-96.
24. Kundu, S.C., Kumar, S. and Gahlawat, N. (2019). Empowering leadership and job performance: mediating role of psychological empowerment. *Management Research Review*. doi:10.1108/mrr-04-2018-0183.
25. Awamleh, N.A. (2013). Enhancing Employees Performance via Empowerment: A Field Survey. *Asian Journal of Business Management*, 5(3), 313-319.
26. Al-Jammal, H., Al-Khasawne, A. and Hamadat, M. (2015). The Impact of the Delegation of Authority on Employees' Performance at Great Irbid Municipality: Case Study. *International Journal of Human Resource Studies*, 5(3), 48-68.
27. Hsu, C. and Sabherwal, R. (2012). Relationship between Intellectual Capital and Knowledge Management: An Empirical Investigation. *Decision Sciences*, 43(3), 489-524.
28. Kasasbeh, W. (2015). The Impact of Knowledge Management in the Development of a Culture of Excellence: An Empirical Study of the Commercial Banks Operating in the City of Tabuk. *Dirasat: Administrative Sciences*, 42(2), 267-287.
29. Bulger M. and Davison P. (2018). *The Promises, Challenges, and Futures of Media Literacy*, Data & Society Research Institute, New York.
30. Calvert, S.L., Jordan, A.B. and Cocking, R.R. (2002). *Children in the digital age: Influences of electronic media on development*. Westport, CT: Praeger.
31. Cramer S. and Inkster D.B. (2017). Status of Mind; Social media and young people's mental health and wellbeing, Royal Society for Public Health, London.
32. Gibney M.J., Lanham-New S.A., Cassidy A. and Vorster H.H. (2009). *Introduction to Human Nutrition* Second Edition, A John Wiley & Sons, Ltd. West Sussex, ISBN 978-1-4051-6807-6.
33. Gilutz, S. and Nielson, J. (2002). *Usability of websites for children: 70 design guidelines*. Nielsen Norman Group, Fremont, CA.
34. Girl Scout Research Institute (2002). *The net effect: Girls and new media*. Girl Scouts of the USA, New York.
35. Goodman R. and Scott S. (2012). *Child and Adolescent Psychiatry*, Third Edition, A John Wiley & Sons, Ltd. Oxford, ISBN 978-1-119-97968-5.
36. Pruneda, G. (2015). Implementation of high-performance work practices in the Spanish private sector: Stronger or weaker during the economic downturn? *Evidence-based HRM: a Global Forum for Empirical Scholarship*, 3(2), 159-180.
37. Murphy, K., Torres, E., Ingram, W. and Hutchinson, J. (2018). A review of high performance work practices (HPWPs) literature and recommendations for future research in the hospitality industry. *International Journal of Contemporary Hospitality Management*, 30(1), 365-388.
38. Arif, S. and Ilyas, M. (2013). Quality of work-life model for teachers of private universities in Pakistan. *Quality Assurance in Education*, 21(3), 282-298.

39. Hong, K., Tan, K. and Bujang, S. (2010). Relationship between Work Life Quality of Teachers with Work Commitments, Stress and Satisfaction: *Journal Teknologi*, 52, 1-15.
40. Liang, G. and Akiba, M. (2015). Characteristics of teacher incentive pay programs: a statewide district survey. *Journal of Educational Administration*, 53(6), 702-717.
41. Garen, J. (2016). Assessing the literature on school reform from an entrepreneurship perspective. *Journal of Entrepreneurship and Public Policy*, 5(3), 383-403.
42. Allen, R.S. and Kilmann, R.H. (2001). The role of the reward system for a total quality management based strategy. *Journal of Organizational Change Management*, 14(2), 110-131.
43. Liu, S., Zhao, D. and Xie, W. (2016). Chinese teachers' attitudes toward performance pay: the cases of three schools. *International Journal of Educational Management*, 30(6), 791-808.
44. Liu, H. and Li, G. (2017). To gain or not to lose? The effect of monetary reward on motivation and knowledge contribution. *Journal of Knowledge Management*, 21(2), 397-415.
45. Arkin, E.B. and Bauer, C. (2010). *Making Health Communication Programs Work SDG*, Department of Health & Human Services Public Health Service, National Institutes of Health National Cancer Institute, USA.
46. Alper, M. (2014). *Digital youth with disabilities*. Cambridge, MA: MIT Press, ISBN 978-0-262-52715-6.
47. Bayer E. and Thompson D. (2015). *Communicating to advance the public's health: Workshop summary*, The National Academies Press, Washington, DC, ISBN 978-0-309-36867-4.
48. Shabbir, T., Nadeemullah, M. and Memon, S. (2020). Uses and Impact of 'Open Data' Technology for Developing Social Sector in Pakistan. *Pakistan Journal of Multidisciplinary Research*, 1(1), 50-64.
49. Buchanan, A.M., Gentile, D.A., Nelson, D.A., Walsh, D.A. and Hensel, J. (2002). What goes in must come out: Children's media violence consumption at home and aggressive behaviors at school, National Institute on Media and the Family, Minneapolis, MN.
50. Buckingham, D. (2002). The electronic generation? Children and new media. In L.A. Lievrouw & S. Livingstone (Eds.), *Handbook of new media: Social shaping and consequences of ICTs* (pp. 77-88), Sage, London.
51. Gross, E.F., Juvonen, J. and Gable, S.L. (2002). Internet use and well-being in adolescence. *Journal of Social Issues*, 58, 75-90.
52. Keles B., McCrae N. and Grealish A. (2018) A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents, *International Journal of Adolescence and Youth*, 25(1), 79-93.
53. Christofferson, J. (2016). *How is Social Networking Sites Effecting Teen's Social and Emotional Development: A Systemic Review*. Retrieved from Sophia, the St. Catherine University repository website: https://sophia.stkate.edu/msw_papers/650.
54. Undiyaundeye, F. (2014). Impact of social media on children, adolescents and families. *Global Journal of Interdisclipinary Social Sciences*, 3(2), 1-4.
55. Canadian Paediatric Society (2017). Screen time and young children: Promoting health and development in a digital world, *Paediatrics & Child Health*, 2017, 461-468.

56. Hamm, M.P., Shulhan, J., Williams, G., Milne, A., Scott, S.D. and Hartling, L. (2014). A systematic review of the use and effectiveness of social media in child health. *BMC pediatrics*, 14(1), 1-15.
57. Ante-Contreras D. (2016). *Distracted Parenting: How Social Media Affects Parent-Child Attachment*. Master of Social Work Dissertations. California State University.
58. O'Keeffe, G.S. and Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4), 800-804.
59. Stiglic, N. and Viner, R.M. (2019). Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. *BMJ open*, 9(1), e023191.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 83-92

A STUDY ON COMPARISONS OF BAYESIAN AND MODIFIED CLASSICAL PARAMETER ESTIMATION METHODS FOR LINDELY DISTRIBUTION

Adeel Anjum, Muhammad Zafar Iqbal, Taswar Iqbal
Muhammad Kashif and **Muhammad Arslan Bhatti**
Department of Mathematics and Statistics, University of Agriculture
Faisalabad, Pakistan
Email: adeelanjum1954@gmail.com
mzts2004@hotmail.com
taswar.iqbal@njc.edu.pk
mkashif@uaf.edu.pk
bhatti.stat.agri@gmail.com

ABSTRACT

Lindley Distribution is commonly used in the development of models for lifetime data. The Lindley Distribution has vast applications in different fields such as insurance, finance, psychology, medical, and some other areas where numerical study and count data modeling is assumed to be difficult. Bayesian estimation technique along with various modified estimation methods like the modification of Moment Estimators and the modification of Maximum Likelihood Estimators, Ordinary least square, weighted least square was used for the estimation of parameters. Extensive simulation study was conducted by using different values of parameters at different sample sizes. R software was used to access the reliability of different methods by using total deviation (TD) and mean square error technique.

1. INTRODUCTION

Modeling and analyzing the lifetime data are crucial in many applied sciences including insurance, medicine, finance, demography, public health and engineering. So, it is difficult to find the statistical models for modeling the real-world phenomena. There are number of continuous models in literature for modeling lifetime data such as exponential model, Lindley model, gamma model, log-normal model and Weibull model. The Lindley model belongs to an exponential family and it can be written as a mixture of exponential model and gamma model (Merovci, 2013). Recently Morscedy et al. (2020) conducted a comprehensive study on two parameters exponential discrete lindley distribution and its application. Iqbal and Iqbal (2020) developed a new form of lindley distribution by using weighted distributions. In many real-life applications where the data show non-monotone shape hazard function Lindley model is applicable. Altun *et al.* (2017) developed the new three parameter Odd Burr Lindley distribution and derived its structural properties. Sangsanit *et al.* (2016) introduced a lifetime model that is Topp-Leone Exponentiated Power Lindley (TL EPL) model. Sub models of Topp-Leone Exponentiated Power Lindley model were also developed such as Topp-Leone Power Lindley model, Topp-Leone Generalized Lindley model and Topp-Leone Lindley model. (Elbatal *et al.*, 2013) presented a generalized Lindley model. The generalized model contained many models such as Lindley, gamma and exponential models. Different properties of model were derived such as hazard function, reverse hazard function r^{th} moments, moment generating function, inequality measures and order statistics. Three estimation methods were derived including maximum likelihood, least square and weighted least square methods. Al-Mutairi et al. (2013) developed the probability density function (pdf) of single parameter Lindley distribution.

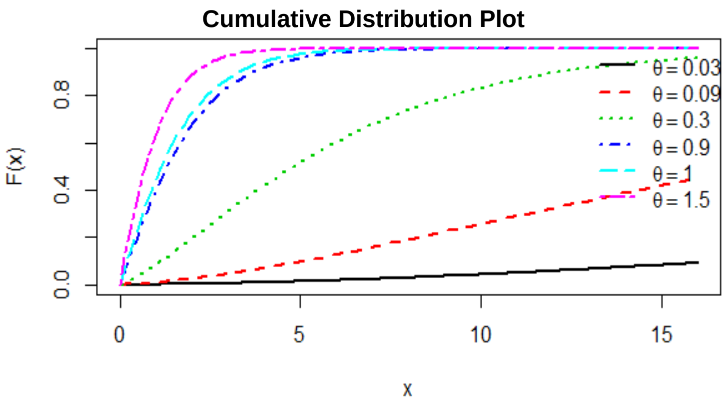
$$f(x,\lambda) = \begin{cases} \left(\frac{\lambda^2}{1+\lambda}\right)(1+x)e^{-\lambda x} & x \geq \lambda > 0 \\ 0 & elsewhere \end{cases} \tag{1}$$

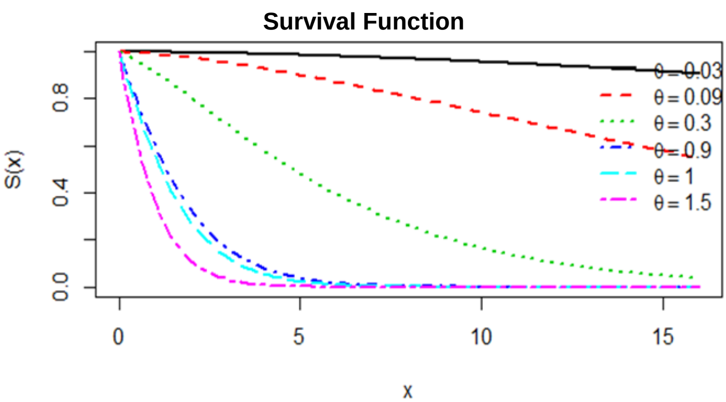
Cumulative Distribution Function:

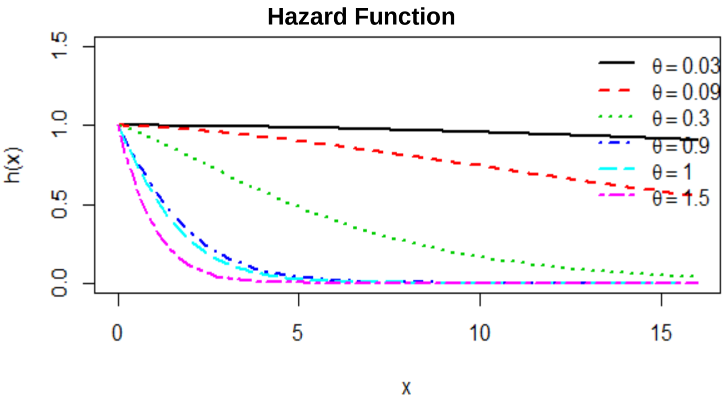
The cumulative distribution function (cdf) of Lindley distribution is given as.

$$F(x;\lambda) = \left(\frac{\lambda^2}{1+\lambda}\right)(1+x)e^{-\lambda x}$$

$$F(x;\lambda) = 1 - e^{-\lambda x}\left(1 + \frac{x\lambda}{1+\lambda}\right) \tag{2}$$







Moments about Origin:

$$(\mu_1') = \left(\frac{\lambda^2}{1+\lambda x}\right)\int_0^\infty x^r\,(1+x)e^{-\lambda x}dx$$

$$\mu_1' = \mathrm{E(X)} = \frac{\lambda+2}{\lambda(1+\lambda)} \tag{3}$$

$$\mu_2' = \frac{\lambda^2}{1+\lambda x}\int_0^\infty x^2(1+x)e^{-\lambda x}\ dx$$

$$\mu_2' = \frac{2(\lambda+3)}{\lambda^2(1+\lambda)}. \tag{4}$$

2. ESTIMATION METHODS

In this section we have discussed different estimation methods to estimate the parameter of Lindley distribution.

Maximum Likelihood Method:

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from Lindley distribution and let λ be the model parameter (Zakerzadeh *et al.*, 2009). The Likelihood function of λ is as follows.

$$L = L\,f(x\,;\lambda) = \frac{\lambda^{2n}}{(1+\lambda)^n}\prod_{i=1}^{n}(1+x)\ e^{-\lambda\sum_{i=1}^{n}x_i} \tag{5}$$

$$\log Lf(x,\lambda) = 2nln\lambda - nln(1+\lambda) + \sum \ln(1+x) - \lambda\sum x \tag{6}$$

By differentiating equation (6) w.r.t λ, we get

$$\frac{\partial}{\partial\lambda}\,(\mathrm{LL}) = \frac{2n}{\lambda} - \frac{n}{1+\lambda} - \sum_{i=1}^{n} x_i \tag{7}$$

By equating above equation to zero we have

$$\lambda^2\bar{x} + \lambda(\bar{x} - 1) - 2 = 0$$

Hence the MLE of parameter can be obtained by solving the polynomials.

$$\hat{\lambda} = \frac{(1 - \bar{x}) \pm \sqrt{(\bar{x} - 1)^2 + 8\bar{x}}}{2\bar{x}} \tag{8}$$

Cramer Rao Lower Bound:

The Cramer Rao Lower Bound for unbiased estimator λ is defined as.

$$VAR_{\lambda(W^*(X))} = \frac{\left(\frac{\partial}{\partial\lambda}(\lambda)\right)^2}{I_n(\lambda)}$$

$$I_{n(\lambda)} = E\left[-\frac{\partial^2}{\partial\lambda^2}\ (LL)\right]$$

By differentiating equation (7) w.r.t λ, we get

$$\frac{\delta^2}{\delta\lambda^2} = \frac{n}{(1+\lambda)^2} - \frac{2n}{\lambda^2} \tag{9}$$

$$I_{n(\lambda)} = \frac{n(\lambda^2 + 4\lambda + 2)}{\lambda^2(1+\lambda)^2} \tag{10}$$

Now Cramer Rao lower bound for unbiased parameter λ is

$$VAR_{\lambda(W^*(X))} = \frac{\lambda^2(1+\lambda)^2}{n(\lambda^2 + 4\lambda + 2)} \tag{11}$$

Method of Moments:

Method of moments is the most popular method to estimate the parameters. Most of the aspects of distribution such as tendency, dispersion, skewness and kurtosis can be studied through method the moments (Shanker *et al.*, 2013).The estimate of the parameters of Lindley distribution using the method of moments are obtained by equating the theoretical moments of the distribution to the corresponding sample moments.

$$m_1' = \bar{x}$$

$$\mu_1' = E(X) = \frac{\lambda + 2}{\lambda(1+\lambda)}$$

By equating to sample moments, we have

$$\bar{x} = \frac{\lambda + 2}{\lambda(1+\lambda)} \tag{12}$$

By solving the above equation, we have

$$\lambda^2\bar{x} + \lambda(\bar{x} - 1) - 2 = 0$$

Hence the MME of λ can be obtained by solving the polynomials.

$$\hat{\lambda} = \frac{(1-\bar{x}) \pm \sqrt{(\bar{x}-1)^2 + 8\bar{x}}}{2\bar{x}} \tag{13}$$

Modify Method of Moments:

In this modification of the moment estimator the second moment of Lindley distribution is replaced by the coefficient of variation (Zaka *et al.*, 2014).

$$S = \sqrt{\frac{\lambda^2 + 4\lambda + 2}{\lambda^2(1+\lambda)^2}} \tag{14}$$

Ordinary and Weighted Least Square Methods:

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from Lindley distribution and suppose that $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denotes the corresponding ordered sample. The expectation and variance of Lindley distribution are independent of the unknown parameter λ is given by.

$$E(F(X_{(i)}) = \frac{i}{n+1}$$

$$V(F\big(X_{(i)}\big) = \frac{i(n-i+1)}{(n+1)^2(n+2)}$$

where $F\big(X_{(i)}\big)$ is the cumulative distribution function (cdf) of Lindley distribution and $x_{(i)}$ is the order statistics. The Least Square estimator can be obtained by minimizing the sum of squares.

$$LS(\lambda, x) = \sum_{i=1}^{n} \left[(F(X_i) - \frac{i}{n+1}\right]^2$$

$$LS(\lambda, x) = \sum_{i=1}^{1} \left[1 - e^{-\lambda x}\left(1 + \frac{\lambda x}{\lambda+1}\right) - \frac{i}{n+1}\right]^2 \tag{15}$$

The derivative of above equation w.r.t λ is not an easy as it cannot be solved analytical so we can use newton Raphson method to solve it. Lu *et al.* (2007) defined the Weighted Least square estimator of λ can be obtained by minimizing.

$$WLs(\lambda, x) = \sum_{i=1}^{n} w_i \left[F x_{i-} \frac{i}{n+1}\right]^2$$

$$w_i = \frac{1}{var(F(x_i\,)} = \frac{(n+1)^2(n+2)}{i(n-i+1)}$$

$$WLS(x;\lambda) = \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left[1 - e^{-\lambda x}\left(1 + \frac{x\lambda}{1+\lambda}\right) - \frac{i}{n+1}\right]^2 \tag{16}$$

The derivative of above equation w.r.t λ is not an easy as it cannot be solved analytical so we can use newton Raphson method to solve it.

Bayesian Parameter Estimation Method:

Bayesian estimation method has received a lot of attention for analyzing failure time data. For this we need to specify prior distributions, $\pi(\lambda)$ for the model parameters. Here we considered a Beta informative prior for unknown parameter λ.

$$P(\lambda) = \frac{\beta^{\alpha}}{\sqrt{\alpha}}\lambda^{\alpha-1}e^{-\beta\lambda}\ \lambda > 0 \tag{17}$$

$$Lf(x;) = \frac{\lambda^{2n}}{(1+\lambda)^n}\prod_{i=1}^{1}(1+x)e^{-\lambda i\sum_{i=1}^{n}x_i} \tag{18}$$

The posterior distribution of λ is

$$\prod(\lambda/x_{1,}x_{2\ldots\ldots.}x_n) = \frac{Lf(x;\lambda)\ P(\lambda)}{\int_0^{\infty} Lf(x;\lambda)\ P(\lambda)d\lambda}$$

$$\prod(\lambda/x_{1,x_{2\ldots\ldots.}x_n}) = \frac{\frac{\lambda^{2n}}{(1+\lambda)^n}\prod_{i=1}^{1}(1+x)\ e^{-\lambda i\sum_{i=1}^{n}x_i}.\frac{\beta^{\alpha}}{\sqrt{\alpha}}\lambda^{\alpha-1}e^{-\beta\lambda}}{\int_0^{\infty}\frac{\lambda^{2n}}{(1+\lambda)^n}\prod_{i=1}^{1}(1+x)\ e^{-\lambda i\sum_{i=1}^{n}x_i}.\frac{\beta^{\alpha}}{\sqrt{\alpha}}\lambda^{\alpha-1}e^{-\beta\lambda}d\lambda}$$

$$P\left(\frac{\lambda}{\underline{x}}\right) \propto \frac{1}{(1+\lambda)^n}\lambda^{(2n+\alpha)-1}e^{-\lambda(\beta+\sum x)}. \tag{19}$$

Bayesian Estimation using Lindley Approximation Method:

The basic idea of Lindley's Approximation is to obtain a Taylor series expansion of function involved in posterior moments (Sharma *et al.*, 2014).

$$E(u(\lambda)|x) = [u + \frac{1}{2}\sum_i\sum_j(u_{ij} + 2\,u_i p_j)\sigma_{ij} + \frac{1}{2}\sum_i\sum_j\sum_k\sum_r L_{ijk\sigma_{ij}\sigma_{kr}]} + O\left(\frac{1}{n^2}\right) \tag{20}$$

The obtain Bayes estimator of λ under prior distribution of Beta

$$U(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$$

Using the approximation for $m = 1$ we have

$U = \lambda\,, U_1 = 1\,, U_2 = 0\ \lambda \propto (\alpha-1)ln\lambda - \beta\lambda$ where $'\beta'$ are constant

$$\phi_1 = \frac{\partial}{\partial\lambda}(\lambda)$$

$$\phi_1 = \frac{\alpha - 1 - \beta\lambda}{\lambda} \tag{21}$$

By using log likelihood of Lindley distribution $\lambda = \hat{\lambda} = \bar{x}$ we have

$$L^2 = \frac{\delta^2(LL)}{\delta\lambda^3} = \frac{n}{(1+n)^2} - \frac{2n}{\lambda^2} \tag{22}$$

$$[-L_2]^{-1} = \sigma^2 = \left[\frac{2n}{\lambda^2} - \frac{n}{(1+\lambda)^2}\right]^{-1} \tag{23}$$

$$L_3 = \frac{-2n}{(1+\lambda)^3} + \frac{4n}{\lambda^3} \tag{24}$$

For single parameter case $m = 1$ the eq (17) reduces to

$$E\left(\frac{U}{x}\right) = \lambda + \phi_1\left[\frac{2n}{\lambda^2} - \frac{n}{(1+\lambda)^2}\right]^{-1} + \left[\frac{2n}{\lambda^2} - \frac{n}{(1+\lambda)^2}\right]^{-2} n\left[\frac{2(1+\lambda)^3 - \lambda^3}{\lambda^3(1+\lambda)^3}\right] + 0\left(\frac{1}{n^2}\right)\Big]\hat{\lambda} \tag{25}$$

By putting the values in eq (23) we have

$$E\left(\frac{U}{x}\right) = \lambda + \frac{\alpha - 1 - \beta\lambda}{\lambda} + \left[\frac{2n(1+\lambda)^2 - n\lambda^2}{\lambda^2(1+\lambda)^2}\right]^{-1} + n\left[\frac{2n(1+\lambda)^2 - n\lambda^2}{\lambda^2(1+\lambda)^2}\right]\left[\frac{2(1+\lambda)^3 - \lambda^3}{\lambda^3(1+\lambda)^3}\right] + 0\left(\frac{1}{n^2}\right)\Big]\hat{\lambda} \tag{26}$$

3. SIMULATION STUDY

In this section we have conducted an extensive simulation study on different sample sizes for the purpose of comparison.

- The following table represents the shape parameter of the Lindley distribution at different sample sizes as $n =$ 20, 60, 100, 300 and various initial values of parameter as $\lambda = 1.5$, 2 and 3.
- Efficiency of the methods is compared on the basis of mean square error (MSE) and Estimate.

Table 1
Comparison of Estimation Methods at $\lambda = 1.5$, 2 and 3

λ	n	MLE		LSE		WLSE	
		Estimate	MSE	Estimate	MSE	Estimate	MSE
1.5	20	1.3940	0.0089	2.8167	0.0167	2.4782	0.0102
	60	1.4263	0.0034	2.0257	0.0158	2.1668	0.0017
	100	1.4541	0.0014	1.7530	0.00081	1.7449	0.00077
	300	1.4951	0.00026	1.5351	0.00015	1.6141	0.00017
2	20	1.8125	0.0048	3.2415	0.0060	3.1973	0.00824
	60	1.8541	0.0024	2.7501	0.0036	2.7301	0.0019
	100	1.9573	0.00199	2.1818	0.00071	2.3615	0.00062
	300	1.9895	0.00041	2.0725	0.00010	2.1097	0.00012
3	20	2.7896	0.0026	4.0843	0.0047	4.9381	0.00369
	60	2.8490	0.0063	3.7932	0.00109	4.0498	0.00109
	100	2.9156	0.0041	3.5329	0.00039	3.3773	0.00037
	300	2.9909	0.00069	3.0532	0.00007	3.1608	0.00007

- By observing the above table, we can conclude that the Maximum likelihood method (MLE) performed better than least square estimation method and weighted least square estimation (WLSE) method at all sample sizes with less mean square error (MSE).
- When we compare the least square estimation (LSE) method with weighted least square estimation (WLSE) method it is concluded that the least square estimation technique produced the comparatively better results as compared to weighted least square as their estimates are close to the true parameter value at all sample size and all values of parameters.

Table 2
Bayesian Lindley Approximation with pair of ($\alpha = 1, \beta = 1$)

λ	n	Estimate	MSE	Bias
1.5	20	1.4324	0.0045	-0.0675
	60	1.4772	0.0005	-0.0227
	100	1.4863	0.0001	-0.0136
	300	1.4954	0.0000	-0.0045
2	20	1.8728	0.0161	-0.1271
	60	1.9573	0.0018	-0.0426
	100	1.9743	0.0006	-0.0256
	300	1.9914	0.0000	-0.0085
3	20	2.6890	0.0966	-0.3109
	60	2.8968	0.0108	-0.1041
	100	2.9374	0.0039	-0.0625
	300	2.9191	0.0004	-0.0208

Table 3
Bayesian Lindley approximation with pair of $(\boldsymbol{\alpha = 2, \beta = 1})$

$\boldsymbol{\lambda}$	$\boldsymbol{n}$	**Estimate**	**MSE**	**Bias**
1.5	20	1.4782	0.0004	-0.0217
	60	1.4925	0.0000	-0.0075
	100	1.4951	0.0000	-0.0045
	300	1.4984	0.00000	-0.0015
2	20	1.9371	0.0039	-0.0628
	60	1.9778	0.0004	-0.0212
	100	1.9872	0.0001	-0.0127
	300	1.9957	0.0000	-0.0042
3	20	2.7934	0.0426	-0.2065
	60	2.9306	0.0048	-0.0693
	100	2.9583	0.0017	-0.0416
	300	2.9960	0.0001	-0.0139

Table 4
Bayesian Lindley approximation with pair of $(\boldsymbol{\alpha = 1, \beta = 2})$)

$\boldsymbol{\lambda}$	$\boldsymbol{n}$	**Estimate**	**MSE**	**Bias**
1.5	20	1.3638	0.01852	-0.1361
	60	1.4543	0.0020	-0.0456
	100	1.4726	0.0007	-0.0273
	300	1.4908	0.0000	-0.0091
2	20	1.7443	0.0653	-0.2556
	60	1.9144	0.0073	-0.0855
	100	1.9486	0.0026	-0.0513
	300	1.9828	0.0002	-0.0171
3	20	2.3760	0.3893	-0.6239
	60	2.7915	0.0434	-0.2084
	100	2.8748	0.01565	-0.1251
	300	2.9782	0.00174	-0.0417

- The above tables represent s the shape parameter of the Bayesian Lindley approximation at different sample sizes as $n =$ 20, 60, 100, 300 and various values of parameter as $\lambda = 1.5$, 2 and 3. Different values of parameters are used to calculate the MSE and their estimates. It is observed that with increasing sample size we get the closer value of true parameter"λ" and minimum MSE.

4. COMMENTS AND CONCLUSION

Lindley Distribution is one of the most popular models for lifetime data. In this article Bayesian estimation technique along with various classical estimation methods like Maximum Likelihood Estimators, Ordinary least square and weighted least square was used for the estimation of parameters. Bayesian Lindley approximation method was used to estimate the hyper-parameter of the distribution. By extensive simulation study the maximum likelihood Estimation method was proved to be more efficient than other competing methods.

5. ACKNOWLEDGEMENT

The authors are deeply thankful to editor and reviewers for their valuable suggestions to improve the quality of the paper.

REFERENCES

1. Al-Mutairi, D.K., Ghitany, M.E. and Kundu, D. (2013). Inferences on stress-strength reliability from Lindley distributions. *Communications in Statistics-Theory and Methods*, 42(8), 1443-1463.
2. Altun, G., Alizadeh, M., Altun, E. and Gamze, O.Z.E.L. (2017). Odd Burr Lindley distribution with properties and applications. *Hacettepe Journal of Mathematics and Statistics*, 46(2), 255-276.
3. Elbatal, I., Merovci, F. and Elgarhy, M. (2013). A new generalized Lindley distribution. *Mathematical theory and Modeling*, 3(13), 30-47.
4. El-Morshedy, M., Eliwa, M.S. and Nagy, H. (2020). A new two-parameter exponentiated discrete Lindley distribution: properties, estimation and applications. *Journal of Applied Statistics*, 47(2), 354-375.
5. Iqbal, T. and Iqbal, M.Z. (2020). On the Mixture of Weighted Exponential and Weighted Gamma Distribution. *International Journal of Analysis and Applications*, 18(3), 396-408.
6. Lu, H.L. and Tao, S.H. (2007). The estimation of Pareto distribution by a weighted least square method. *Quality & Quantity*, 41(6), 913-926.
7. Merovci, F. (2013). Transmuted lindley distribution. *International Journal of Open Problems in Computer Science and Mathematics*, 6(2), 63-72.
8. Sangsanit, Y. and Bodhisuwan, W. (2016). The Topp-Leone generator of distributions: properties and inferences. *Songklanakarin Journal of Science & Technology*, 38(5), 537-548.
9. Shanker, R. and Mishra, A. (2013). A quasi Lindley distribution. *African Journal of Mathematics and Computer Science Research*, 6(4), 64-71.
10. Sharma, V.K., Singh, S.K., Singh, U. and Agiwal, V. (2014). *The inverse Lindley distribution: A stress-strength reliability model*. arXiv preprint arXiv:1405.6268.
11. Zaka, A. and Akhter, A.S. (2014). Modified moment, maximum likelihood and percentile estimators for the parameters of the power function distribution. *Pakistan Journal of Statistics and Operation Research*, 10(4), 369-388.
12. Zakerzadeh, H. and Dolati, A. (2009). Generalized Lindley Distribution. *Journal of Mathematical Extension*, 3(2), 13-25.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 93-102

THE IMPACT OF HOUSEHOLD SOCIO-ECONOMIC STATUS ON CHILD STUNTING IN PAKISTAN

Abid Ali[1] and **Syed Adil Hussain**[1,2]

[1] Department of Mathematics & Statistics, International Islamic University Islamabad, Pakistan. Email: abid55.aa@gmail.com

[2] Department of Statistics, University of Gujrat, Gujrat, Pakistan
Email syed.adil@uog.edu.pk, adil.phdst9@iiu.edu.pk

ABSTRACT

Child stunting is one of the serious issues not only in Pakistan also in developing countries. Documenting it statistically is very important for policymakers. The main objective of the study to explore the impact of household socio-economic conditions on child stunting. We used the secondary data which is taken from the Pakistan Demographic and Health Survey 2017-18, a total of 4226 available samples were used for the analysis. Used Binary Logistic Regression model to assess the results and for statistical analysis used STATA 15. The results found that there are significantly 54%, 53%, and 39% more chances of child stunting in AJK, Sindh, and Balochistan respectively as compared to Punjab at a 1% level of significance. The chances of child stunting significantly decrease 29% and 54% among mothers education level secondary and higher respectively at 1% level of significance. There are more chances of decreasing child stunting ratio with improvement in household wealth status as for poorer, middle, richer and richest class families which are 39%, 45%, 55% are 67% respectively as compared to the poorest families at 1% level of significance. So, the household socio-economic status significantly affects child stunting in Pakistan. From the recent study, we recommended that there is a need for a health care center throughout Pakistan to provide basic needs and sufficient nutrient intake and frequent infections to children especially whose belong to the poorest families. Also, this is very important for government and health care agencies to provide sufficient education to parents about child nutrition and motivate them to provide a healthy environment to their children.

KEY WORDS

PHDS, Child Stunting, Socio- Economic Status

INTRODUCTION

There is a well-known quotation “To eat is a necessity, but to eat intelligently is an art”. The process of supplying or acquiring the food needed for health, development, reproduction, protection, and repair is known as nutrition. Child nutrition requires making certain that children eat healthy foods to help them grow physically and mentally, as well as to prevent future diseases. Malnutrition is a nutritional disorder characterized by a lack of, excess, or deficiency in energy, protein, and other nutrients needed for a healthy and strong life. Child morbidity and mortality have their major cause as undernutrition of age below five years. Children from developing countries account for 90% of the total number of underweight or stunted children. Deaths occur under-five children 41% causes due to malnutrition. Every year, about 10 billion children between the ages of five suffer from illnesses that are preventable and treatable. Malnutrition causes a number of deaths and those that recover suffer from frequent illnesses that result in sickness, stunted growth, and weak learning abilities. Improving a child's diet can help in a variety of ways. It has the potential to make them not only disease-free but also emotionally and psychologically healthy (Anilkumar, *et al.,* 2017).

Poor nutrition during childhood has a detrimental impact on a person's early life and also their entire life span. The importance of adequate nutrition and treatment among the first thousand days of childhood, from conception to the age of two, has been consistently stressed in nutrition and child welfare literature. Low socioeconomic status, insufficient governance, work status, low educational status, poor sanitation, and demographic characteristics such as household size, the gender of the head of household, birth order of children, infant gender, and housing status as a possible description for undernutrition and infant mortality (Sinha, *et al.,* 2017).

In the region, a healthy and strong population is the wealthiest and happiness of the nation. Children are a country's most valuable human resource. Children are the ones who keep a country's potential alive and set the limits for its future growth. “Better the Nutritional Status of the children, higher will be the Nation rise”. Children today are the people of tomorrow, and they should be well. As a result, the nutritional status of a child is critical to the country's growth. Proper nutrition is important for safe growth and development as well as long-term health. Malnutrition is a significant public health concern that affects millions of children around the world, with the majority of cases resulting in death or long-term disability. Children nominated a greater proportion of the global population today (Devi, & Gupta, 2016).

Child undernutrition is one of the world's most severe health issues. According to their figures, undernutrition is a factor in fifty percent of infant deaths in developing countries each year, and it leads to poor long-term growth. Some research that looked into the relationship between food supply and child nutritional status found incongruent results. In developing countries, malnutrition is responsible for the deaths of more than half of preschool children. Although poverty is the primary cause of malnutrition, it often contributes to poverty by raising morbidity, impairing cognitive growth, and reducing job ability and efficiency as people get older. Development organizations and governments all over the world realize that poverty remains the most significant barrier to development, and in the 1990s, world leaders agreed to a standard series of millennium development goals to monitor progress toward reducing poverty and related socio-economic imbalances from 1990 to 2015, including a two-thirds decrease in infant mortality. Various low-income households in developing countries lack the financial capacity to increase or buy adequate food, a necessary condition for good nutritional status; as a result, their diets are low in energy and nutrients (Torlesse, *et al.,* 2003).

Whenever it comes to health issues in developing countries, the nutritional welfare of children has long been a research subject, and malnutrition in children is often regarded as national wellbeing and living conditions are primarily measured by this predictor. The effects of poor nutrition on child growth are well known, including an increased risk of stunting, morbidity, underweight, chronic diseases, even lower working ability, and poor academic performance as children grow older. In recent years, policymakers in developing countries, including China, have paid increased attention to children's nutrition, health, and education, prioritizing human capital expenditure as a crucial strategy for preventing interglacial hunger transmission and closing the nutrition and health gap between urban and rural children (Wu, 2019).

There are some types of malnutrition includes kwashiorkor, marasmus, stunting, wasting, and underweight. Stunting or low height for age is caused by a lack of nutrients for a long time and repeated infections. Stunting usually begins before the age of two, and the consequences are often permanent. It starts before birth as a result of poor maternal health and results in stunted growth in a child with otherwise normal proportions.

Between 1992 and 2002, the prevalence of stunting in children under the age of five raised by 42%, stated by the China National Nutrition Survey and the China National Nutrition and Health Survey. During that decade, this was the largest drop in undernutrition in any region (Bredenkamp, 2009).

In 2012, 162 million children under the age of five were reported to be stunted, accounting for 25% of all children under the age of five. More than 90 percent of the globe's stunted children are born in Africa and Asia, with 36 percent and 56 percent of children, respectively, living in these two regions (Web, 2013). According to a UNICEF survey, 25% and 8% of under five years old children are stunted and wasted, respectively, and an approximate 6.3 million live-born children died before attaining the age of five in 2013 (Ali, *et al.,* 2016).

In 2016, Asia had 87 million stunted infants, Africa had 59 million, and Latin America and the Caribbean had 6 million. Stunting affects a significant number of children around the world, according to these global patterns (Bangoura, 2018). In the report of WHO, there were at least 155, 52, and 99 million stunted, wasted, and underweight children under the age of five worldwide in 2016. Approximately 6 million children were also confirmed to be stunted and wasting as a result. Malnutrition is a disorder that affects developing countries, especially those in Africa and South Asia. India, Pakistan, and Bangladesh, three countries in South Asia, have an especially high incidence of the disease. Malnutrition is a leading cause of morbidity and mortality among children under the age of five in Pakistan, which ranks 22nd in the world for infant deaths under the age of five (Khan, *et al.,* 2019).

According to a 2011 survey, malnutrition affects people of all ages in Pakistan, and the trend has not slowed in the last decade. One-third of Pakistan's children are underweight, 44 percent are stunted, 15 percent are wasting, and half are anemic, and approximately one-third suffer from iron deficiency anemia. The incidence of stunting varies substantially across Pakistan's districts, ranging from 22 percent to 76 percent (Das, *et al.,* 2016). The Pakistan Demographic and Health Survey 2017-18 reported that 38% of Pakistani children are stunted (below -2 SD), with 17% severely stunted (below -3 SD). Stunting as children get older rises, peaking at 48% between the ages of 24 and 35 months. In rural areas, the ratio of stunted infants is high (41%) than in urban areas (31%). Children whose mothers are illiterate are more likely to be stunted than those whose mothers are educated (PDHS, 2017-18).

Khan, S., *et al.,* (2019) recommended that the majority of the factors investigated that led to malnutrition in Pakistani infants, such as mother's age at marriage, educational level, and mothers' nutritional status, are preventable. As a result, countermeasures such as community-based development and direct dietary intervention are needed to reduce the burden of malnutrition. Das, *et al.,* (2016) suggested that the situation is optimal for improvement, with a recent concentration on nutrition and the implementation of numerous national and regional nutrition policies; nutrition plans must be seen as a net investment in the country's future by policymakers and developers, who must understand the importance of better child health and nutrition for national growth. Mian, R.M., *et al.,* (2002) asserted that malnutrition is more common in lower-income households. According to Laghari, Z.A., *et al.,* (2015) explained that female children have a higher average incidence and seriousness of malnutrition than male children. According to Khan, G.N., *et al.,* (2016) malnutrition was widespread among children under the age of five, according to the findings. To strengthen the country's malnutrition condition, food/nutrient-based participation should be combined with better hygiene habits and household income.

A study conducted by Bangoura A. (2018) in Guinea. Because of its connection to increased morbidity and mortality of infants, he advised that in developing countries, child stunting is also a significant public health problem. Between 1999 and 2012, the prevalence of child stunting rose from 30.5 percent to 31.2 percent. He also looked into the causes of stunting and extreme stunting in Guinean children under the age of five. According to 2016 WHO recommendations, the height for age Z-score (HAZ) was used to define stunting into two categories: moderate stunting (HAZ<-2 standard deviation) and severe stunting (HAZ<-3 standard deviation). To identify risk factors for stunting, researchers used statistical approaches such as multivariate and multinomial logistic regressions. Boys (32.69 percent) were found to be more stunted than girls in this report (29.12%). Other variables that contributed to fluctuating development included the wealth index, short birth intervals, the mother's low BMI, and unimproved drinking water supplies. He concludes that reducing hunger, improving women's education, child feeding habits, household hygiene, and family planning should all be priorities in addressing childhood stunting in Guinea. Devi, & Gupta, (2016) evaluate the nutritional condition of preschool children in an urban and rural community. They also discovered that rural girls had greater nutritional status than urban girls. Ekbrand, H. & Halleröd, B., (2018) recommended that gender equality in education and jobs reduces malnutrition among children and that women’s empowerment reduces health distress for children with uneducated mothers. According to Chikhungu, & Madise (2014), an infant's nutritional condition changes with the seasons and maintains a seasonal cycle of childhood disease, but not of household food availability. Wu, Yichao., (2019) recommended that from 1991 to 2009, the ratio of undernourished infants who are stunted or underweight has declined in China. Often, say that economic growth is a crucial factor in enhancing disadvantaged children's nutritional status. Cetthakrikul, N., *et al.,* (2018) stated that childhood stunting was caused by several causes. Also concluded that children living in low-income families were more likely to be stunted than rich families.

MATERIAL AND METHOD

Data

Data for the given study taken from Pakistan Demographic and Health Survey (PDHS) 2017-18. In this survey, rural and urban areas of Pakistan are included. The PDHS 2017-18 conducted in all provinces and regions in Pakistan includes Punjab, Sindh, Khyber Pakhtunkhwa, Balochistan, AJK, FATA, Gilgit Baltistan, and Islamabad (ICT). A total of 4226 available samples were used for the analysis.

The Child Growth Standards of the World Health Organization (WHO) from 2006 were used. Children's chronic nutritional condition were measured using height for age z-scores. The height for age z-score determines when a child's height is above or below the median height of healthy children in the same age group or a comparison group in standard deviations (SDs). Children who were <-2 SD from the median were considered short for their age (stunted).

Model Specification

We have to specify the effect of household socio-economic status and other associated factors on child stunting. Hence, the dependent variable of our model is child stunting status. The following model is specified.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \qquad (3.1)$$

where Y represent child stunting status, while X_1, X_2, X_3, X_4, and X_5 represent the region, place of residence, highest education level of mothers, wealth index, and sex of child respectively.

Statistical Methodology

Logistic Regression Analysis

Since most of the response variables in social sciences are binary, we use binary logistic regression to describe such data since the dependent variable in the given study has one of two categories: child stunted or not stunted (normal). As compared to the probit model, it is less sensitive to outliers (Copas, 1988). Where the explanatory variables are categorical or a mixture of continuous and categorical, logistic regression is favored over discriminant analysis (Agresti, 2018). It is also not yet required the normality of variables nor claims the linearity in the relationship among independent and the dependent variables (Hilbe, 2009).

The Logit Model

Since our study's dependent variable is of the dichotomous kind, there are two potential outcomes that are “child stunted” (taken as 0) or “child not stunted” (taken as 1). Instead of using the traditional least square method to estimate the model's parameters, the maximum likelihood (ML) method is used.

Consider the equation as

$$Y_i^* = \beta' X_i + U_i \qquad (3.2)$$

where $i = 1,2,3,\ldots\ldots,n$ & $\beta' = \beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5$ are the regression coefficients and X_i are explanatory variables.

Odd-Ratio

The purpose of relative odd ratios is to see whether the child nutrition status of the infant changes with a unit shift in the particular explanatory variable while holding the data balanced on all the remaining variables. As

$$Ln\ (odd_2) - Ln\ (odd_1) = Ln\left(\frac{odd_2}{odd_1}\right) = \text{B},$$

where,

$$odd_2 = P \text{ (child not stunted or normal) and } odd_1 = P \text{ (child stunted)}$$
$$odd_2/odd_1 = exp\ (B),$$

So we can write that;

$$Odd\text{-}Ratio = exp\ (B).$$

Test of significance

Wald test

The Wald test is used to explore the importance of individual coefficients of predictors in a logit model. The null hypothesis is being tested in this research; $H_0: \beta_i = 0$, where i= 1,2,…,5 on the basis of test statistic;

$$W = \left(\frac{\widehat{\beta_i}}{SE(\widehat{\beta_i})}\right)^2 \tag{3.3}$$

For a large sample size, the Wald test assumes the chi-square distribution of one degree of freedom (Long and Fareese, 2006).

RESULTS AND DISCUSSION

Results

Characteristics of the Sample

The prevalence of stunting recorded 37.39% from the available data of PDHS 2017-18. The child stunting ratio is high in province Sindh which is (22.91%) followed by Punjab (17.09%), KPK (15.38%), Balochistan (15.13%), FATA (11.39%), AJK (7.66%), Gilgit Baltistan (6.9%) and Islamabad (3.54%) respectively. Those children whose mothers have no education their ratio is high which is (64.30%), primary education (13.16%), secondary education (15.44%), and higher education (7.09%). The stunting ratio is high in males which is (51.58%) than females which have (49.24%). In rural areas stunting ratio are high than in urban which is (60.44%) and (39.56%) respectively. In the poorest families, the child stunting ratio is high (32.09%) followed by the poorer, middle, richer and richest which are (27.09%), (17.85%), (13.86%) and (9.11%) respectively.

Logistic Regression Results

The following table 1 shows the results of logistic regression analysis of variables. Also presenting the odd ratio to compare the results with each other.

Table 1
Results of Logistic Regression Analysis of Child Stunting and their Characteristics

Characteristics/variable	**Stunting**		
	Odd Ratios	**P-Value**	**C.I**
Region			
Punjab	ref		
Sindh	0.47	0.000*	1.243927 1.90165
Khyber Pakhtunkhwa	0.9	0.366	.8878092 1.381066
Balochitan	0.61	0.007*	1.093619 1.777161
Gilgit Baltistan	0.75	0.133	.9333274 1.683593
Islamabad (ICT)	1.05	0.78	.6729484 1.346443
AJK	0.46	0.001*	.6694549 1.135201
FATA	1.13	0.308	1.183245 2.023091
Mothers Education			
No Education	ref		
Primary	1.13	0.211	.7153919 1.076743
Secondary	1.29	0.001*	.5807272 .8689879
Higher	1.54	0.000*	.3580998 .6069046
Sex of Child			
Male	ref		
Female	1.11	0.144	.7967887 1.033802
Type of Residence			
Urban	ref		
Rural	1.09	0.284	.7859866 1.07306
Socio Economic Status			
Poorest	ref		
Poorer	1.39	0.000*	.5060778 .7392037
Middle	1.45	0.000*	.4481524 .6983409
Richer	1.55	0.000*	.3537493 .5858709
Richest	1.67	0.000*	.2458897 .4456953
* represent significance at 1%			

The main findings of the recent study shown in table 1. First, discuss the ratio of stunting in different regions of Pakistan, there are significantly 53%, 39%, and 54% more chances of child stunting in Sindh, Balochistan, and AJK respectively as compared to Punjab at a 1% level of significance. The odd ratios of KPK, Gilgit Baltistan, Islamabad, and FATA are insignificant for child stunting status. The chances of child stunting significantly decrease 29% and 54% among mothers education level secondary and higher respectively at 1% level of significance, also have same results are found look for details Khan, S., *et al.,* (2019). The odd ratios of the place of residence and sex of the child are insignificant for child stunting status.

According to table 1 there is a significant impact of socio-economic status of household on child stunting in Pakistan. There are more significant chances of decreasing child stunting ratio with improvement in household wealth status as for poorer, middle, richer and richest class families which are 39%, 45%, 55% are 67% respectively as compared to the poorest family at 1% level of significance, same results are found see for more details Mian, R.M., *et al.,* (2002), Bangoura A. (2018) and Cetthakrikul, N., *et al.,* (2018). The household socio-economic condition is significantly affecting the child's nutrition status because poor families have no access to safe food so there is a lack of nutrition for under-five children.

In the table column of C.I show the confidence interval of odds. From the test of significance, the value of the Wald Test is 271.49 with a p-value of 0.0000 which indicates that the final model with all explanatory variables is more effective than the null model.

CONCLUSION AND RECOMMENDATION

It is observed from the study that the household socio-economic status significantly affects child stunting in Pakistan. The ratio of stunting in the poorest household is very high as compared to the richest household which is 67% more than the richest. There are significantly 53%, 39%, and 54% more chances of child stunting in Sindh, Balochistan, and AJK respectively as compared to Punjab at 1% level of significance. Also education status of mothers affects the ratio of child stunting, the ratio of child stunting is high in those children whose mothers have no education as compare to educated mothers. From the recent study, we recommended that there is a need for a health care center in each union council, tehsil and districts throughout Pakistan to provide basic needs and sufficient nutrient intake and frequent infections to children especially whose belong to low-income families. Also, this is very important for government and health care agencies to provide sufficient education to parents about child nutrition and motivate them to provide a healthy environment to their children.

REFERENCES

1. Anilkumar, N.A., Gupta, D., Khare, S., Gopalkrishna, D.M. and Jyotishi, A. (2017). Characteristics and causes of malnutrition across Indian states: a cluster analysis based on Indian demographic and health survey data. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2115-2123). IEEE.
2. Ali, A.M., Batu, M.M. and Kaushik, K.K. (2016). Socio-economic determinants of nutritional status of children in Ethiopia. *Int. J. Sci. Res. Publications*, 6(3), 166-76.
3. Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.
4. Bredenkamp, C. (2009). Policy-related determinants of child nutritional status in China: The effect of only-child status and access to healthcare. *Social Science & Medicine*, 69(10), 1531-1538.
5. Bangoura, A. (2018). *Determinants of Childhood Stunting in Guinea: Further Analysis of Demographic and Health Survey 2012*. https://scholarworks.gsu.edu/iph_theses/579
6. Chikhungu, L.C. and Madise, N.J. (2014). Seasonal variation of child under nutrition in Malawi: is seasonal food availability an important factor? Findings from a national level cross-sectional study. *BMC Public Health*, 14(1), 1-11.

7. Cetthakrikul, N., Topothai, C., Suphanchaimat, R., Tisayaticom, K., Limwattananon, S. and Tangcharoensathien, V. (2018). Childhood stunting in Thailand: when prolonged breastfeeding interacts with household poverty. *BMC Pediatrics*, 18(1), 1-9.
8. Copas, J.B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2), 225-253.
9. Devi, S. and Gupta, S. A descriptive study to assess the nutritional status of preschool children in urban and rural community at Karnal. International Journal of Advanced Education and Research, 1(4), 26-28.
10. Das, J.K., Achakzai, A.B.K. and Bhutta, Z.A. (2016). Stop stunting: Pakistan perspective on how this could be realized. *Maternal & Child Nutrition*, 12(Suppl 1), 253-256.
11. De Onis, M. (2006). WHO child growth standards. *Geneva: WHO*, 1Á336.
12. Ekbrand, H. and Halleröd, B. (2018). The more gender equity, the less child poverty? A multilevel analysis of malnutrition and health deprivation in 49 low-and middle-income countries. *World Development*, 108, 221-230.
13. Hilbe, J.M. (2009). *Logistic regression models*. CRC press.
14. Khan, S., Zaheer, S. and Safdar, N.F. (2019). Determinants of stunting, underweight and wasting among children< 5 years of age: evidence from 2012-2013 Pakistan demographic and health survey. *BMC Public Health*, 19(1), 358.
15. Khan, G.N., Turab, A., Khan, M.I., Rizvi, A., Shaheen, F., Ullah, A., Hussain, A., Hussain, I., Ahmed, I., Yaqoob, M. and Ariff, S. (2016). Prevalence and associated factors of malnutrition among children under-five years in Sindh, Pakistan: A cross-sectional study. *BMC Nutrition*, 2(1), 69. https://doi.org/10.1186/s40795-016-0112-4.
16. Laghari, Z.A., Soomro, A.M., Tunio, S.A., Lashari, K., Baloach, F.G., Baig, N.M. and Bano, S. (2015). Malnutrition among children under five years in district Sanghar, Sindh, Pakistan. *Gomal Journal of Medical Sciences*, 13(1), 54-57.
17. Long, J.S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (Vol. 7). Stata Press.
18. Mian, R.M., Ali, M., Ferroni, P.A. and Underwood, P. (2002). The nutritional status of school-aged children in an urban squatter settlement in Pakistan. *Pakistan Journal of Nutrition*, 1(3), 121-123.
19. McCullagh, P. and Nelder, J.A. (1989). Generalized linear models, 2nd Ed. (Chapman and Hall: London). *Standard book on generalized linear models*.
20. Nkurunziza, S., Meessen, B. and Korachais, C. (2017). Determinants of stunting and severe stunting among Burundian children aged 6-23 months: evidence from a national cross-sectional household survey, 2014. *BMC Pediatrics*, 17(1), 176.
21. National Institute of Population Studies (NIPS) [Pakistan] and ICF (2019). *Pakistan Demographic and Health Survey 2017-18*. Islamabad, Pakistan, and Rockville, Maryland, USA: NIPS and ICF.
22. Park, H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43(2), 154-164.
23. Ren, W., Rammohan, A. and Wu, Y. (2014). Is there a gender gap in child nutritional outcomes in rural China? *China Economic Review*, 31, 145-155.

24. Sinha, A., McRoy, R.G., Berkman, B. and Sutherland, M. (2017). Drivers of change: Examining the effects of gender equality on child nutrition. *Children and Youth Services Review*, 76, 203-212.
25. Shroff, M., Griffiths, P., Adair, L., Suchindran, C. and Bentley, M. (2009). Maternal autonomy is inversely related to child stunting in Andhra Pradesh, India. *Maternal & Child Nutrition*, 5(1), 64-74.
26. Torlesse, H., Kiess, L. and Bloem, M.W. (2003). Association of household rice expenditure with child nutritional status indicates a role for macroeconomic food policy in combating malnutrition. *The Journal of Nutrition*, 133(5), 1320-1325.
27. Wu, Y. (2019). What is known about children's undernutrition and health levels in China–An empirical analysis from 1991 to 2009. *Children and Youth Services Review*, 96, 372-380.
28. United Nations Children's Fund, World Health Organization, The World Bank (2013). *UNICEF-WHO-World Bank Joint Child Malnutrition Estimates*. Retrieved from (http://data.unicef.org/resources/2013/webapps/nutrition).
29. Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. Online available at http://pkg.cs.ovgu.de/LNF/i386/5.10/R/LNFr-lmtest/reloc/R-2.10/library/lmtest/doc/lmtest-intro.pdf

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 103-117

FORECASTING PRODUCTION OF SESAMUM CROP IN PAKISTAN: A TIME SERIES APPROACH

Zahid Iqbal[1] and **Jalal-ud-Din**[2]
[1] Allama Iqbal Open University, Islamabad, Pakistan
Email: zahid.iqbal@aiou.edu.pk
[2] Crop Reporting Service, Agriculture Department,
Rawalpindi, Pakistan. Email: jalalmusayyab@gmail.com

ABSTRACT

Particular study is carried out to forecast the production of Sesamum crop in Pakistan by using time series analysis. After satisfying stationary and normality condition. One of the prerequisite assumptions of linearity is checked by a popular Brock, Dechert and Scheinkman (BDS) test on the residuals of Autoregressive integrated moving average (ARIMA) models (0,1,1). The BDS test shows that ARIMA (0,1,1) model fulfills the linearity condition. Then ARIMA (0,1,1)) model is used for forecasting the production of Sesamum crop in Pakistan for the year 2016-17 to 2021-22, which shows an increasing trend in future.

KEYWORDS

Time series ARIMA, Linearity, BDS test, Sesamum production, forecast

INTRODUCTION

Pakistan is one of the beautiful countries of the world which have four seasons. No doubt that Pakistan has potential and suitable atmosphere for agriculture. But unfortunately, no remarkable attention is given to agriculture sector by the previous Governments in Pakistan. For example, it spends a lot of money on the import of Palm and Soybean oil. Although many oilseed crops are cultivated in Pakistan, but which are not sufficient to meet agriculture needs. How much we should grow and which agriculture commodities? Are very important task to government. By better planning and correct knowledge of needs maybe helpful to decrease imports of oilseed crop in Pakistan. If we see the imports of Pakistan, it is come to our knowledge that only in 2017-18 we spent 223879 million rupees for palm oil and 14683 million rupees on Soybean oil (Agriculture Statistics 2017-18). We are utilizing 38.5% of our workforce in agriculture sector and this sector of Pakistan contributes 18.5% to Gross Domestic Product (Pakistan Economics Survey 2018-19).

One of the important kharif crops in Pakistan is Sesamum crop. Sesamum (Sesame) is known as queen of oilseed crop because of its many prominent properties over other oilseed crops. Its family of Pedaliaceae is known as one of the oldest oilseed crop of that region its cultivation history is more than 5000 before Christ (B.C) old in Harrapa valley (Zhong 2013).

Sesamum in Pakistan is locally called Til and is cultivated in both irrigated and un-irrigated areas. The need is to increase in the area and production of such important oilseed crop in the country. From last few years the production of the Sesamum crop is decreasing due to lack of government's interest toward this human health friendly crop. Increase in the production of Sesamum crop can be made by proper care of the crop and giving free good quality of seed to the farmer.

Our study purpose is to provide correct estimate about production of Sesamum crop by using time series analysis. We tried to choose such a model for forecasting purpose, which fulfill all prerequisites assumptions including linearity of residuals. For checking linearity, we applied a popular and reliable Brock Dechart Sheinkman (BDS) test in this study.

MATERIALS AND METHODS

The data for production of Sesamum crop from 1947-48 to 2016-17 retrieved from www.amis.pk the website of agriculture marketing service, Punjab, Pakistan.

In time series analysis the most reliable and well recognized technique, Box-Jenkins ARIMA approach is used for finding appropriate ARIMA model for forecasting production of Sesamum crop.

The reason behind choosing ARIMA model for forecasting purpose is that it assumes non-zero autocorrelations within the values of data (Kumar and Anand, 2014). The linear feature of the data is captured by ARIMA model and it is very important to satisfy linearity condition before using the appropriate chosen ARIMA model for forecasting the production of Sesamum crop.

We checked normality condition of the series by histogram. To achieve for normality condition, we used logarithmic transformation. The graph of the time series, correlogram and unit root test is used to for stationary condition. To overcome the problem of stationary differencing technique is used. After satisfying the both condition of stationary and normality of the series. The correlogram helps us to find the values of autoregressive (p) and moving average (q) for the model. The values of AIC, Normalized BIC etc. helps us to choose the appropriate model and the BDS test on the residuals from the fitted ARIMA models for confirming linearity condition is also done.

Null hypothesis H_0 : Time series is linear

The BDS test was developed for testing the independence and identical distribution (iid). After further studies it is noted that it can performs well to check linearity also (Brock 1991).

At first the series are as $x_t^m = \left(x_1,\, x_{t+1}, \ldots,\, x_{t-m-1}\right)$, where $t = 1,2,3,\ldots,t-m$ for integer embedding dimension m which is greater than or equal to two. The correlation integral $= \left(C_m T\ \left(\varepsilon\right)\right)$ is then computed with distance = epsilon of each other.

$$C_M T(\varepsilon) = \frac{2}{(T-m+1)} \sum_{t<m} I_\varepsilon \left(X_t^m - X_s^m\right)$$

where I_ε is the indicator function.

$$I_\varepsilon = \begin{cases} 1 \; \|X_t^m - X_s^m\| < \varepsilon \\ 0 \; Otherwise \end{cases}$$

The observed X_t are independent identical distributed (iid) under the null hypothesis, then $C_{m,I}(\varepsilon) - C_{I,t}(\varepsilon)^m$ with probability one as sample size and ε tends to infinity and zero respectively.

Brock et al., (1996) shows, the BDS test statistics as,

$$V_{m,I}\,(\varepsilon) = \frac{\sqrt{T}\{C_{m,I}(\varepsilon) - C_{I,T}\,(\varepsilon)^m\}}{\sigma_{m,I}(\varepsilon)}$$

where

$$\sigma^2{}_{m,I}(\varepsilon) = 4\left[k^m + 2\left\{\sum_{j-i}^{m-1} k^{m-j}\, C_{1,T}(\varepsilon)^{2j}\right\} - (m-1)^2 C_{1,T}(\varepsilon)^{2m} - m^2 k(\varepsilon) C_{1,T}(\varepsilon)^{2m-2}\right]$$

where,

$$k = k(\varepsilon) = \frac{6\sum_{t<s<r} h_\varepsilon(\varepsilon) x_t^m,\; x_r^m, x_s^m}{[(T-m+1)(T-m)(T-m-1)]}$$

$$C = C_{1,T(\varepsilon)}$$

$$h_\varepsilon(i,j,k) = \frac{[I_\varepsilon(i,j)I_\varepsilon(j,k) + I_\varepsilon(i,k)I\varepsilon(k,j) + I_\varepsilon(j,i)I_\varepsilon(i,k)]}{3}$$

$C_{I,t}(\varepsilon)$ and $k(\varepsilon)$ both are consistent estimators and belongs to U-statistics and also efficient estimator of C and k respectively (Kanzler 1999).

As BDS test is 2 tail test, at level of significance $\alpha = 0.05$ the null H_0 will be rejected if the calculated value of test statistics fails to fall in the acceptance region.

A test of joint hypothesis of all autocorrelation coefficients (ρ_k) equal to zero for specific lags, which is done by Q-Statistic by Box & Pierce.

$$Q = n\sum_{k=1}^{m} \hat{\rho}_k^2$$

where

$n =$ sample size

$k = 1, 2, \ldots, m$

$\hat{\rho}_k =$ autocorrelation of the estimated residuals

$m =$ lags length

(Gujrati & Porter 2009).

The Q statistic is distributed as χ^2 for large samples with degree of freedom $= m$, we may reject null H_0 when calculated value of Q does not falls in acceptance region. A variant of Q statistics is Ljung Box statistics which is

$$\text{Ljung Box} = n\,(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\rho}^2_{n-k}}{n-k}\right) \sim \chi^2 m$$

(Gujrati & Porter 2009).

For large sample size Q and LB statistic follows Chi distribution with $d.f = m$ but for small size properties Ljung Box statistic is considered more powerful test.

After model diagnostic testing of appropriate ARIMA model which fallows the linearity condition used for forecasting.

RESULTS

The production of Sesamum crop for the period 1947-48 to 2012-13 is used for model building and for model testing for 2013-14 and 2016-17 is done. Finally forecasting for the year 2017-18 to 2021-22 is made.

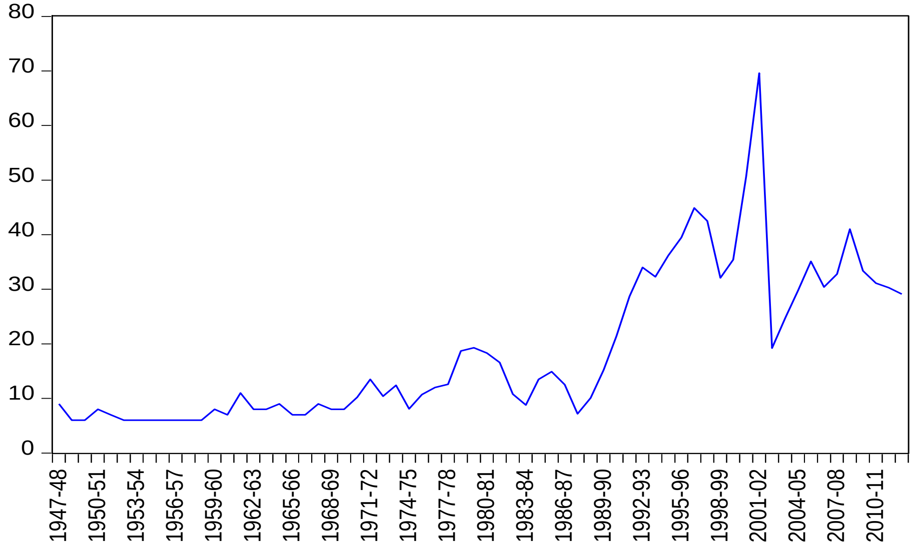


Figure 1: Production of Sesamum Crop in Pakistan from 1947-48 to 2012-13

The graph for production of Sesamum crop is shown in figure 1 from the year 1947-48 to 2012-13. The graph shows a clearly upward trend after 1988-89. First of all we checked its normality by histogram.

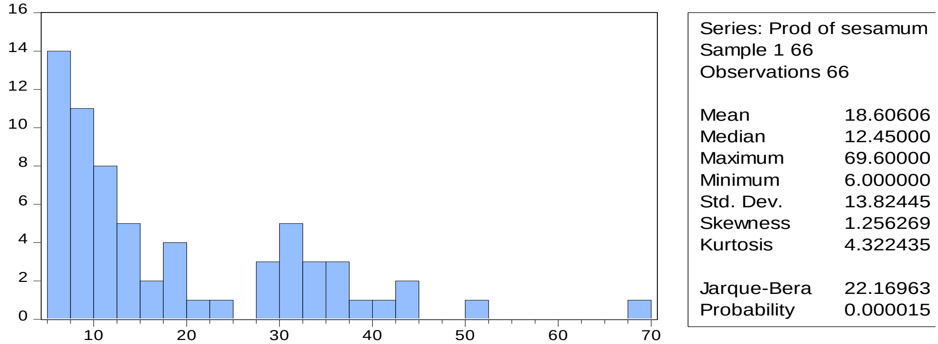


Figure 2: Histogram for Production of Sesamum Crop from 1947-48 to 2012-13

As shown in Figure 2 the value of Jarque Bera is 22.16963 and the probability is 0.000015. This clearly shows us that the series is not normal. Now we have to make the series normal, for this we use logarithmic transformation.

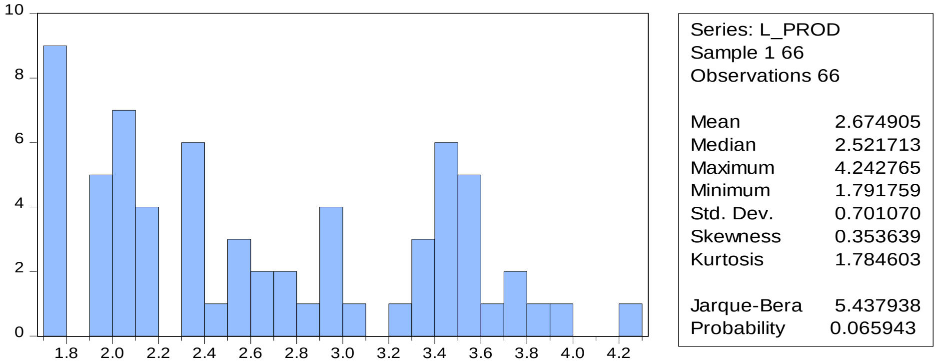


Figure 3: Histogram for L(Production) of Sesamum Crop from 1947-48 to 2012-13

As shown in the Figure 3, the transformed series has p value 0.065943 This shows normality of the transformed series. Now we will check its stationary by correlogram and unit root test.

Table 1
Correlogram of L(Production) of Sesamum in Pakistan

Date: 09/27/20 Time: 11:13
Sample: 1 66
Included observations: 66

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.910	0.910	57.119	0.000
		2	0.832	0.030	105.71	0.000
		3	0.787	0.145	149.81	0.000
		4	0.759	0.107	191.53	0.000
		5	0.706	-0.115	228.24	0.000
		6	0.649	-0.039	259.72	0.000
		7	0.602	-0.001	287.27	0.000
		8	0.554	-0.060	311.02	0.000
		9	0.512	0.024	331.66	0.000
		10	0.476	0.027	349.84	0.000
		11	0.453	0.058	366.57	0.000
		12	0.395	-0.187	379.55	0.000
		13	0.347	0.014	389.72	0.000
		14	0.325	0.087	398.85	0.000
		15	0.313	0.019	407.47	0.000
		16	0.266	-0.135	413.82	0.000
		17	0.215	-0.037	418.04	0.000
		18	0.159	-0.151	420.40	0.000
		19	0.126	0.044	421.90	0.000
		20	0.085	-0.054	422.61	0.000
		21	0.030	-0.106	422.70	0.000
		22	-0.023	-0.046	422.76	0.000
		23	-0.063	0.033	423.18	0.000
		24	-0.094	0.007	424.12	0.000
		25	-0.103	0.122	425.27	0.000
		26	-0.115	-0.030	426.75	0.000
		27	-0.127	0.087	428.60	0.000
		28	-0.133	0.022	430.69	0.000

Table 2
Unit Root Test for Transformed Series

Null Hypothesis: L_PROD has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.535673	0.5094
Test critical values:	1% level	-3.534868	
	5% level	-2.906923	
	10% level	-2.591006	

*MacKinnon (1996) one-sided p-values.

In Table 1 it is shown that series is nonstationary and in Table 2 the P-value is .5094 which clearly indicates that the null H_0 cannot be rejected and the particular series is nonstationary. Now we have to make it stationary before applying further statistical methods. To make the series stationary a popular approach is differencing.

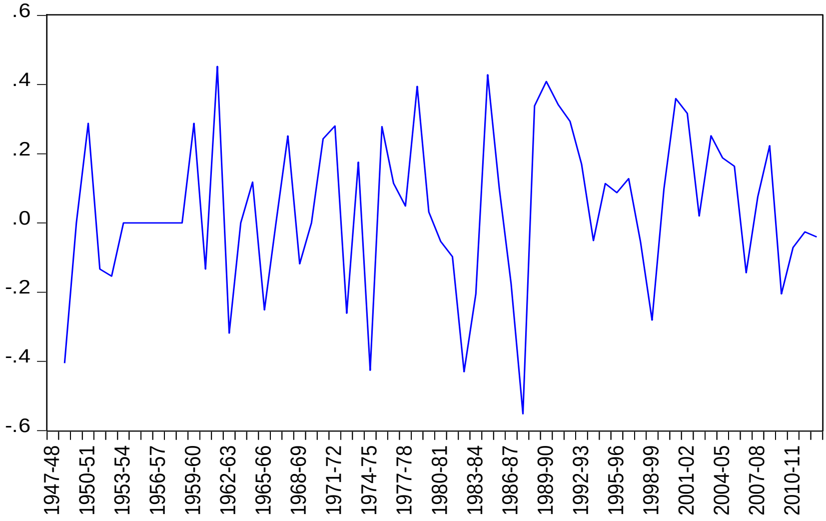


Figure 4: Graph of DL (production) of Sesamum in Pakistan

For making the series stationary we use differencing technique and take first difference of transformed series as shown in the Figure 4. We can see that the mean and variance is looking constant, which means that series is stationary at first difference. To confirm stationary condition we use unit root test.

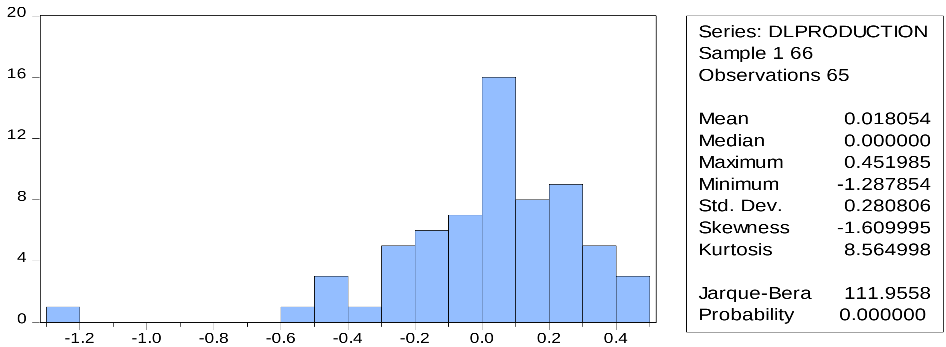


Figure 5: Histogram for DL (production) of Sesamum Crop in Pakistan

It is shown in Figure 5 that the P value = 0.0000, which indicates that we cannot consider series as normal. In histogram we can clearly see outlier and outlier may be the obstacle for achieving normality. We can see the value of Kurtosis = 8.564998, which is much greater than 3 and similarly value of Skewness = -1.609995 is also less from 0 but acceptable. This may be due to outlier or nonlinearity. First, we remove outlier and we replace it by the mean of DLProduction series as mean = 0.018054 and then we will draw histogram of that series if basic diagnostic test fails then we will move to nonlinear model.

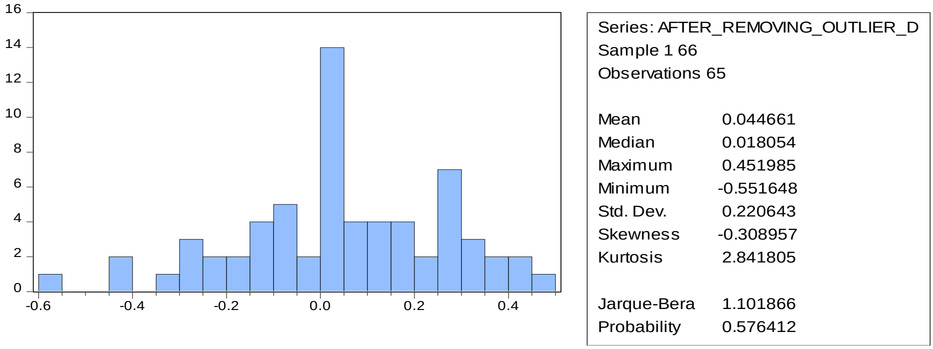


Figure 6: **Histogram for DL(Production) of Sesamum Crop in Pakistan after Removing Outlier**

In Figure 6 after removing outliers we can see the value of skewness and kurtosis are become close to 0 and 3 respectively, the value of Jarque-Bera = 1.311804 and prob value = 0.598608 which indicate that the series is normal. Thus, series passes the basic diagnostic test we will move towards ARIMA model.

Table 3
Correlogram of DL (Production) of Sesamum in Pakistan after Removing Outlier

Date: 09/29/20 Time: 19:54
Sample: 1 66
Included observations: 65

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.013	-0.013	0.0108	0.917
		2	-0.103	-0.103	0.7439	0.689
		3	-0.106	-0.110	1.5388	0.673
		4	-0.055	-0.071	1.7519	0.781
		5	0.039	0.013	1.8612	0.868
		6	0.010	-0.014	1.8687	0.931
		7	-0.123	-0.135	3.0066	0.884
		8	-0.051	-0.060	3.2023	0.921
		9	-0.101	-0.136	3.9885	0.912
		10	-0.056	-0.116	4.2340	0.936
		11	0.181	0.122	6.8683	0.810
		12	0.115	0.086	7.9479	0.789
		13	-0.017	-0.010	7.9724	0.845
		14	0.069	0.109	8.3746	0.869
		15	-0.019	0.019	8.4058	0.906
		16	-0.058	-0.075	8.7006	0.925
		17	0.064	0.054	9.0759	0.938
		18	-0.104	-0.085	10.079	0.929
		19	0.012	0.023	10.092	0.951
		20	-0.050	-0.024	10.338	0.962
		21	-0.075	-0.026	10.895	0.965
		22	-0.076	-0.118	11.481	0.967
		23	0.040	-0.017	11.645	0.976
		24	0.030	-0.001	11.743	0.983
		25	0.105	0.027	12.935	0.977
		26	-0.130	-0.166	14.813	0.961
		27	-0.028	-0.019	14.904	0.971
		28	-0.016	-0.089	14.933	0.979

In Table 3 the correlogram is show which clearly indicating that the series is stationary, we can confirm this by unit root test.

Table 4
Unit Root Test for DL(Production) of Sesamum in Pakistan after Removing Outlier
Null Hypothesis: After_Removing_Outlier_DL production has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
A D F test statistic		-7.965653	0.0000
Test critical values:	1% level	-3.536587	
	5% level	-2.907660	
	10% level	-2.591396	

*MacKinnon (1996) one-sided p-values.

In Table 4 the p value = 0.0000 which clearly indicates that we reject the H_0 that series has a unit root. Thus, we are now able to satisfy both normality and stationary conditions for the production of Sesamum in Pakistan. With the help of correlogram we choose some appropriate models which are shown in table 5.

Table 5
Summary Values of AIC, Schwarz and Hannan-Quinn of Different Models for Production of Sesamum in Pakistan

S#	Model	R^2	Adj R^2	AIC	Schwarz	Hannan-Quinn
1	(1,1,0)	0.00016	-0.0321	-0.1079	-0.0075	-0.0683
2	**(0,1,1)**	**0.0002**	**-0.0321**	**-0.1079**	**-0.0076**	**-0.0683**
3	**(1,1,1)**	**0.06255**	**0.01645**	**-0.1164**	**0.0174**	**-0.0636**
4	(0,1,2)	0.01404	-0.0345	-0.0906	0.04323	-0.0378
5	(2,1,0)	0.01085	-0.0378	-0.0876	0.04625	-0.0348
6	(1,1,2)	0.03753	-0.0266	-0.0821	0.08518	-0.0161
7	(2,1,1)	0.03777	-0.0264	-0.0824	0.08487	-0.0164
8	(2,1,2)	0.03781	-0.0437	-0.0517	0.14904	0.02752
9	(0,1,3)	0.02772	-0.0371	-0.073	0.09429	-0.007
10	(3,1,0)	0.02327	-0.0419	-0.0689	0.09839	-0.0029
11	(1,1,3)	0.03806	-0.0435	-0.0519	0.14878	0.02726
12	(3,1,1)	0.03784	-0.0437	-0.0517	0.14901	0.02749
13	(2,1,3)	0.03807	-0.0614	-0.0212	0.213	0.07122
14	(3,1,2)	0.03796	-0.0616	-0.021	0.21312	0.07135
15	(3,1,3)	0.11798	0.00966	-0.0243	0.24337	0.08135

In Table 5 different fifteen models summary of AIC, Schwarz and Hannan-Quinn values are presented. For choosing the appropriate model among different models table 5 helps us. Model (0,1,1) has least values of Schwarz and Hannan-Quinn tests and model (1,1,1) has least AIC value. So we initially choose the model (0,1,1) and (1,1,1) because these have least values of AIC, Schwarz and Hannan-Quinn as compared to the all other models.

Table 6
ARIMA (0,1,1) Model

Dependent Variable: After_Replacing_Outlier_DL Production				
Sample: 2 66		Included observations: 65		
Variable	Coefficient	SE	t-Statistic	Prob.
C	0.044688	0.027462	1.627283	0.1087
MA(1)	-0.015717	0.114151	-0.137684	0.8909
SIGMASQ	0.047925	0.009021	5.312708	0
R-squared	0.000201	Mean dependent var		0.04466
Adjusted R-squared	-0.03205	S.D. dependent var		0.22064
S.E. of regression	0.224151	Akaike info criterion		-0.1079
Sum squared resid	3.115114	Schwarz criterion		-0.0076
Log likelihood	6.507821	Hannan-Quinn criter.		-0.0683
F-statistic	0.006243	Durbin-Watson stat		1.99452
Prob(F-statistic)	0.993778			

Table 6 shows that prob value for coefficient of constant term and the coefficient for MA(1) term have p values greater than 0.05 and are not significant. The value of Akaike information is -0.107933. The values of Schwarz criterion is -0.007577 and for Hannan-Quinn criterion is -0.068336 in model (0,1,1) which are least as compared to all other models. The value of Durbin-Watson statistics is 1.994516 which is quite reasonable.

Table 7
ARIMA (1,1,1) Model

Dependent Variable: After_Replacing_Outlier_DL Production				
Sample: 2 66		Included observations: 65		
Variable	Coefficient	SE	t-Statistic	Prob.
C	0.047289	0.017098	2.765692	0.0075
AR(1)	0.880524	0.180999	4.864793	0
MA(1)	-1	1056.287	-0.00095	0.9992
SIGMASQ	0.044936	1.057136	0.042507	0.9662
R-squared	0.062552	Mean dependent var		0.04466
Adjusted R-squared	0.016447	S.D. dependent var		0.22064
S.E. of regression	0.218821	Akaike info criterion		-0.1164
Sum squared resid	2.920847	Schwarz criterion		0.0174
Log likelihood	7.783169	Hannan-Quinn criter.		-0.0636
F-statistic	1.356748	Durbin-Watson stat		1.91477
Prob(F-statistic)	0.264492			

Table 7 shows that prob value for coefficient of constant term and the coefficient for AR(1) terms have p value less than 0.05, so are significant. But the value of MA(1) term has prob value greater than 5 percent and is not significant. The value of Akaike information is -0.116405 which is least as compared to all other model in the table. The value of Durbin-Watson statistics is 1.914771.

Table 8
ARIMA Model Fit Statistics of (1,1,1)

Model	NBIC	Ljung-Box Q(18)		
		Statistics	DF	Sig.
(1,1,1)	3.759	9.766	16	0.879

Table 9
ARIMA Model Fit Statistics of (0,1,1)

Model	NBIC	Ljung-Box Q(18)		
		Statistics	DF	Sig.
(0,1,1)	3.667	13.349	17	0.713

In Table 8 and 9 models fit statistic are shown for both models (1,1,1) and (0,1,1) respectively. The LB statistic is quite high for both models which give evidence that the series are white noise and model fits best. The NBIC is 3.667 for model (0,1,1) which is least, this leads us for choosing model (0,0,1) for forecasting purpose.

Now with the help of BDS test we will check linearity condition for (0,1,1) model if it fulfill than we can use this model for forecasting purpose.

Table 10
BDS Test for Residual of ARIMA Model (0,1,1)

BDS Test for RESID					
Sample: 1 66		Included observations: 66			
Dimension	BDS Statistic	Std. Error	z-Statistic	Prob.	
2	0.006566	0.007844	0.837145	0.4025	
3	0.011986	0.012559	0.954372	0.3399	
4	0.019288	0.015065	1.280342	0.2004	
5	0.023808	0.015816	1.505258	0.1323	
Raw epsilon		0.325975			
Pairs within epsilon		2983.000	V-Statistic	0.706036	
Triples within epsilon		145513.0	V-Statistic	0.529861	
Dimension	C(m,n)	c(m,n)	C(1,n-(m-1))	c(1,n-(m-1))	c(1,n-(m-1))^k
2	998.0000	0.495040	1409.000	0.698909	0.488473
3	680.0000	0.348182	1358.000	0.695341	0.336196
4	476.0000	0.251719	1313.000	0.694342	0.232430
5	344.0000	0.187978	1275.000	0.696721	0.164171

From table 10 it is clear that for all dimensions p value leads to do not reject our null hypothesis, so which fulfill linearity condition. Now we can use the model for further statistical analysis as it satisfies all prerequisite basic conditions.

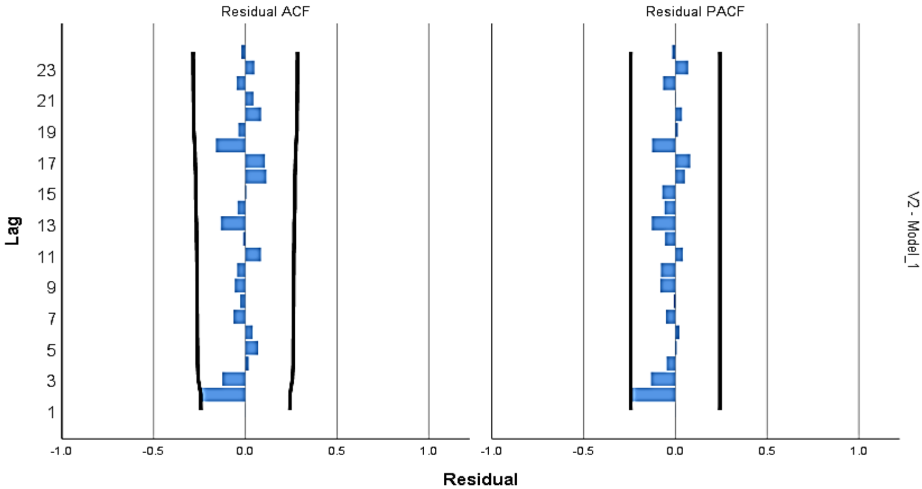


Figure 7: Residuals plot of PACF & ACF for ARIMA model (0,1,1)

In Figure 7 the Residual plot of PACF and ACF for the ARIMA model (0,1,1) are shown, we can clearly see that all the values are within the bounds.

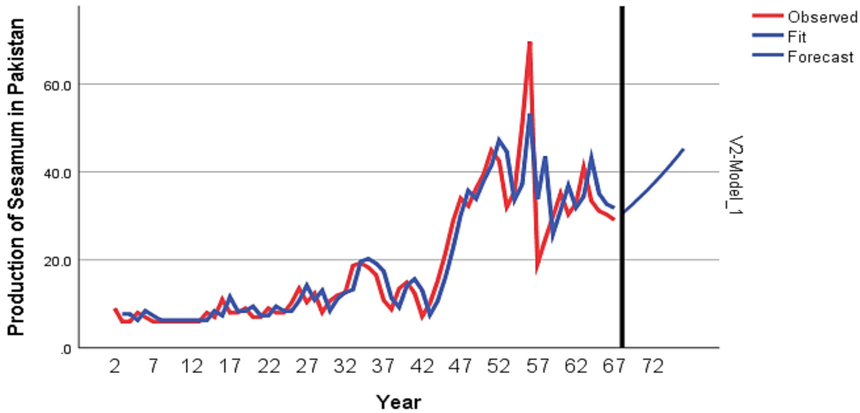


Figure 8: Plot of Observed, Fit and Forecast Values of ARIMA Model (0,1,1)

The forecasted, fitted and observed values for ARIMA model (0,1,1) are shown in the figure 8, which gives us clear picture of how well the model fits and forecasted production of Sesamum crop in Pakistan has an increasing trend in future.

Table 11
Cross Validated and Forecasted Values for Production of Sesamum Crop in Pakistan

Year	Observed	Forecasted (0,1,1)
2013-14	32.6	30.6
2014-15	34.3	32.1
2015-16	31.8	33.7
2016-17	34.1	35.4
2017-18		37.2
2018-19		39.1
2019-20		41
2020-21		43.1
2021-22		45.3

Hence the Table 11 shows cross validation and forecasted values for ARIMA model (0,1,1). From the Table 11, we conclude that there is an increase in the production of Sesamum crop in Pakistan in future.

SUMMARY AND CONCLUSION

Here in the study of production of Sesamum crop in Pakistan, we tried to find out an appropriate ARIMA model. We choose (0,1,1) and (1,1,1) models initially on the basis of AIC, Schwarz, Hannan-Quinn criterion and NBIC. As (0,1,1) model is chosen for forecasting because its residuals are linear and it has least values of Schwarz, Hannan-Quinn and NBIC. We also conclude that the production of Sesamum crop in Pakistan has an increasing trend in future.

RECOMMENDATIONS

Production of Sesamum crop can be increased by using best quality of seed and better management of the crop. Government has large agricultural forms in different cities of Pakistan, these can be used in Rabi season for oilseed crop like rapeseed and in kharif season for Sesamum. Because these are the native crop in this region and having many advantages over other oilseed crops. We can save a lot of foreign exchange by decreasing import of edible oil. We have one of the world's best irrigation system the need is to make new Dams to fulfill storing capacity of water.

REFERENCES

1. Ali, S., Badar, N. and Fatima, H. (2015). Forecasting production and yield of sugar cane and cotton crops of Pakistan for 2013-2030. *Sarhad J. Agric.*, 31(1), 1-9.
2. Amiri, E. (2015). Forecasting daily river flows using nonlinear time series models. *Journal of Hydrology*, 527, 1054-1072.
3. Anilakumar, K.R., Pal, A., Khanum, F. and Bawa, A.S. (2010). Nutritional, medicinal and industrial uses of sesame (Sesamum indicum L.) seeds-an overview. *Agriculturae Conspectus Scientificus*, 75(4), 159-168.

4. Badmus, M.A. and Ariyo, O.S. (2011). Forecasting cultivated areas and production of maize in Nigerian using ARIMA Model. *Asian Journal of Agricultural Sciences*, 3(3), 171-176.
5. Box, G. and Jenkins, G. (1970). *Time Series Analysis-Forecasting and Control*. San Francisco: Holden Day. 553 p.
6. Brock, W., Hsieh, D. and LeBaron, B. (1991). *Nonlinear Dynamics, Chaos, and Instability Statistical Theory and Economics Evidence*. MIT Press.
7. Broock, W.A., Scheinkman, J.A., Dechert, W.D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3), 197-235.
8. Clement, E.P. (2014). Using normalized bayesian information criterion (BIC) to improve box-jenkins model building. *American Journal of Mathematics and Statistics*, 4(5), 214-221.
9. Government of Pakistan, *Agricultural Statistics of Pakistan* (various issues) Islamabad, Ministry of Food, Agriculture and Livestock.
10. Government of Pakistan, *Pakistan Economic Survey* (various issues).
11. Gujarati, D.N. (2005). *Basic Econometrics* 4th edition. McGraw Hill. Higher Education: New Dehli, India.
12. Gujarati, D.N. and Porter, D. (2009). *Basic Econometrics*, Mc Graw-Hill International Edition.
13. Islam, T., Mohib, A.A. and Haque, S.Z. (2017). Econometric models for forecasting remittances of Bangladesh. *Business and Management Studies*, 4(1), 1-9.
14. Jaffar, A.H., Maqsood, H., Sarwat, Z., Muhammad, A., Habib, A. and Muhammad, A.I. (2016). Estimating the technical efficiency in rapeseed and mustard production: A case study of District Okara. *Germany Journal of Agricultural Economics*, 54(4), 801-811.
15. Kanzler, L. (1999). *Very fast and correctly sized estimation of the BDS statistic*. Available at SSRN 151669.
16. Khan, D., Ullah, A., Bibi, Z., Ullah, I., Zulfiqar, M. and Khan, Z.U. (2020). Forecasting area and production of guava in Pakistan: An econometric analysis. *Sarhad Journal of Agriculture*, 36(1), 272-281.
17. Kumar, M. and Anand, M. (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. *Studies in Business and Economics*, 9(1), 81-94.
18. Ljung, G.M. and Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.
19. Luukkonen, R., Saikkonen, P. and Teräsvirta, T. (1988). Testing linearity in univariate time series models. *Scandinavian Journal of Statistics*, 15(3), 161-175.
20. Mithiya, D., Datta, L. and Mandal, K. (2019). Time Series Analysis and Forecasting of Oilseeds Production in India: Using Autoregressive Integrated Moving Average and Group Method of Data Handling–Neural Network. *Asian Journal of Agricultural Extension*, *Economics & Sociology*, 30(2), 1-14.
21. Nath, B., Dhakre, D.S. and Bhattacharya, D. (2019). Forecasting wheat production in India: An ARIMA modelling approach. *Journal of Pharmacognosy and Phytochemistry*, 8(1), 2158-2165.

22. Nyoni, T. (2019). A Box-Jenkins ARIMA approach to the population question in Pakistan: A reliable prognosis. *Munich Personal RePEc Archive (MPRA*), Paper No.92434.
23. Rahman, N.M.F., Hasan, M.M., Hossain, M.I., Baten, M.A., Hosen, S., Ali, M.A. and Kabir, M.S. (2016). Forecasting Aus Rice Area and Production in Bangladesh using Box-Jenkins Approach. *Bangladesh Rice Journal*, 20(1), 1-10.
24. Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425), 208-218.
25. Umer, U.M., Sevil, T. and Sevil, G. (2019). Forecasting performance of smooth transition autoregressive (STAR) model on travel and leisure stock index. *The Journal of Finance and Data Science*, *5*(1), 12-21.
26. United Nation Department of Economics and Social Science (2019). World Population Projection. Retrieved From http://www.un.org/en/development/desa/news/population/2019-report.html
27. Yan, Z. (1990). Overview of rapeseed production and research in China. In *Proceedings of the International Canola Conference Potash and Phosphate Institute*, *Atlanta*, *Georgia* (pp. 29-35). Retrieved From https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/19116/IDL-19116.pdf?sequence=1
28. Zhang, H., Miao, H., Wang, L., Qu, L., Liu, H., Wang, Q. and Yue, M. (2013). Genome sequencing of the important oilseed crop Sesamum indicum L. *Genome Biology*, 14(1), 401.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 119-136

KERNEL FUNCTIONS FOR SEASONAL AND NON-SEASONAL TIME SERIES

Haris Khurram[1,2] and **Muhammad Mutahir Iqbal**[2]
[1] Department of Sciences and Humanities, National University of Computer and Emerging Sciences, Chiniot, Pakistan.
Email: haris.khurram@nu.edu.pk
[2] Department of Statistics, Bahauddin Zakariya University Multan, Pakistan.

ABSTRACT

Gaussian Process is a Bayesian non-parametric approach used for regression and classification purpose. This model is also used for time series forecasting. In this paper we have proposed some multi-featured kernels functions for versatile time series. These proposed kernel functions will be helpful in forecasting of seasonal and non-season time series. The empirical properties of these kernel functions are also discussed.

1. INTRODUCTION

Suppose, $x_t \in X_t$ is an input space and the mapping $f : X \to \mathbb{R}$ is the mapping from input to reals space then for series of interest y_t a time series model will be

$$y_t = f(X_t) + e_t \tag{1}$$

where X_t can be autoregressive lag series and e_t is disturbance term with $e_t \sim N(0, \sigma^2)$.

The term $f(X_t)$ is distributed as Gaussian Process. A Gaussian Process (GP) is a Stochastic process, indexed by a variable whose finite distribution is multivariate Gaussian. GP is not a distribution of random variable but a distribution over function with finite domain (Khurram and Iqbal 2019; Williams and Rasmussen, 2006). So we can write it as

$$f(X_t) \sim GP\left(m(X_t), g(X_t, X_t)\right) \tag{2}$$

where $m(X_t)$ is a mean function and $g(X_t, X_t)$ is a covariance function also known as kernel function. For autoregressive series this covariance function will be an autocovariance function. The posterior predictive distribution is also a Gaussian Process. Thus the posterior distribution is

$$p\left(y_{t+1} \mid y_t, X_{t+1}, X_t\right) \sim GP\left[E\left(y_{t+1} \mid y_t, X_{t+1}, X_t\right), COV\left(y_{t+1} \mid y_t, X_{t+1}, X_t\right)\right] \tag{3}$$

with mean function

$$E\left(y_{t+1}\right) = g(X_t, X_{t+1})'\left[g(X_t, X_t) + \sigma_n^2 I\right]^{-1} y_t, \tag{4}$$

and covariance function

$$COV\left(y_{t+1}\right)=g(X_{t+1},X_{t+1})-g(X_{t+1},X_t)\left[g(X_t,X_t)+\sigma_n^2 I\right]^{-1}g(X_t,X_{t+1}) \quad (5)$$

The kernel function has a key role and importance in the construction of a gaussian process and prediction by using it. Kernel function provides a matrix which is used as a covariance matrix in the Gaussian process. Since, the mean function and covariance function of gaussian process incorporates the kernel module so there is a need to define the structure, properties and flexibility of the covariance function in details (Zhong et al. 2013).

Consider a norm of input variable X_t defined on vector space $X_t \in \mathbb{X}$ with the inner product as $\|X_t\|_2=\sqrt{\langle X_t,X_t\rangle}$, then the metric between two sets of vectors X_t and X_t' is $d\left(X_t,X_t'\right)=\|X_t-X_t'\|_2$ defined on Hilbert space H. Thus, for two input variables X_t and X_t' such that $\left(X_t,X_t'\right)\in\mathbb{X}$, a mapping $g:\mathbb{X}\times\mathbb{X}\to\mathbb{R}$ where

$$g\left(X_t,X_t'\right)=\left\langle\phi\left(X_t\right),\phi\left(X_t'\right)\right\rangle_H \quad (6)$$

is called kernel function defined on Hilbert space H. And $\phi(\cdot)$ is the feature matrix of some defined input space (Fasshauer, 2011). This kernel function gives a matrix knows as kernel matrix which expresses the similarities between two sets of points. If two points are similar, then their corresponding functional points are also similar (Evangelista et al. 2007).

2. PROPERTIES OF KERNELS MATRIX

A kernel matrix is considered to be a covariance matrix of the Gaussian process if the kernel matrix has the following properties (Hofmann et al., 2008).

2.1 Symmetric and Invertible

A Kernel matrix $\boldsymbol{G}$ should be symmetric because of the basic property of a covariance matrix. Thus,

$$G=g\left(X_i,X_j'\right)=g\left(X_j',X_i\right)\ \forall\ i=1,2,...,n \text{ and } j=1,2,...,m. \quad (7)$$

Similarly, this covariance matrix should be invertible.

2.2 Positive Semidefinite

Symmetric and invertible kernel matrix is considered as a covariance matrix if it is at least positive semidefinite. In other words, kernel matrix $G=g\left(X_i,X_j'\right)$ is positive definite if for any vector $\boldsymbol{a}=a_1,a_2,...,a_n\in\mathbb{R}^n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j g\left(X_i,X_j'\right)>0 \quad (8)$$

and positive semidefinite

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j g\left(X_i,X_j'\right)\geq 0\,. \quad (9)$$

Theorem 1

A function $g:\mathbb{X}\times\mathbb{X}\to\mathbb{R}$ defined on some Hilbert space is

At least positive semidefinite if it is a kernel function.

A kernel if at least a positive semidefinite.

Proof:

Consider the spectral decomposition of the matrix $G=V\Lambda V'$ where $\boldsymbol{V}$ is the matrix based on eigenvectors "$\boldsymbol{v}$". Λ is the diagonal matrix with λ_i as the eigenvalues of this diagonal matrix.

For first, consider the g as a kernel. Then

$$\begin{aligned}\left\langle v, g\left(X,X'\right)v'\right\rangle_H &= \sum_{i,j=1}^{n} v_i g\left(X_i, X_j'\right)v_j \\ &= \sum_{i,j=1}^{n} v_i \left\langle \phi\left(X_i\right), \phi\left(X_j\right)\right\rangle_H v_j \\ &= \left\langle \sum_{i=1}^{n} v_i\phi\left(X_i\right), \sum_{j=1}^{n} v_j\phi\left(X_j\right)\right\rangle_H \\ &= \left\| \sum_{i=1}^{n} v_i\phi\left(X_i\right)\right\|_H^2 \geq 0\end{aligned}$$

Thus, g is positive semidefinite.

For the second, consider g as a positive semidefinite function then $\boldsymbol{V}$ is also positive semidefinite. Now assume $\phi\left(X\right)=\Lambda^{1/2}V$, then,

$$\begin{aligned}\left\langle \phi\left(X_i\right), \phi\left(X_j\right)\right\rangle_H &= \Lambda^{1/2}V\left(\Lambda^{1/2}V\right)' \\ &= V\Lambda V' \\ &= g\left(X_i, X_j\right)\end{aligned}$$

■

Consider $L^2\left(\mathcal{X}\right)$ as Hilbert space of square-integrable function on vector space $\mathcal{X}\in\mathbb{R}$. For a kernel g, consider a Hilbert-Schmidt operator $T_g : L^2\left(\mathcal{X}\right)\to L^2\left(\mathcal{X}\right)$ defined as

$$T_g f\left(X\right)=\int_{\mathcal{X}} g\left(X,X'\right)f\left(X'\right)dX' \qquad (10)$$

and this operator is positive semidefinite $\forall f\in L^2\left(\mathcal{X}\right)$ that is,

$$\int_{\mathcal{X}\times\mathcal{X}} g\left(X,X'\right)f\left(X\right)f\left(X'\right)dXdX'\geq 0 . \qquad (11)$$

Theorem 2

(Mercers) Consider g as a continuous kernel and Hilbert-Schmidt operator Tg has an eigenfunction $\psi_i \in L^2(\mathbb{X})$ with eigenvalue $\lambda_i \geq 0$, then $g(X, X')$ is as

$$g(X, X') = \sum_{i=1}^{\infty} \lambda_i \Psi_i(X) \Psi_i(X') \tag{12}$$

where the convergence is absolute and uniform, as

$$\lim_{n\to\infty} \sup \left| g(X, X') - \sum_{i=1}^{n} \lambda_i \Psi_i(X) \Psi_i(X') \right| = 0 .$$

3. CLASSES OF KERNEL FUNCTIONS

A kernel function is generally divided into the following categories (Genton, 2001).

3.1 Stationary Kernel Functions

A kernel function is said to be stationary if it is invariant to translation (shift-invariant) of input. Thus, for a kernel $g(X, X')$ defined on some input space is stationary if

$$g(X, X') = g(X - X') . \tag{13}$$

In other words, the kernel is depending on the difference between inputs. To be a valid covariance function this kernel function follow the property of positive semidefinite as define in eq. (9) as

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j g(X_i - X_j) \geq 0 . \tag{14}$$

A stationary kernel can be constructed from their spectral densities $\boldsymbol{S}$. By using inverse Fourier transformation one can convert the spectral density S into the positive semidefinite kernel function. So, one to one mapping from stationary kernel to finite measures can be achieved using inverse Fourier transformation by using the following theorem.

Theorem 3

(Bochner) A complex-valued positive semidefinite function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ is at least weekly stationary kernel if and only if it is the transformation of non-negative Borel measure $\boldsymbol{\mu}$. So, it can be represented as

$$g(X - X') = \int_{\mathbb{R}^d} e^{2\pi i (X - X')\omega} d\mu(\omega) . \tag{15}$$

where $\mu(\cdot)$ a probability measure with the associated spectral distribution of g. The one to one correspondence of this translation-invariant kernel with spectral density S implies that

$$g(X - X') = \int_{\mathbb{R}^d} e^{2\pi i (X - X')\omega} S(\omega) d\omega . \tag{16}$$

where

$$S(\omega) = \int_{\mathbb{R}^d} g(X - X')e^{-2\pi i(X - X')\omega} d(X - X'). \tag{17}$$

On the other hand, if the stationary kernel $g(X, X')$ depends on the Euclidian distance as

$$g(X, X') = g(\|X - X'\|), \tag{18}$$

then it is called isotropic stationary kernels with Euclidean norms. Yaglom (1957) derived the spectral representation of this isotropic stationary kernels with Euclidean norms by extending the above theorem.

Stationary kernels are used in defining the properties of the stochastic process where a variable is indexed over time. As time series modelling involves in this situation so we used these kernels for defining the autocorrelation functions.

3.2 Non-Stationary Kernel Functions

A non-stationary kernel function is dependent only on the input coordinates. The property of positive semidefinite of the nonstationary kernel function can be ensured by spectral density. The spectral representation of a nonstationary kernel $g(X, X')$ for some input point can be characterised by using the following theorem

Theorem 4

(Yeglom 1987) A complex-valued positive semidefinite function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ is nonstationary kernel if and only if it can be represented as

$$g(X, X') = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{2\pi i(X\omega_1 - X'\omega_2)} \mu(d\omega_1, d\omega_2). \tag{19}$$

The spectral density $S(\omega_1, \omega_2)$ is associated with the kernel function using generalized Fourier transformation. μ is a bounded and positive measure. If the function $\mu(d\omega_1, d\omega_2)$ is based on the diagonal where $\omega_1 = \omega_2$ then this representation reduces to the spectral representation of the stationary kernel defined in Eq. (15). Many of the nonstationary kernels have been used to model the time series data and to access the different variation of the dataset.

4. CONSTRUCTİNG OF NEW KERNEL THROUGH OPERATİONS

Kernel function can also be constructed by applying some specific operation on the regularly available kernels. This makes kernels more comprehensive and complicated which make it easier to capture the versatility of a data. Considered two valid kernel functions $g_1(X, X')$ and $g_2(X, X')$ which are covariance functions. The following operations for the kernel are valid.

4.1 Multiplication by a Scalar

Consider a scalar $c \in \mathbb{R}^+$, then for a kernel function $g_1(X, X')$

$$g(X, X') = c \cdot g_1(X, X') \tag{20}$$

using Eq. (9) and as $c \in \mathbb{R}^+$, thus

$$c \cdot a' g_1\left(X, X'\right) a \geq 0$$

which means that it is a kernel function.

4.2 Sum of Kernel

Considered the kernel function $g_1\left(X, X'\right)$ and $g_2\left(X, X'\right)$ their sum is also a valid kernel function

$$g\left(X, X'\right) = g_1\left(X, X'\right) + g_2\left(X, X'\right) \tag{21}$$

As $g_1\left(X, X'\right)$ and $g_2\left(X, X'\right)$ both follow the Eq. (2.17) thus

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[g_1\left(X_i, X_j\right) + g_1\left(X_i, X_j\right)\right] \geq 0$$

which show that it is a kernel function.

4.3 Product of Kernel

Considered the kernel function $g_1\left(X, X'\right)$ and $g_2\left(X, X'\right)$ their product is also a valid kernel function

$$g\left(X, X'\right) = g_1\left(X, X'\right) \cdot g_2\left(X, X'\right) \tag{22}$$

This product of kernel functions is the tensor product of two kernels matrices. As $g_1\left(X, X'\right)$ and $g_2\left(X, X'\right)$ both follow the Eq. (2.17) thus

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[g_1\left(X_i, X_j\right) \otimes g_1\left(X_i, X_j\right)\right] \geq 0$$

which show that it is a kernel function.

4.4 Composition of Function

Considered the function $\boldsymbol{f}$ of some input $\boldsymbol{X}$, then a kernel function is the composition of function.

$$g\left(X, X'\right) = f\left(X\right) \cdot f\left(X'\right) \tag{23}$$

As $g\left(X, X'\right) = \left\langle \phi\left(X\right), \phi\left(X'\right)\right\rangle_H$ and $\phi : X \rightarrow f\left(X\right)$ thus it is a kernel function.

4.5 Exponential of Kernel

Considered the kernel function $g_1\left(X, X'\right)$ is also a valid kernel function if it is in exponential form.

$$g(X, X') = e^{g_1(X, X')} \tag{24}$$

By using Taylor expansion, we have

$$e^{g_1(X, X')} = 1 + g_1(X, X') + \frac{1}{2}\left[g_1(X, X')\right]^2 + \cdots$$

which is the infinite series of sum and product of kernel function. Thus, it is also a kernel.

5. KERNEL FUNCTIONS

5.1 Squared Exponential Kernel Function

Squared exponential (SE) is one of the important and very commonly used kernel functions in kernel-based modelling. It is also known as redial based function (RBF) kernel. SE kernel function for input X_t is

$$g_{SE}(X_t, X_t') = \sigma^2 \exp\left(-\frac{\|X_t - X_t'\|}{2l^2}\right) \tag{25}$$

where $\boldsymbol{\sigma}$ is the scale and l is the length-scale parameters. This kernel function is an isotropic stationary kernel function. Additionally, the major feature of this kernel function is that it can be infinitely differentiable.

5.2 Rational Quadratic Kernel Function

Rational quadratic kernel (RQ) function is the scale mixture, or in other words infinite sum, of the SE kernel. RQ kernel function for series X_t is

$$g_{RQ}(X_t, X_t') = \left(1 + \frac{\|X_t - X_t'\|}{2\alpha l^2}\right)^{-\alpha} \tag{26}$$

where $\boldsymbol{\alpha}$ is the scale mixture parameter and l is the length-scale parameter. Both parameters $\alpha, l > 0$ and for $\alpha \to \infty$ the RQ kernel function will be converted into SE kernel function. This is also an isotropic stationary kernel function.

5.3 Periodic Kernel

The periodic (P) kernel was firstly used by MacKay (1998) to access the periodic repetition in data. For a series X_t the kernel is defined as

$$g_P(X_t, X_t') = \exp\left[\frac{-2}{l^2}\left[\sin\left(\frac{\pi}{w}\|X_t - X_t'\|\right)\right]^2\right] \tag{27}$$

where w is the periodic parameter, l is the length-scale parameters and $w, l > 0$. This also an isotropic stationary kernel function.

5.4 Linear Kernel

Linear kernel (L) itself not much important but it provides an attractive feature when it is combined with other kernels. So, a simple linear kernel for a series X_t is defined as

$$g_L\left(X_t, X_t'\right) = \left(X_t - C\right)\left(X_t' - C\right) \tag{28}$$

where C is an optional constant. Using only the linear kernel in the gaussian process is Bayesian regression modelling. It is a non-stationary kernel function.

5.5 The Matérn Class of Kernel Function

The Matérn (MP) class of kernel was proposed by Matérn (1960) is the generalization of RBF kernel function. For an input X, the general Matérn class of kernel function in the form of modified Bessel function is defined as

$$g_M(X_t, X_t') = \frac{1}{2^{1-v}\Gamma(v)}\left(\frac{\sqrt{2v}\left\|X_t - X_t'\right\|}{l}\right)^{v} K_v\left(\frac{\sqrt{2v}\left\|X_t - X_t'\right\|}{l}\right) \tag{29}$$

where $\boldsymbol{v}$ is the order parameter and l is the length-scale parameter with $v, l > 0$. K_v is the modified Bessel function and $\Gamma(.)$ is a gamma function. A simplified form of Matérn class of kernel is adopted by using $v = P + \frac{1}{2}$. This type of functional form is as

$$g_{M_p}(X_t, X_t') = \exp\left(\frac{\sqrt{2v}\left\|X_t - X_t'\right\|}{l}\right)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+1)!}{i!(p-i)!}\left(\frac{\sqrt{8v}\left\|X_t - X_t'\right\|}{l}\right)^{p-i} \tag{30}$$

For $P = 1$ we have $v = \frac{3}{2}$ which is a very common form of this kernel and written as

$$g_{M_1}(X_t, X_t') = \left(1 + \frac{\sqrt{3}\left\|X_t - X_t'\right\|}{l}\right)\exp\left(-\frac{\sqrt{3}\left\|X_t - X_t'\right\|}{l}\right), \tag{31}$$

and for $P = 2$ we have $v = \frac{5}{2}$ which is also a common form of this kernel as

$$g_{M_2}(X_t, X_t') = \left(1 + \frac{\sqrt{5}\left\|X_t - X_t'\right\|}{l} + \frac{5}{3}\left(\frac{\left\|X_t - X_t'\right\|}{l}\right)^2\right)\exp\left(-\frac{\sqrt{5}\left\|X_t - X_t'\right\|}{l}\right) \tag{32}$$

This simplified form is less than v times differentiable. Similarly, eq. (31) is at least one time differentiable while the form in Eq. (32) is at least two times differentiable. The eq. (31) and eq. (32) are a very popular choice for in data modelling. For $v < \frac{3}{2}$ the shape of function become rough and for $v \to \infty$ this kernel converted into RBF type kernels. Also, this type of kernel is the isotropic stationary kernel.

5.6 Wavelet Kernel Function

A wavelet (W) kernel function based on some mother wavelet function is considered by Zhang et al. (2004). For input X_t, the translation-invariant W kernel function is

$$g_W\left(X_t, X_t'\right) = \prod_{i=1}^{p} h\left(\frac{x_t - x_t'}{a}\right) \tag{33}$$

and for mother wavelet, it can be defined as

$$g_W\left(X_t, X_t'\right) = \prod_{i=1}^{p}\left[\cos\left(\frac{1.75\left(x_t - x_t'\right)}{a}\right)\exp\left(\frac{-\left\|x_t - x_t'\right\|}{2a^2}\right)\right] \tag{34}$$

where a is the parameter and x is the realization in one dimension (Antoniadis, 2006).

6. PROPOSED KERNEL FUNCTIONS

6.1 Proposed Kernel Functions for Non-Seasonal Series

As the data of crop production generally have some trend so we have defined four different kernel function for the crop production datasets which are multifeatured and able to capture maximum information of the data set. These four proposed kernel functions are as follows

6.1.1 Periodic with Linear Kernel

It is a simply defined kernel which is the sum of the periodic and linear $(P + L)$ kernel functions. This is a locally stationary kernel function. For the lag series, $L^l y$ the functional form of this kernel is as

$$\gamma_{P+L}\left(L^l y, L^l y'\right) = \sigma_f^2 \exp\left[\frac{-2}{l^2}\mathrm{Sin}\left(\frac{\pi\left\|L^l y - L^l y'\right\|}{w}\right)^2\right] + \left(L^l y - C\right)\left(L^l y' - C\right) \tag{35}$$

where σ, l, w and C are the hyperparameters and tune using an optimization algorithm. By using the property in eq. (21) it is clear that the proposed kernel is at least positive semidefinite.

6.1.2 Squared Exponential Periodic Linear Kernel

This kernel is the squared exponential times periodic time linear (SEPL) kernel function. This is also a locally stationary kernel function. For the lag series, $L^l y$ the functional form of this kernel is as

$$\gamma_{SE\times P\times L}\left(L^l y, L^l y'\right) = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|^2}{2l^2} - \frac{2}{l^2}\mathrm{Sin}\left(\frac{\pi\left\|L^l y - L^l y'\right\|}{w}\right)^2\right)\left(L^l y - C\right)\left(L^l y_t' - C\right) \tag{36}$$

where σ, l, w and C are the hyperparameters for LDGP and tune using an optimization algorithm.

As all these kernel functions are defined on Hilbert space and by using the property in Eq. (22) the proposed kernel is at least positive semidefinite.

6.1.3 Squared Exponential with Periodic Linear Kernel

This kernel function is the sum of squared exponential and periodic time linear ($SE + PL$) kernel function. This is also a locally stationary kernel function and for the lag series $L^l y$ the functional form of this kernel is as

$$\gamma_{SE+P\times L}\left(L^l y, L^l y'\right) = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|^2}{2l^2}\right) + \exp\left[\frac{-2}{l^2}\mathrm{Sin}\left(\frac{\pi\left\|L^l y - L^l y'\right\|}{w}\right)^2\right]\left(L^l y - C\right)\left(L^l y' - C\right) \quad (37)$$

where σ, l, w and C are the hyperparameters for LDGP and tune using an optimization algorithm.

As all these kernel functions are defined on Hilbert space and by using the properties in Eq. (21) and Eq. (22) the proposed kernel is at least positive semidefinite.

6.1.4 Squared Exponential Periodic with Linear Kernel

This kernel is the sum of the squared exponential times Periodic and linear ($SEP + L$) Kernel functions. This is not exactly stationary but is considered as the locally stationary kernel. for the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE\times P+L}(L^l y, L^l y') = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|^2}{2l^2} - \frac{2}{l^2}\mathrm{Sin}\left(\frac{\pi\left\|L^l y - L^l y'\right\|}{w}\right)^2\right) + \left(L^l y - C\right)\left(L^l y_t' - C\right) \quad (38)$$

where σ, l, w and C are the hyperparameters which will be tune using an optimization algorithm. As all these Kernel functions are defined on Hilbert space and by using the properties in Eq. (21) and Eq. (22) the proposed kernel is at least positive semidefinite.

6.2 Proposed Kernel Functions for Seasonal Series

We have defined seven different kernel function for the seasonal series which are multifeatured and able to capture maximum information of the series. These Kernel function can be used for any seasonal series. The proposed kernel functions are as follows

6.2.1 Squared Exponential Periodic Kernel

This kernel is the squared exponential times Periodic (SEP) Kernel. This is the stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE\times P}\left(L^l y, L^l y'\right) = \sigma_f{}^2 \exp\left(-\frac{\left\| L^l y - L^l y' \right\|}{2l^2} - \frac{2}{l_1^2}\left[\sin\left(\frac{\pi}{w}\left\| L^l y - L^l y' \right\|\right)\right]^2 \right) \quad (39)$$

where $\sigma, l, l1$ and w are the hyperparameters which will be tune using an optimization algorithm. Using the properties in Eq. (21) the proposed kernel is at least positive semidefinite.

6.2.2 Squared Exponential with Wavelet Kernel

This special kernel is the sum of the squared exponential and wavelet ($SE + W$) kernel where the mother wavelet is same as defined by Zhang et al. (2004). It is the locally stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE+W}\left(L^l y, L^l y'\right) = \sigma_f{}^2 \exp\left(-\frac{\left\| L^l y - L^l y' \right\|}{2l^2} \right) + \prod_{i=1}^{p}\left[\cos\left(\frac{1.75\left(L^l y - L^l y'\right)}{a}\right)\exp\left(\frac{-\left\| L^l y - L^l y' \right\|}{2a_1^2}\right)\right] \quad (40)$$

where σ, l, a and $a1$, are the hyperparameters which will be tune using an optimization algorithm. Using the properties in Eq. (21) the proposed kernel is at least positive semidefinite.

6.2.3 Rational Quadratic with Wavelet Kernel

This kernel is the sum of the Rational Quadratic and wavelet (RQ+W) kernel function. It is also a locally stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{RQ+W}\left(L^l y, L^l y'\right) = \left(1 + \frac{\left\| L^l y - L^l y' \right\|}{2\alpha l^2}\right)^{-\alpha} + \prod_{i=1}^{p}\left[\cos\left(\frac{1.75\left(L^l y - L^l y'\right)}{a}\right)\exp\left(\frac{-\left\| L^l y - L^l y' \right\|}{2a_1^2}\right)\right] \quad (41)$$

where α, l, a and $a1$, are the hyperparameters which will be tune using an optimization algorithm.

6.2.4 Squared Exponential Periodic with Matérn Class Kernel

This kernel is the sum of the Squared Exponential times periodic and Matern $(SEP + M2)$ kernel function. It is a stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE\times P+M2}\left(L^l y, L^l y'\right) = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|}{2l^2} - \frac{2}{l_1^2}\left[\sin\left(\frac{\pi}{w}\left\|L^l y - L^l y'\right\|\right)\right]^2\right) + \left(1 + \frac{\sqrt{5}\left\|L^l y - L^l y'\right\|}{l_2} + \frac{5}{3}\left(\frac{\left\|L^l y - L^l y'\right\|}{l_2}\right)^2\right)\exp\left(-\frac{\sqrt{5}\left\|L^l y - L^l y'\right\|}{l_3}\right) \quad (42)$$

where $\sigma, l, l1, w, l2$ and $l3$, are the hyperparameters which will be tune using an optimization algorithm.

6.2.5 Squared Squared Exponential Periodic Kernel

This kernel is the squared of Squared Exponential kernel times periodic (SSEP) kernel function. It is a stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE^2\times P}\left(L^l y - L^l y'\right) = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|}{l^2} - \frac{2}{l_1^2}\left[\sin\left(\frac{\pi}{w}\left\|L^l y - L^l y'\right\|\right)\right]^2\right) \quad (43)$$

where $\sigma, l, l1$ and w, are the hyperparameters which will be tune using an optimization algorithm.

6.2.6 Squared Exponential Squared Periodic

This kernel is the Squared Exponential kernel times squared of periodic (SESP) kernel function. It is also a stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{SE\times P^2}\left(L^l y, L^l y'\right) = \sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|}{2l^2} - \frac{4}{l_1^2}\left[\sin\left(\frac{\pi}{w}\left\|L^l y - L^l y'\right\|\right)\right]^2\right) \quad (44)$$

where $\sigma, l, l1$ and w, are the hyperparameters which will be tune using an optimization algorithm.

6.2.7 Squared Exponential with Squared Exponential Periodic

This kernel is the sum of the Squared Exponential and squared exponential times periodic (2SEP) kernel function. It is also a stationary kernel function. For the lag series, $L^l y$ the functional form of the kernel is as

$$\gamma_{2SE\times P}\left(L^l y, L^l y'\right) = 2\sigma_f^2 \exp\left(-\frac{\left\|L^l y - L^l y'\right\|}{2l^2}\right)\exp\left(-\frac{2}{l_1^2}\left[\sin\left(\frac{\pi}{w}\left\|L^l y - L^l y'\right\|\right)\right]^2\right) \tag{45}$$

where $\sigma, l, l1$ and w, are the hyperparameters which will be tune using an optimization algorithm.

7. RANDOM SAMPLE FROM GAUSSAIN PROCESS PRIOR

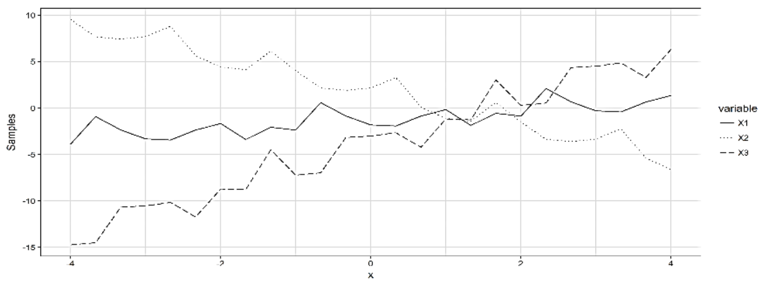
We have drawn three different random samples from gaussain process prior using proposed kernel functions. For modelling the data is simulated by assumming sample size 25. The response y is genrated by assuming the follwing model

$$y = \sin\left(X\right) + e$$

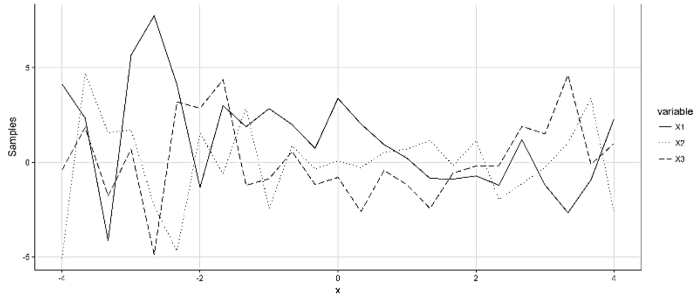
where e is the standard normally distributed disterbance term. X is taken as a seq from -4 to 4 of size 25.

Figure 1 shows the random samples of GP prior drawn from proposed season kernel function. There are four different kernels and from Figure 1 it can be assess that the Figure 1(a) and 1(c) are nonseasonal kernel function with linear trend. While the kernel functions in Figure 1(b) and 1(d) the sample from these kernels are nonseasonal and without linear trend.

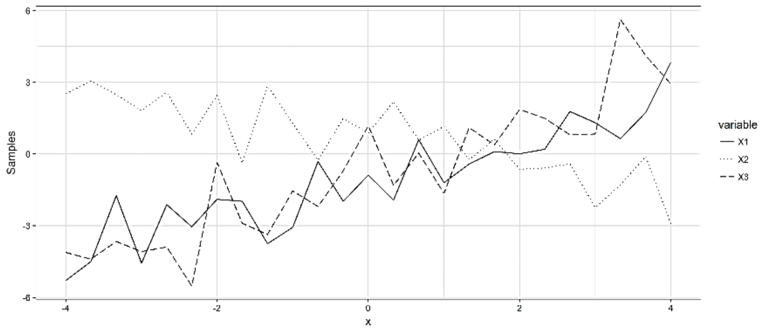
Figure 2 shows the three sample drawn from GP Prior using seven different proposed seasonal kernel function. The samples from these complex kernels are seem to be seasonal in pattern and are able to adjeted for some posterior GP.



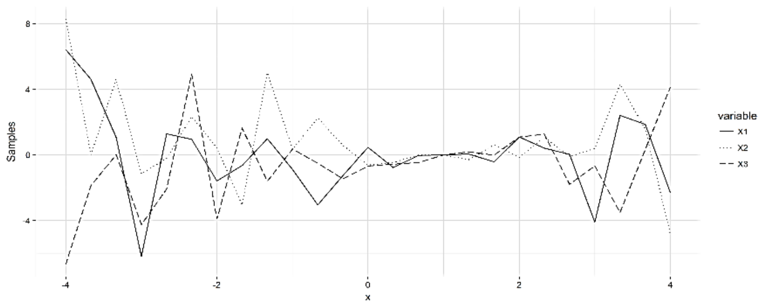
1(a)



1(b)

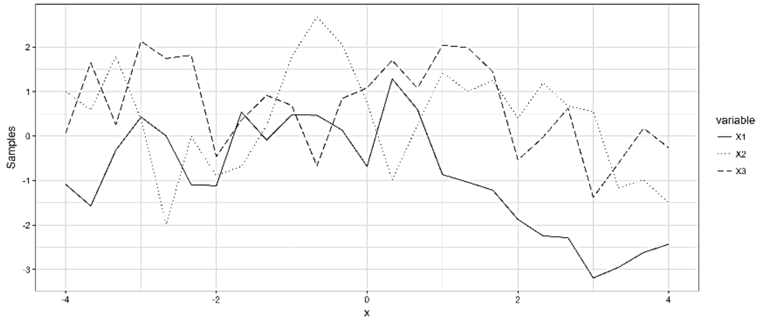


1(c)

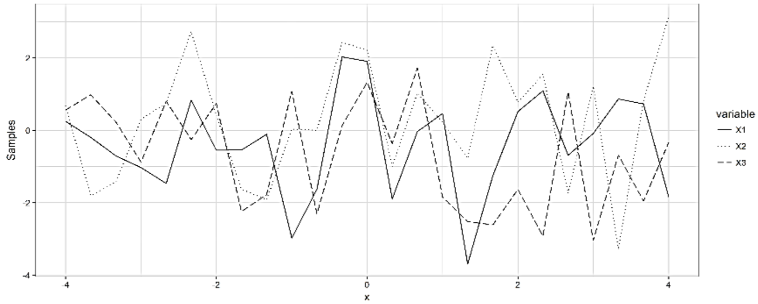


1(d)

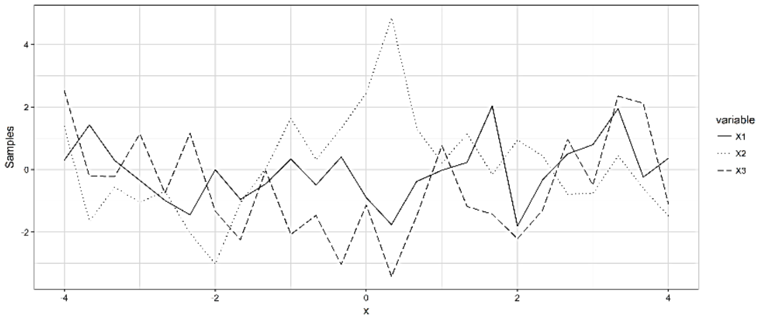
Figure 1: Three Random Samples Drawn from GP Prior using Each Non-seasonal kernel Function
a. Periodic with Linear Kernel;
b. Squared Exponential with Periodic Linear kernel;
c. Squared Exponential Periodic with Linear Kernel;
d. Squared Exponential Periodic Linear Kernel


Samples
x
variable
X1
X2
X3

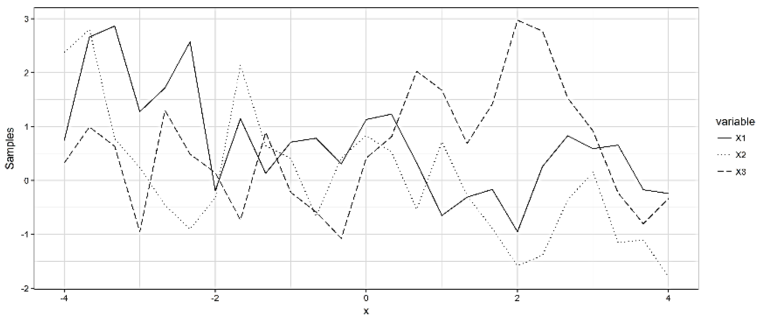
2(a)


Samples
x
variable
X1
X2
X3

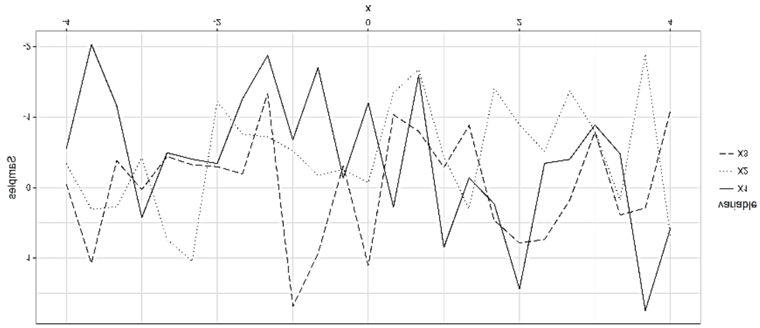
2(b)


Samples
x
variable
X1
X2
X3

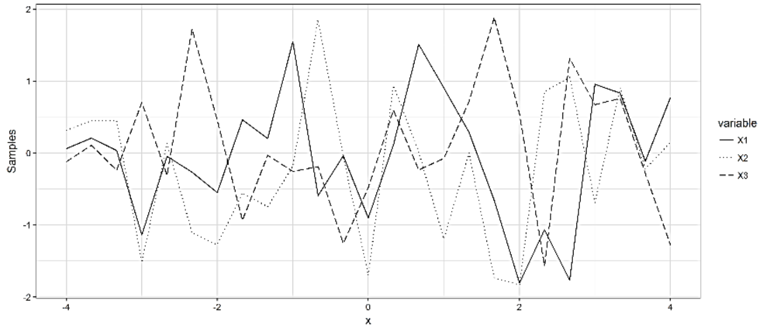
2(c)


Samples
x
variable
X1
X2
X3

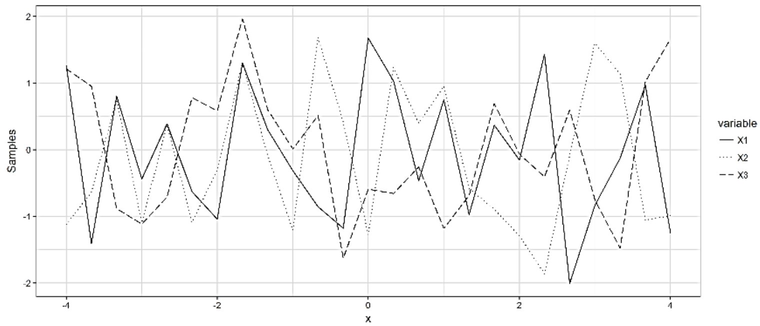
2(d)



2(e)



2(f)



2(g)

Figure 2: Three Random Samples Drawn from GP Prior Using Each Non-seasonal kernel Function
a. Rational Quadratic with Wavelet Kernel;
b. Squared Exponential Periodic with Matérn Class Kernel;
c. Squared Exponential with Squared Exponential Periodic;
d. Squared Exponential with Wavelet Kernel;
e. Squared Exponential Periodic Kernel;
f. Squared Exponential Squared Periodic;
g. Squared Exponential with Squared Exponential Periodic

4. COMMENTS AND CONCLUSION

In this work, we have present some of our proposed kernel functions. These kernel function can be used for modelling any data by using Gaussian process posterior predictive distribution. These kernel function can be used with simple Gaussian process and also can be used for time series modelling and forecasting using Gaussian process. These are multi-featured kernel which are able to adjust for any type of data set. Using these kernel can leads in efficient modeling and predication in general statistical modelling and for machine learning problem.

REFERENCES

1. Evangelista, P.F., Embrechts, M.J. and Szymanski, B.K. (2007). Some properties of the Gaussian kernel for one class learning. In *International Conference on Artificial Neural Networks* (pp. 269-278). Springer, Berlin, Heidelberg.
2. Fasshauer, G.E. (2011). Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation*, 4, 21-63.
3. Genton, M.G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2(Dec), 299-312.
4. Hofmann, T., Schölkopf, B. and Smola, A.J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3), 1171-1220.
5. Khurram, H. and Iqbal, M.M. (2019). Forecasting Potato Production in Pakistan Using Bayeian Non Parametric Modelling. *International Journal of Biology, Pharmacy and Allied Sciences*. 8(12), 2287-2295.
6. MacKay, D.J. (1998). Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168, 133-166.
7. Williams, C.K. and Rasmussen, C.E. (2006). *Gaussian Processes for Machine Learning*, Vol. 2, MIT Press Cambridge,
8. Williams, C.K. and Rasmussen, C.E. (2006). *Gaussian processes for machine learning*, vol. 2. MIT Press Cambridge,
9. Yaglom, A.M. (1957). Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory of Probability & Its Applications*, 2(3), 273-320.
10. Yuan, J., Wang, K., Yu, T. and Fang, M. (2008). Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression. *International Journal of Machine Tools and Manufacture*, 48(1), 47-60.
11. Zhang, L., Zhou, W. and Jiao, L. (2004). Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1), 34-39.
12. Zhong, S., Chen, D., Xu, Q. and Chen, T. (2013). Optimizing the Gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification. *Pattern Recognition*, 46(7), 2045-2054.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 137-143

HEALTH INFORMATION SYSTEMS AT HOSPITALS AT PESHAWAR & ABBOTTABAD, PAKISTAN

Imran Anwar Ujan[1], **Abdul Hamid Orakzai**[2] and **Asadulah Shah**[2]

[1] Department of IT, Faculty of Engineering & Technology,
University of Sindh, Jamshoro, Pakistan.

[2] KICT, International Islamic University, Malaysia

ABSTRACT

In the advancement of the healthcare industry, health information technologies, specifically the computerized provider order entry and electronic health records are deemed essential by consumers, lawmakers, and healthcare experts alike as the handling of information is a top priority in order to carry out proper healthcare.

The Survey be targeted health care professionals at public Sector hospitals of Peshawar and Abbottabad, KPK, Pakistan to investigate factors affecting health care professional's adoption of health information systems. The influence in using the quantitative approach is its applicability to the study, given that it is a very useful mechanism in both natural and social sciences.

The Questionnaire based on conceptual model. Seven-point Likert scale be used throughout the questionnaire. The survey from different hospitals of KPK conducted. The factors following factors are using and tested Behavioural Intention (BI), Perceived Usefulness (PU), Usability (UB), Cost effectiveness (CE), Internetwork (IN), Facilitating Conditions (FC), Performance Expectancy (PE), Compatibility (CMP), Complexity (COM), Data Security (DS), Data Privacy (DP), Hardware Modularity (HM), Software Modularity (SM).The legitimate reactions coded and dissected by utilizing Statistical Package for the Social Sciences (SPSS) version 25., descriptive statistics was performed.

INTRODUCTION

The HIS intends to gain ground in the current arrangement of medical services through the improvement of the work process as far as its reportage. The systems are intended to facilitate the medical care circumstance led under the HIS. Though, several literatures suggested that positive barriers were encountered for the implementation of health information systems, Shefter & Black noted some strengths and weaknesses of the application (Shefter, 2006, Black et al., 2011).

Healthcare professionals and decision-making bodies require several information such as (1) *health determinants* namely, socio-economic, environmental behavioral, genetic factors, together with the contextual surroundings where the health system operates; (2) *inputs* including policy and organization, health infrastructure, facilities and equipment, costs, human and financial resources, as well as health information systems; (3) Output, for example, accessibility, openness, quality and utilization of wellbeing data and administrations, responsiveness of the framework to client needs, and monetary danger assurance; (4) *outcomes* on mortality, mobility, outbreaks, and general health status; and (5)disparities in medical care, as far as inclusion on administrations and wellbeing results including key separation, explicitly, sex, financial status, nationality, geographic area, and other pertinent variables (Chaudhry et al., 2006). So, these factors will be used for further study in the context of private hospitals in Pakistan.

LITERATURE REVIEW

In the field of medicine, information is of utmost importance. Physicians are constantly processing information. Their practice involves the recording, collating, testing, and modifying information. The specific situation of clinical informatics is at the crossing point of data innovation, psychological science, man-made brainpower, and medication. Thus, this is definitely not a straightforward field including just a single perspective, for example, clinical processing, broadcast communications, or data designing, yet rather it is an exchange between doctors, patients, and clinical informaticians in clinical data frameworks and online data assets. It investigates and grows new information, assembles new hypotheses, and arranges standards and arrangements. Wellbeing data advancements and data frameworks are to emphatically build the result of clinical consideration. Computers, information systems and evidence-based decision makings are vital pillars of health informatics. To take full advantage of all ICT applications in health, physicians have to learn the skills for framing, analyzing and integrating the healthcare information into clinical practices (Pauker and Stahl, 1997; David Blumenthal, M.D., M.P.P, 2009).

It is essential in the research field of health informatics that a good fit persists in an ICT system and clinical practices (Chilana, P.K, et al., 2011). Previous studies identified the intellectual abilities as well as other skills required from medical practitioners who will utilize this technology***. (***Bhattacherjee, 2001; Brooks and Menachemi, 2006) In addition, studies conducted were able to identify the leading ICT professional in the different fields (Cohen, Manion and Morrison, 2000; Cooper and Weaver, 2003, Gefen, and Straub, 2000). The effectiveness of the instruction conducted for HIS operators were likewise investigated. More so, studies showed that in Health Information Technology (HIT) and HIS encompass various information technology applications used in the field of medicine.

RESEARCH METHODOLOGY

This study was carried out to determine the factors that affect the acceptance of health information systems, and to map out the correlations among these elements. Other theories and frameworks in technology acceptance have developed a health information system acceptance. Given its consistency to the topic, this study made use of the quantitative approach to be able to validate hypotheses in the models. Hussey and Hussey (1997) stated that the process of positivistic or quantitative approach is to review the literature and to form a theory and hypotheses.

The influence in using the quantitative approach is its applicability to the study, given that it is a very useful mechanism in both natural and social sciences. It enables the researcher to establish the reliability and validity of past researches on theoretical schemes and hypotheses – solely dependent on experiments (Patton, 1990; Blumberg, Cooper, and Schindler, 2005). Furthermore, this research method to explore the relationship between factors and the quantitative approach is highly recommended by Hussey and Hussey (1997).

The term population is defined as a whole group of people, occasions, or things that the researchers want to examine. For this research, the population be chosen from the hospitals of two major cities of KPK, Peshawar and Abbottabad. The functionality and reliability of appropriate technology is the main reason for this selection. Deciding among hospitals would rely on the emphasis of research, which relates to technicians and physicians who are acquainted with the utilization of technology.

RESULTS

Researcher personally visited the Leady Reading Hospital Peshawar, Hayatabad Medical Complex Peshawar, Khyber Teaching Hospital Peshawar and Ayoub Teaching Hospital Abbottabad for collecting the data. On the collected data descriptive statistics and correlation tests are applied using SPSS 25. There are 200 questionnaires were distributed among the healthcare professionals who are using the HMIS at their hospitals and returned responses were 170, in which 131 were male and 39 were female.

Gender					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Male**	131	77.1	77.1	77.1
	Female	39	22.9	22.9	100.0
	Total	170	100.0	100.0	

The age group of participants were 25-30 years having 80 which was large respondents, 31-35 have 58 , 36-40 was 24 and above 40 years was only 8 having lower respondents.

Age					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**25-30 Years**	80	47.1	47.1	47.1
	31-35 Years	58	34.1	34.1	81.2
	36-40 Years	24	14.1	14.1	95.3
	Greater than 40 Years	8	4.7	4.7	100.0
	Total	170	100.0	100.0	

When the researcher wants to see the experience side higher rate was 1-2 years having 56 and lowest was less than 1 year was 26 respondents.

Experience					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Less than one year**	26	15.3	15.3	15.3
	1-2 years	56	32.9	32.9	48.2
	2-4 years	44	25.9	25.9	74.1
	More than 4 Years	44	25.9	25.9	100.0
	Total	170	100.0	100.0	

In qualification demographic information Bachelor have 50, Masters were 90, MBBS degree holder are 16 and others qualification holder were 14. So highest level of participants was master's level.

Qualification					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Bachelor**	50	29.4	29.4	29.4
	Masters	90	52.9	52.9	82.4
	MBBS	16	9.4	9.4	91.8
	Other	14	8.2	8.2	100.0
	Total	170	100.0	100.0	

Experience					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Less than one year**	26	15.3	15.3	15.3
	1-2 years	56	32.9	32.9	48.2
	2-4 years	44	25.9	25.9	74.1
	More than 4 years	44	25.9	25.9	100.0
	Total	170	100.0	100.0	

The descriptive statistics of Behavioural Intention (BI), Perceived usefulness (PU), Usability (UB), Cost effectiveness (CE), Internetwork /(IN), Facilitating Conditions (FC), Performance Expectancy (PE), Compatibility (COMP), Data Security (DS), Data Privacy (DP), Hardware Modularity (HM), Software Modularity (SM). According to results all values of mean score are almost 5 which is acceptable level, which shows that healthcare professionals of Khyber Pakhtunkhwa (KPK) accept the health information systems in their hospitals.

Descriptive Statistics			
	N	Mean	Std. Deviation
BI1	170	5.4059	1.61537
BI2	170	5.7882	1.53534
BI3	169	5.5799	1.51023
BI4	169	5.7160	1.41077
BI5	170	5.7471	1.51154
UB1	170	5.1412	1.56631
UB2	170	5.3941	3.35352
UB3	169	5.1893	1.55445
UB4	169	5.0000	1.65112
UB5	170	5.1294	1.49388
UB6	170	5.2118	1.32874
UB7	170	5.1412	1.38592
UB8	170	5.2588	1.48884
PU1	170	5.4235	1.59790
PU2	170	5.6471	1.37351
PU3	170	6.2588	6.68506
PU4	170	5.7235	1.34121
PU5	170	5.8588	1.52419
PU6	170	5.7647	1.39862
PU7	170	5.8647	1.21111
CE1	170	4.8176	1.49426
CE2	170	4.8765	1.20749
CE3	170	4.8706	1.34814
CE4	170	4.9941	1.27561
CE5	170	5.0765	1.45139
CE6	170	5.1235	1.30635
INET1	170	4.8941	1.51148
INET2	170	4.9882	1.34555
INET3	170	4.8588	1.61467
INET4	170	5.0765	1.55758
INET5	170	5.5000	1.32901
FC1	170	5.1235	1.48046
FC2	170	4.8529	1.33528
FC3	170	4.7118	1.55924
FC4	170	5.1294	1.51746
FC5	170	5.2412	1.47787
FC6	170	4.8235	1.64029

Descriptive Statistics

	N	Mean	Std. Deviation
PE1	170	5.2765	1.53462
PE2	170	5.2118	1.41081
PE3	170	5.2765	1.37175
PE4	170	5.2059	1.34515
CMP1	170	4.9706	1.37787
CMP2	170	5.0941	1.15255
CMP3	170	5.0765	1.19666
CMP4	170	5.0765	1.42257
CMP5	170	4.6647	1.68159
CMP6	170	5.1059	1.41859
COMP1	170	4.1647	1.87065
COMP2	170	3.9882	1.85571
COMP3	170	4.1176	1.73316
COMP4	170	4.3000	1.70918
COMP5	170	3.9765	1.74720
COMP6	170	4.0176	1.78247
DS1	170	5.1294	1.56355
DS2	170	5.1706	1.42267
DS3	170	5.2647	1.28046
DS4	170	4.7529	1.46684
DP1	170	4.6647	5.82683
DP2	168	4.5417	2.66009
DP3	170	4.4588	1.69268
DP4	170	4.4765	1.58470
DP5	170	5.1353	1.28231
DP6	170	5.0294	1.41181
DP7	170	4.6647	1.63158
HM1	170	4.9529	1.64555
HM2	170	5.2059	1.34075
HM3	170	4.9412	1.64517
HM4	170	4.8471	1.25934
HM5	170	5.3529	1.21826
SM1	170	4.8235	1.65108
SM2	170	4.9647	1.34073
SM3	170	4.9000	1.49020
SM4	170	5.0412	1.51269
SM5	170	5.2353	1.39013
Valid N (listwise)	166		

CONCLUSION

Studies dependent on components influencing clients' acknowledgment of Health Information System by Healthcare Professionals in KPK was roused by the striking headway of innovation. This had all the earmarks of being one of the principal impulses for change in the healthcare area.

There is practically nothing, or few observational information presents in a creating economy like Pakistan about the variables that decide the utilization of the health data arrangement of the medical services experts.

REFERENCES

1. Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly*, 351-370.
2. Black, A.D., Car, J., Pagliari, C., Anandan, C., Cresswell, K., Bokun, T., McKinstry, B., Procter, R., Majeed, A. and Sheikh, A. (2011). The impact of eHealth on the quality and safety of health care: a systematic overview. *PLoS Med.*, 8(1), e1000387.
3. Chaudhry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S.C. and Shekelle, P.G. (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of* internal medicine, 144(10), 742-752.
4. Chilana, P. K., Ko, A. J., Wobbrock, J. O., Grossman, T. and Fitzmaurice, G. (2011). Post-deployment usability: a survey of current practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2243-2246).
5. Cooper, J. and Weaver, K. (2003). *Gender and Computers: Understanding the Digital Divide*, Lawrence Erlbaum Associates, Mahwah, NJ.
6. Gefen, D. & Straub, D. (2000). The relative importance of perceived ease of use in IS adoption: A study of e-commerce adoption. *Journal of the Association for Information Systems*, 1(8), 1-30.
7. Gefen, D. and Straub, D.W. (2004). Consumer trust in B2C e-commerce and the importance of social presence: experiments in e-products and e-services. *Omega*, 32(6), 407-424.
8. Hussey, J. and Hussey, R. (1997). *Business Research: A Practical Guide for Undergraduate and Postgraduate Students*, Macmillan, London.
9. Patton, M.Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, Inc.
10. Pauker, S.G. and Stahl, J.E. (1997). Medical informatics: where the action is. *Western Journal of Medicine*, 166(2), 148-150.
11. Shefter, S.M. (2006). Workflow Technology: The new frontier. How to overcome the barriers and join the future. *Lippincott's Case Management*, 11, 25-34.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 145-159

CHILD LABOUR AND THE ROLE OF PAKISTAN BAIT-UL-MAL

Mohammad Rafiq Khan and **Harris Jamil Alam Khan**
Lahore School of Economics, Lahore, Pakistan
Email: drrafiq@lahoreschool.edu.pk
hjalam@gmail.com

ABSTRACT

This study was conducted on the project titled, “Child Labour in Pakistan and the Role of Pakistan Bait-ul-Mal”. The purpose of this study was to investigate the effects of political and economic instability on the National Centres for Rehabilitation of Child Labour. These National Centres for Rehabilitation of Child Labour are supervised by the Pakistan Bait-ul-Mal. The paper also elucidates whether or not these National Centres can help curtail the problem of child labour or not. The paper sets out to explore, inter alia, the policy implications of the Pakistan Bait-ul-Mal. It sets out guidelines and recommendations that can help improve the efficiency and implementation of the policies that are in place to combat child labour. The data has been collected over a time span of thirteen years (1995-2008). The variables have been considered to check for the effectiveness of safety nets in combating child labour. Regressions have been run in order to check for the efficiency measures.

The results are unequivocal that there has been a marked improvement over the years. Pakistan Bait-ul-Mal has been considerably successful in curbing child labour. However, the study reveals that there is still room for improvement. The same can only be achieved if broad surveys are conducted by the government, tailor-made educational and awareness programs are introduced and a more transparency is introduced in implementation of policies.

Child labour is an evil that gnaws at the very roots of the society and that is why it is all the more important to combat it at every corner to make our country a better place for posterity

KEYWORDS

Bait-ul-Mal, Child Labour, Treasury, Rehabilitation of Child Labour, Education, Political, Economic Impact, Performance, Pakistan

1. INTRODUCTION

A child anywhere in the world has the God given right to be protected, nourished, loved and cared for but unfortunately the real picture is a harrowing blend of child slavery, labour and abuse. The focus of this study is, however limited to child labour. Child labour broadly defines children under twelve years of age who work on a regular basis for which they are paid or whose output is destined for the market. Child labour has been part of our society since the eighteenth century. It gained momentum during the Industrial Revolution and is mostly associated with the Third World countries. Although the process of globalization has called for measures to curb this evil, it is still prevalent around the world especially in South Asia.

The evil of child labour is rampant due to certain core factors such as failure of the educational system, the objectives of households to maximize present income etc. However, it is undeniable that the major factor is abject poverty which compels people to have large families and make them go out and earn their own livelihood. Since Asia contains the largest pool of child workers, it has all the same compelled it to pay special attention to its causes and consequences. Pakistan is one such country facing this problem and has actually charted out steps to tackle this problem. A distinguishing feature of child workers in Pakistan is that a large number of them are child domestic workers. These include children working as child minders, maids, cooks, cleaners, gardeners and general house-helps.

This study will be focusing on the measures and steps taken by the Pakistan Bait-ul-Mal (PBM) in combating this problem in the past years. Pakistan Bait-ul-Mal is one institution that is working earnestly to tackle this problem and has been able to control it to a certain extent. PBM was established in February 1992 under the provisions of Pakistan Bait-ul-Mal Act 1991 mainly to provide assistance to those in need and who are not covered by Zakat. It is administered by an autonomous board of management consisting of the chairman, five non-official and three official members. PBM has established more than thirty rehabilitation centres throughout the country to educate children withdrawn from work. Despite the limited capacity of these centres/schools they are playing an important role in the withdrawal and rehabilitation of children in labour.

Extensive legislation has been done all over the world to combat child slavery. Same has been the case in Pakistan where laws have been adopted and amended to facilitate the obviation of child slavery. The UN Convention on the Rights of the Child (CRC) Convention was adopted by the United Nations General Assembly on the 20th of November, 1989. The scale of this convention extends to persons up to the age of 18. Pakistan ratified this Convention in 1990 to become a part of the global effort in combating child labour. Article 32 of Convention of the Rights of the Child reads:

"State Parties recognise the right of the child to be protected from economic exploitation and from performing any work that is likely to be hazardous or to interfere with the child's education, or to be harmful to the child's health or physical, mental, spiritual, moral or social development".

Unfortunately, the enthusiasm with which these laws were adopted did not channel into actual output. Recent studies have not been conducted into the incidence of child labour and the impact the legislation for fighting child labour over the decades has not been properly assessed. However it is no hidden fact that child labour is still rampant in Pakistan.

The PBM primarily provides grant to NGO’s engaged in welfare projects under different strategies. National centres(s) for rehabilitation of child labour have been established country wide since 1995. The basic objectives of this centre is to wean away children from the hazardous child labour environment, impart primary education to them in conducive class room environments and afford opportunities for secondary and higher education to bring them into the mainstream of social milieu.

This research sets out to explore various activities envisaged by the PBM to eradicate child labour such as:

- Awareness raising
- Withdrawal of children from exploitive labour and rehabilitation through education and vocational training
- Community mobilization
- Situation Analysis
- Development of database on child labour
- Law enforcement
- Establishment of multipurpose complexes for vulnerable children drawn from work
- Empowerment of families and
- Poverty alleviation.

Apart from this, the study also sets out to highlight the steps that need be taken in near future. In doing so, the study will present a road map that may be taken up by the Pakistan Bait-ul-Mal in solving this problem. In general this article will focus on what has been done and what can be done for child labour rehabilitation in Pakistan.

2. REVIEW OF LITERATURE

A number of studies have been undertaken to thrash out its causes and effects and prescribe certain measures to combat this evil. Some of these studies have been carried out in Pakistan but most of them have been carried out in the developed countries. A brief outline of these studies and their implications are given below.

Suvarchala, (1992) studied the problem in an international perspective and elucidated on the legislative problem pertaining to the developing economies. The study highlighted that the child labour cannot be abolished in the developing economies, but instead, it can only be regulated as ‘large number of families are dependent upon this prevailing evil’.

Pierik and Houwerzijl (2006), lodged an enquiry and underlined the problem of child labour in the developing countries and concluded that it was due to the advent of globalization. The researchers discussed the policies formulated by the first world and their application in the third world in context of their appropriateness. They concluded that implementing western policies in the third world was not entirely efficient, and banning the products produced by child labor may prove counterproductive. The problem of child labour, in their opinion, could be reduced by collaborative measures. The study stressed on addressing the issue of global poverty in order to reduce child labor worldwide.

Bhalotra (2006) examined whether the child labour is compelled by poverty or not. The findings suggested that it was a work ethic problem as boys do tend to work in order to curtail the poverty compulsions. It found that the evidence for female child laborers remained ambiguous. A gender specific regression model was created for each regression. The results interpreted that the problem of female child labourers had significant relation with education as females are not sensitive to education. Finally, the study stressed upon the fact that if the return on education is increased this would help in instilling an education seeking attitude among the female child labour.

These are just some of the studies that amply outline the objectives, as well as, the constraints on mitigation of child labor globally. These studies also highlight the glaring fact that child labor is a global evil with almost the same roots and consequences and hence it would not be unwise to look at the countries which have successfully combated child labor and have designed legislated and implemented projects to achieve satisfactory results. The same was the context of the work being reported here. Moreover, this work was due as no work seems to have been reported in literature.

3. GOAL AND OBJECTIVES OF RESEARCH

The goal of research undertaken here was the assessment of the performance and the effectiveness of the National Centres for Rehabilitation of Child Labour (NCRCLs) with PBM support and for successful implementation of the child labour laws in Pakistan.

The objectives of the research were as follows:

1. Assessment of the proposed schemes whether they are successful or not.
2. Examination of extreme paradigms to know whether an outright ban on child labor would be an effective policy tool.
3. Identification of core problems to highlight the role of PBM regarding its expenditure and donation to different National Centers
4. Assessment of the centers whether these have been successful in raising the required finance to fight the problem.

4. RESEARCH METHODOLOGY

Multiple research techniques were used which required a well-defined theoretical framework and both primary and secondary data.

4.1 Theoretical Framework

In order to assess the growth of NCRCLs, relationship between the former and the political, economic and social stability of the government was studied by regression analysis. Thus the growth of the NCRCLs was defined as dependent variable and latter were defined as the independent variables. The timeframe was from the establishment of NCRCLs to the year of enquiry i.e.1995- 2007 and their growth relating to the growth of the country's economy per year.

The study tends to highlight a descriptive study on how political repression and economic stability can increase the effectiveness of safety nets such as PBM. Thus NCRCLs were defined as dependent Variable while political and economic repressions were defined as independent variables.

The equation for the Multiple Linear Regression was:

$$Number\ of\ NCRCLs = \alpha + \beta Political\ repression + \gamma Economic\ repression$$

Hypothesis 1

H_0 = The NCRCLs are not affected by political activity.

$H_1 \neq$ The NCRCLs are affected by political activity.

Hypothesis 2

H_0 = The NCRCLs are not affected by economic activity.

$H_1 \neq$ The NCRCLs are affected by economic activity.

Similarly, it may be statistically checked whether or not the presence or creation of NCRCLs, safety nets of the Government are efficient in reducing child labour. Data of the last thirteen years has been considered to this end. Here the dependent variable was the Child Labour while independent variables were Number of students and number of NCRCLs

The equation to be tested for this regression is:

$$Child\ Labour = \propto + \beta(No. of\ Students) + \gamma(No. of\ NCRCLs)$$

Hypothesis 3

H_0 = Increase in the Number of students enrolled in NCRCLs cannot help reduce child labour.

$H_1 \neq$ Increase in the Number of students enrolled in NCRCLs can help reduce child labour.

Hypothesis 4

H_0 = Increase in the Number of NCRCLs cannot help reduce child labour.

$H_1 \neq$ Increase in the Number of NCRCLs can help reduce child labour.

4.2 Secondary Data

The secondary data were collected from different government statistics institutions and from Statistics Section of Bait-ul-Mal itself. The statistics procured form the PBM and other institutions was consolidated and verified to check its credibility. Data for the past ten years was obtained from the PBM. Other statistical sources included World Bank reports, newspaper articles, multiple online, commentaries, history, encyclopaedias, bibliographies, textbooks and already existing thesis done on the study of child labour.

4.3 Primary Data

The primary data was collected through surveys focused on the administrative staff of the institutions of six National Centres for rehabilitation of child labour whom Bait-ul-Mal provides funding in various forms to gather their views on the performance of the Bait-ul-Mal over the years. Subsequently, the responses of child labourers were gathered. The students included ex-students of the National Centres and the existing students to know PBM's performance and the efficiency for providing support to the students and their families. Well designed and pre-checked questionnaires were used to get responses which were handed over to the respondents in person. The responses of executives of PBM provided an insight into the performance of the PBM from the National Centres' point of view while those. From their staff and PBM itself helped provide an in-depth analysis of the working of the institution and its overall efficiency.

The interviews were held in person. Those who could not make themselves available in person were interviewed via telephonic conversation. The surveys contained both structured and unstructured questions (closed ended and open ended). The data were tabulated and analyzed.

4.4 Analysis of Data

The data were processed using different statistical software's; Microsoft Excel and Minitab for tabulation, SPSS, E-views and Stat-graphics for coding and frequency and regression analysis.

5. RESULTS AND DISCUSSION

The results are reported at two levels. At the first level, the information collected through the interviews is reported as descriptive research while at the second the results of the quantitative statistical analysis are reported.

5.1 Responses of Administrative Staff

Twelve members of the administrative staff from 6 national centres were asked questions aimed at the performance of PBM. When asked whether the environment was conducive enough to carry out work effectively, the administrative staff gave a 100% positive reply as PBM had created a good and workable environment. When the respondents were asked about their pay scale being satisfactory 75% of the respondents were not satisfied and 25% were so. On scrutiny it was found that it was satisfactory at start but due to the economic instability of the country and the rising standards of living the income received was becoming unsatisfactory. When asked about the ease of doing desk and field jobs and meeting of administrative needs by PBM the reply from all was 100% positive. This showed that the level of communication between PBM and national centres was good. When asked about the facilities being provided in time and the overall experience with the PBM the respondents again gave a 100% positive answer. When asked about the projects being completed well in time, answer was 83.3% positive and 16.7% negative. The delay was due to the fact that children are tough to teach and get along with at their primary level and have to be rehabilitated first before they can be taught; this process is cumbersome and time consuming. In response to the question whether the amount of funds was adequate 75% disagreed and 25% agreed. This revealed that the entire cost of the project was increasing and thus causing a problem for the effective allocation and utilization of funds. When the respondents were asked about the project coordination or independent execution in the national centres, 83.3% replied in positive and 16.7% were not sure about the answer. Further enquiry revealed that they were not sure whether they did know what the overall curriculum was but the uncertainty was there due to different districts having different agendas and those could only be looked at by the PBM. When asked about any improvement they would suggest to the PBM 50% said no improvements were required, 25% said that there was a need for an increase in the pay scale, 8.3% suggested an increase in the budget allocation to the NCRCLs and 16.7% said that the enrolment should be increased, as the schools could accommodate even more students.

5.2 Responses of the Ex-Students of National Centres for Rehabilitation of Child Labour

The responses of ex-students against queries about the benefits of the training they had received, in-time provisions of what the students needed, care of all expenses by PBU or spent by themselves, subsistence income given to the parents of the students, their happiness at the education imparted to them and the PBM's support to them in the matters of further education were 100% positive. This provides evidence that the PBM has been providing adequate support to the current students as well as to the ex-students, who are now in different secondary schools. On questioning about the mode of instruction being satisfactory, the ex-students gave 88% positive and 12% negative reply. Of course, on informal questioning they said that some teachers did not pay full attention. When asked had the training helped them better understand the outside world to interact with it in a better way, 84% said yes while 16% were not sure indicating that the knowledge they had been imparted was satisfactory for them. When the ex-students were questioned as to whether the teachers and the administration were accessible to them at all time 92% replied in positive and 8% said that they were not accessible every time. When asked whose choice was it to continue their education to the secondary level (parents, own choice or both), 44% said that it was their parents choice, 44% said it was their own choice and 12% said that it was a mutual decision. This question was included to estimate the role of parents in their education and elimination of child labour and also to know whether it is the greed of the parents or their will to educate their children. The question was set in their research enquiry by Lopez-Calva and Miyamoto (2004) as whether it was the duty of the parents to educate their child or it was "Filial Obligation". Their results indicated that individuals faired as "purely self-interested", and therefore altruism was left out of the study but in fact altruism is quite common in the cultural realm of Pakistan. Therefore, it cannot be excluded from the analysis. The scope of the study revolved around the national centres effectiveness and not around the causes of child labour. Due to time limitations the interviews with the parents were not conducted which could have led to clarifications regarding the question of filial obligations, duty or self-interest on part of the parents or the children. When the ex-students were asked about their views on educating themselves further after the completion of their secondary education, 60% were sure that they would but the 40% were not. On further questioning they replied that they were the breadwinners and the family needed more income for that purpose.

5.3 Responses of the Existing Students of National Centres for Rehabilitation of Child labour

When the current students were asked about the incentives being provided to them and the provisions being adequate enough, the respondents gave a 100% positive reply. Their responses when consolidated with the rest indicated that the performance of the PBM was efficient. When they were questioned about their mode of instruction being satisfactory, 78% gave a positive reply while 10% said no and 12% didn't know. When they were asked whether the traditional skills being taught were helpful or not, 50% gave a positive reply, 30% gave a negative response and 20% did not know. When asked about whether the teachers and administrative staff were friendly and cooperative or not, 70% gave a positive reply while the 30% disagreed. They were further questioned whether environment of the centres were favourable to them to carry out their tasks and work effectively, 48% said that it was good while 52% disagreed. When questioned about the curriculum and the skills they were being taught would help them in the outside world 46% agreed while the 54% were not sure about the education helping them in the outside world. When the students were questioned whether the education standard was the same as the curriculum being taught to other primary education schools, 32% said yes while the 68% were not sure that they were being taught the same course. When asked that the students wanted to educate themselves further after completing their studies in this national centre, 62% said yes, 32% said no and 6% did not know. When the students were asked whether their parents wanted them to continue their education, 72% said yes and 28% said that they did not know. The only problem that could be traced back was whether the parents were forcing their children to work, or the child himself was not yet fully rehabilitated to understand that bonded labour was a danger.

5.4 Results of Regression

Analysis 1

Table 1
Output of Regression Analysis 1 (Included Observations: 13)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	85.57143	20.87470	4.099289	0.0021
POLREP	5.857143	26.94912	0.217341	0.8323
ECOREP	-53.14286	26.94912	-1.971970	0.0769
R-squared	0.315432	Mean dependent var.		67.38462
Adjusted R-squared	0.178518	S.D. dependent var.		49.75362
S.E. of regression	45.09450	Akaike info criterion		10.65457
Sum squared resid.	20335.14	Schwarz criterion		10.78494
Log likelihood	-66.25472	F-statistic		2.303877
Durbin-Watson stat	0.513656	Prob. (F-statistic)		0.150343

The output of regression model is shown in Table 1.

The output shows the results of fitting a multiple linear regression model to describe the relationship between NCRCLs and 2 independent variables. The equation of the fitted model is Number of NCRCLs = 85.57143 +5.857143*Political Repression - 53.14286*Economic Repression.

The R-Squared statistic indicates that the model explains 31.5432% of the variability in NCRCLs. The adjusted R-squared statistic, which is more suitable for comparing models with different number of independent variables, explains only 17.8518% variability. The standard error of the estimate shows the standard deviation of the residuals to be 45.0945. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in data file. Since the DW value is less than 1.4 due to presence of the dummy variables, there may be some indication of serial correlation.

To test Hypothesis 1 let us take use P-value which is 0.8323, for Political Repression. Since it is greater than 0.10, the relationship between NCRCLs and political repression is not statistically significant at the 90% confidence level. Therefore, the Alternative Hypothesis H_1 is rejected. Similarly, Hypothesis 2 can be checked. In this case the P-value is 0.0769 which is less than 0.10 indicating that the relationship between NCRCLs and economic repression is statistically significant at the 90% level and thus alternative hypothesis H_1 is accepted.

As observed in Fig.1 that from 1995 till 1999 both economic and political stability bring about rapid growth in the national centres for rehabilitation of child labour and the growth from year to year takes place in leaps and bounds, but from the years 1999-2000 growth of the NCRCLs becomes stagnant. This was the period when the first martial law was declared since the establishment of the NCRCLs wing of the PBM. The period showed high political instability and economic downturn. The government was directly influenced by such a political endeavour which caused not only the political situation to become unstable but also led to the volatility of the economic activity in the country. In 2001 the events of 9/11 led to an increase in remittances, foreign aid and loans worth 12 billion dollars which kicked the economy (Khalid Bhatti, 2007.). The study sees that the foreign aid and remittances restored the economic stability in the country and this led to a jump in the yearly growth of NCRCLs in the country. The year 2007-2008 was a period of strong economic and political unrest. The emergency declaration caused a great disturbance in the Country, which also led to the increase in the economic instability. The growth of NCRCLs became stagnant throughout these years again. The factors such as oil price rising, the value of rupee declining and the cost of living skyrocketing could easily defeat the purpose of NCRCLs because it is observed that in times like these child labour takes preference over the NCRCL programme, as the subsistence allowance provided by PBM is not enough to meet the day to day living expenses of the poor and needy.

The physical evidence in this study and other researches suggests that if the incidence of child labour increases the opportunity of adult labour decreases. This starts a chain reaction that in the end affects the economy of the country. The political influences somehow allow child labour rather than solving the problem.

5.5 Results of Regression

Analysis 2

Table 2
Table 1: Output of regression analysis 2
(Included observations: 13)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1771033.	275855.3	6.420151	0.0001
Students	-216.1233	250.7759	-0.861819	0.4090
NCRCLS	30979.30	26265.29	1.179477	0.2655
R-squared	0.474586	Mean dependent var		2231186.
Adjusted R-squared	0.369503	S.D. dependent var		639766.6
S.E. of regression	507999.4	Akaike info criterion		29.31352
Sum squared resid	2.58E+12	Schwarz criterion		29.44390
Log likelihood	-187.5379	F-statistic		4.516298
Durbin-Watson stat	2.574224	Prob(F-statistic)		0.040041

The equation of the fitted model is:

*Child labour force = 1.77103E6 - 216.123*Number of students + 30979.3*Number of NCRCL*

The output of the Regression Analysis 1 is displayed in Table 2. Since the P-value in the ANOVA table (0.0001) is less than 0.05, there is a statistically significant relationship between the variables at the 95% confidence level.

The R-Squared statistic indicates that the model explains 47.4586% of the variability in child labour force. The adjusted R-squared statistic, explains 36.9503% variation. The standard error of the estimate shows the standard deviation of the residuals to be 507999. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in the data file. Since the DW value is greater than 1.4, there is probably not any serious autocorrelation in the residuals of the students enrolled and the number of NCRCLS

The P-value corresponding to the independent variables to number of students is 0.4090. Since it is greater than 0.10, the relationship is not statistically significant at the 90% level of confidence. Thus the model is not a good fit and is rejecting the alternative hypothesis H_3. Similarly, the p-value corresponding to NCRCLS is 0.2655 which is greater than 0.10, Thus the relationship is not statistically significant at the 90% level of confidence. Therefore the null hypothesis $H_4$is rejected and the alternate hypotheses accepted.

5.6 The Hypothesis Discussion and Limitations

Regression 1:

The Alternative hypothesis H_1 was rejected on the basis of the multiple linear regression. However, the physical evidence showed that there was a relationship between economic and political activity on the performance of the PBM as the expansion of the NCRCLs became stagnant when faced with unfavourable circumstances. This could have been due to the presence of external factors due to which the variables could not sort out the problem significantly. The situation may be compared with that reported by Richard Jong-A-Pin (2006). His methodology could be used to calculate the amount of political repression as an index he has created. This would give an insight on how political repression affects the economy and how it can be effectively calculated for defining the variables more accurately. The same can be done by consulting results of other researches on economic repression saying that a new index should be created to better measure the effects on the economy.

Regression 2:

This was rejected due to the fact that the actual census data was not available and growth rates were used to estimate the number of child labourers that would be present at a given year.

The limiting factors that rejected the hypothesis are, the timeframe, as it was not enough data to establish a relationship between the variables. It is also due to the fact the Pakistan Bait-ul-mal focuses on bonded labour rather than child labour as a whole. Therefore, the amount of child labour may be increasing the amount of bonded labour is decreasing. This can also be contributed to the fact that there is no actual survey or census taking place that measures the amount of bonded labour taking place in Pakistan.

Policy Discussion

PBM needs to expand its horizon over the invisible and concentrated child labour and start to focus over the Concentrated and visible child labour, dispersed and visible child labour and Dispersed or invisible child labour. The Pakistan Bait-ul-mal needs to see that as the amount of bonded labour is decreasing the amount of the dispersed and visible child labour is increasing. This could pose a threat in the future if the Pakistan Bait-ul-mal focuses on the bonded child labour needs only. The selection criteria should be changed to the places. Pakistan Bait-ul-mal currently focuses on only one area which is not feasible with respect to the population growth rate of Pakistan. A quick and effective policy measure should be executed to stop the growth of the visible child labour.

The number of national centres needs to be increased. Considering the number of children who are engaged in child labour in Pakistan, the present number of centres is not enough to wean away children from this evil. Therefore, efforts are required to establish more centres countrywide so as to curb this evil as much as possible. The current policy that is being dictated by the quota set by the government limiting the number of National Centres should be abolished. This problem is in correlation with the enrolment issue of the national centres, the current strength of students at any centre is 120 children only. Steps and measures are needed to increase the strength of students and enroll in more children to curtail most hazardous forms of child labour. This is only possible if the centres have appropriate infrastructure to accommodate more children.

One of the key policy measures that need to be taken into consideration by the Bait-ul-mal is the need to advertise their role in curtailing child labour. The sensible use of right words in their advertisement campaign can make a lot of difference. For instance, physical and mental exploitation of children can be eliminated by providing a conducive environment for children at Bait-ul-mal. The right kind of publicity campaign can be beneficial as well. Bait-ul-Mal needs to highlight their role in curbing this evil that is slowly eating up our society. The role of a firm advertising campaign is highly related to the interest the private parties would take into this matter. Upon an informal interview with the Deputy Director of the Punjab Bait-ul-Mal, it was found that the there were once two parties interested in giving small donations. But a Management Committee of the Pakistan Bait-ul-mal has to be approached and an approval is sent to confirm the acceptance of the donation. The lengthy and cumbersome process does not help the cause at all. Pakistan Bait-ul-mal needs to engage private parties that can help expand the number of centres, increase the enrolment and cater to the needs of the children. These parties can help Pakistan Bait-ul-mal increase their opportunities, by expanding their education level to secondary education, which has been on the policy agenda for past three years without any sight of implementation. The problem that Pakistan Bait-ul-mal also needs to resolve is the number of the governing bodies controlling the activities of the institutes. The Punjab Bait-ul-mal is having management issues with controlling all the national centres established in Punjab.

The current budget of Bait-ul-mal is being taken into consideration which is being increased from the usual 1.9 million rupees per national centre to 2.5 million rupees. The problem is that even when the budget was 1.9 million, the expenditure was always underutilised; the administrative staff of the National Centres of Rehabilitation of Child Labour 75% demanded a better pays scale and they also added that the overall cost of the project was increasing. The *Figure.2.F* shows that the expenditure has always been underutilised. But considering the economy and the rising inflation rates in Pakistan the increase in the budget would help the future of these national centres with the administrative satisfaction and the increase in the overall benefits of the children studying in the National centres for Rehabilitation of Child Labour

6. CONCLUDING REMARKS

It is no hidden fact that child labour is rampant in Pakistan. It thrives in the province of Punjab. The Pakistan Bait-ul-Mal has been successful in combating child labour to some extent but it has still to go a long way to eradicate child labour. Ample studies have not been conducted and lack of data continues to hide the true picture. Not to mention the fact that a plethora of legislative measures can be taken but they will not help the cause if they are not properly implemented.

First and foremost it is necessary to identify the root and try to uproot the problem with some basic decisive measures which I have broadly divided into five glaring recommendations. These recommendations are as follows:

1) Proper procedures should be introduced to effectively implement the legislation that has been carried out to eradicate child labour.
2) These procedures should include but are not limited to better quota allocation, in depth investigations by a child labour task force, reporting any incidence and taking decisive action against it.
3) Introducing legislation that levies severe fines and effective punishment on anyone promoting or furthering child labour
4) Introduction of third parties and NGOs to fund, implement and further the cause of combating child labour
5) Household services should be provided to identify the problems plaguing the family. Time frame of studies/data should envisage factors over a minimum five year period to understand the complexities and variables involved in child labour. Actual census data of bonded child labour should be procured and analyzed to curb child labour more effectively.

These are just some of the primary steps that need to be taken to effectively expand the base from where PBM and other representatives who are against child labour can combat the incidence of child labour.

The issue is also one that needs to create awareness amongst the people of our country and sensitize them towards this great evil. For when you send a child to work, you just don’t send him to perform some work in a field or factory. Along with his labour, you murder his dreams, his innocence and his promise of a better future. It is thus our collective responsibility to protect and nourish the soul of the child and give it a better environment for prosperity.

REFERENCES

1. Akhtar, S. and Razzaq, S. (2005). *Child Domestic Labour in Pakistan: overview, issues and testable hypothesis*. Center for research and poverty reduction and income distribution, 1-24.
2. Amin, S., Quayes, M.S. and Rives, J.M. (2004). Poverty and Other Determinants of Child Labor in Bangladesh. *Southern Economic Journal*, 70(4), 876-892.
3. Basu, K. and Van, P.H. (1998). The Economics of Child Labour, *American Economic Review*, 88(3), 412-27.
4. Basu, K. and Van, P.H. (1999). The Economics of Child Labour: Reply. *American Economic Review*, 89(5), 1386-1388.
5. Basu, K. and Tzannatos, Z. (2000). Child Labour and Development: An Introduction. *The World Bank Economic Review*, 17(2), 145-46.
6. Basu, K. and Tzannatos, Z. (2000). Child Labour Problem: What Do We Know and What Can We Do? *The World Bank Economic Review*, 17(2), 147-73
7. Basu, K. (1999). Child Labour: Cause, Consequence, and Cure, with Remarks on International Labour Standards. *Journal of Economic Literature*, 37(3), 1083-1119.
8. Bellettini, G. and Ceroni, C.B. (2004). Compulsory Schooling Laws and the Cure for Child Labour. *Bulletin of Economic Research*, 56(3), 227-239.
9. Bhatti, K. (2007). Pakistan Independence: 60 years of Shattered Dreams. *Socialism Today (Socialist Party magazine),* 111(July-August). Online available at http://socialismtoday.org/archive/111/pakistan.html
10. Cigno, A. (2003). Globalisation can help reduce child Labour. *CESifo Economic Studies*, 49(4), 515-526.
11. Dessy, S.E. and Pallage, S. (2005). A theory of the worst forms of Child Labour. *The Economic Journal*, 115, 68-87.
12. Dumas, C. (2007), Why do parents make their children work? A test of the poverty hypothesis in rural areas of Burkina Faso. *Oxford Economic Papers*, 59(2), 301-29.
13. Emerson, P.M. and Knabb, S.D. (2006). Opportunity, Inequality and the Intergenerational Transmission of Child Labour. *Economica*, 73, 413-34.
14. Fan, C.S. (2004). Relative wage, Child Labor, and Human Capital, *Oxford Economic Papers*. 56(4), 687-700.
15. Government of Pakistan, *Economic Survey of Pakistan 2005-2006*.
16. Government of Pakistan, *Pakistan Demographic Survey 2001-2002*.
17. Government of Pakistan, *Pakistan Demographic Survey 2003-2004*.
18. Humphires, J. (2003). Child Labour Lessons from the Historical Experience of Today's Industrial Economies. T*he World Bank Economic Review*, 17(2), 175-196.
19. Jong-A-Pin, R. (2006). *On the Measurement of Political Instability and its Impact on Economic Growth*, Research Report 06C05, University of Groningen, Research Institute SOM (Systems, Organisations and Management).
20. Kis-Katos, K. and Schulze, G.G. (2005). Regulation of Child Labour. *Institute of Economic Affairs*, 24-30.
21. Kurosaki, T., Ito, S., Fuwa, N., Kubo, K. and Sawada, Y. (2006). Child Labour and School Enrollment in Rural India: Whose education matters. *The Developing Economies*, XLIV(4), 440-64.

22. Lopez-Calva, L.F. and Miyamoto, K. (2004). Filial Obligations and Child Labour. *Review of Development Economics*, 8(3), 489-504.
23. Maffei, A., Raabe, N. and Ursprung, H.W. (2004). Political Repression and Child Labour: Theory and Empirical Evidence. *The World Economy*, 29(2), 211-239.
24. Official Website (n.d). *Pakistan Bait-ul-Mal, National Centres for Rehabilitation of Child Labour*, http://www.pbm.gov.pk/ncrcl.html
25. Official Website (n.d). *CIA World FactBook. Pakistan*. https://www.cia.gov/the-world-factbook/countries/pakistan/
26. Official Website (2008). *Pakistan Finance Ministry* [Online]. Economic Survey. http://www.finance.gov.pk/survey_archieve.html
27. Official Website (2008). *Pakistan Law site* [online]. Available at www.pakistanlawsite.com
28. Official Website (2008). *Pakistan Federal Bureau of Statistics* [Online]. Available at www.statpak.gov.pk
29. Pakistan, Child Labour Resource Centre, Training Module on Child Labour (2007). Center for the Improvement of Working Conditions and Environment Industrial Institute, Labour and Human Resource Department, Government of Punjab.
30. Child Labour Resource Centre (n.d). *Training Kit on Child Labour, ILO-IPEC action programme No. P.095.00.342.150, entitled "Strengthening of Punjab Council for the Rights of Children and Child Labour Resource Centre (CLRC).*
31. Summary results of Child Labour Survey in Pakistan, Federal Bureau of Statistics (FBS), Statistics Division Ministry of Labour, Manpower and Overseas Pakistanis, International Labour Organization (ILO) and International Programme on the Elimination of Child Labour (IPEC). (n.d) https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---sronew_delhi/documents/publication/wcms_436435.pdf
32. *Pakistan, Statistics Pakistan, Census 1998* (n.d). [Online]. https://www.pbs.gov.pk/content/population-census.
33. Pierik, R. and Houwerzijil, M.S. (2006). Western policies on Child Labour Abroad, *Ethics & International Affairs*, 20(2), 193-218.
34. Rogers, C.A. and Swinnerton, K.A. (1999). A theory of exploitive child labour. *Oxford Economic Papers*, 60(1), 20-41.
35. Rosati, F.C. and Rossi, M. (2003). Children's Working Hours and School Enrollment: Evidence from Pakistan and Nicaragua. *The World Bank Economic Review*, 17(2), 283-295.
36. Satz, D. (2003). Child Labour: a Normative Perspective, *the World Bank Economic Review*, 17(2), 297-309.
37. Suvarchala, G. (1992). Legislation to combat child labour: An international perspective. *Industrial Relations Journal*, 23(2), 144-154.
38. Swinnerton, K.A. and Rogers, C.A. (1999). The economics of child labor: Comment. *American Economic Review*, 89(5), 1382-1385.
39. Jafri, S.Y. and Raishad, S.Y. (1997). Some dimensions of child labour in Pakistan. *The Pakistan Development Review*, 36(1), 69-86.

FIGURES

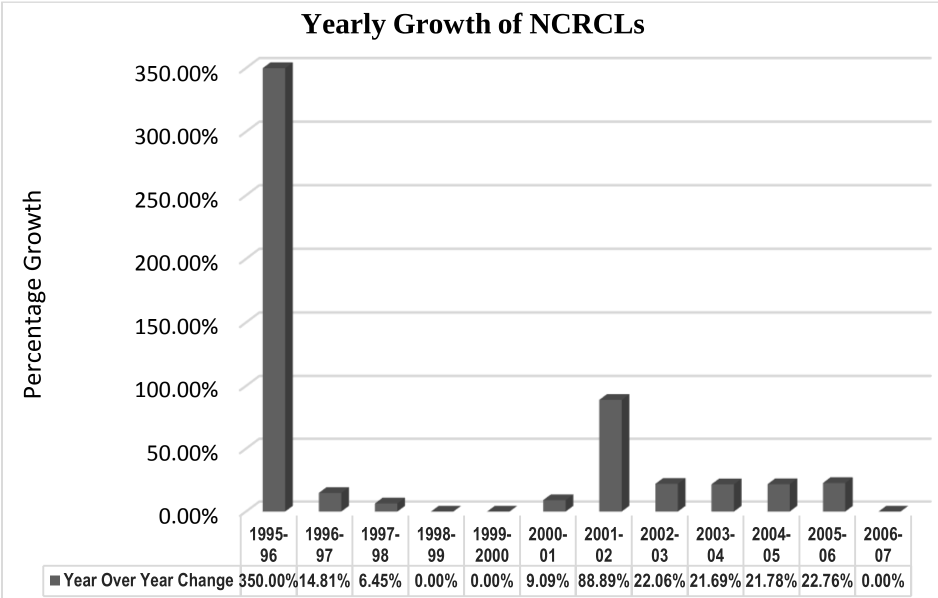


Figure 1.F

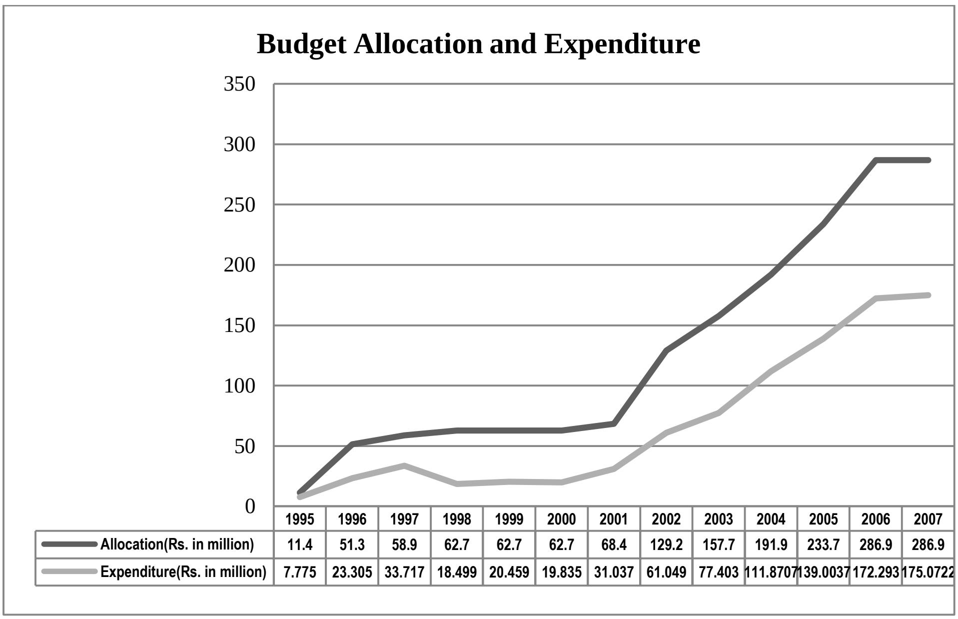


Figure 2.F

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 161-168

PREFERENCE OF TAKING MEDICAL INTERVENTIONS BY DIFFERENT PROFESSIONALS

Muhammad Waqar Hussain[1], **Wajiha Nasir**[2] and **Maria**[1]
[1] Department of Statistics, University of Lahore, Lahore, Pakistan
Email: m.waqarhussain90@gmail.com; maria.afzal71@yahoo.com
[2] Department of Statistics, Govt. College Women University,
Sialkot, Pakistan. Email: wajiha.nasir@gcwus.edu.pk

ABSTRACT

In this paper we studies the Pakistan's medical field is not more established. So, the people preference of taking medical interventions by different professionals. The objective of this study is to investigate the how many people's taking medicine for specialist and how many people's relief disease with in two day in district Hafizabad. From this study we conclude that majority of respondent's preference the specialist for health provider and majority of respondent relief disease within two days.

1. INTRODUCTION

The IOM defines patient-centered care as Health care that establishes a partnership among practitioners, patients, and their families (when appropriate) to ensure that decisions respect patients wants, needs, and preferences and that patients have the education and support they need to make decisions and participate in their own care. Studies show that orienting health care around the preferences and needs of patients has the potential to improve patients' satisfaction with their care, as well as their clinical outcomes. Patient-centered care also has been shown to reduce both underuse and overuse of medical services. Despite the recent prominence given to patient-centered care, and the growing evidence of its importance, the nation's health care system appears to fall short of achieving it. For example, according to a recent Commonwealth Fund survey of patients in five countries (Australia, Canada, New Zealand, the United Kingdom, and the U.S.), one-third of sick patients in the U.S. leave the doctor's office without getting answers to important questions. And across all countries in the study, one-third to one-half of respondents said their doctors sometimes, rarely, or never tells them about treatment options or involves them in making decisions about their care. In landmark 2001, crossing the Quality gap, the Institute of Medicine (IOM) named patient-centered care as one of the six original aims of the U.S. health care system.

The picker institute to explore what it will take to achieve more rapid and widespread implementation of patient centered care in both in patient and ambulatory health care settings. The Picker Institute was an early leader in developing surveys designed to measure patients' experience with their care. Since the late 1980s, the Picker surveys and those modeled after them, such as the Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys, have been used to gather information from millions of patients in hospitals and physician practices in the U.S., Canada, the United Kingdom, Germany, and other European countries. After all this investment in measurement, a clear need remains to determine how such information can be used to actually make and sustain improvements in the patient's experience with care. In adopting this definition in 1998, the International Pharmaceutical Federation (FIP) added one significant amendment: "achieving Definite out comes that improve or maintain a patient's quality of life". Medicines play a very important role in maintaining health, preventing illness, managing chronic conditions and curing disease. In an era of significant economic, demographic and technological challenge it is crucial that patients get the best quality outcomes from medicines. However, there is a growing body of evidence that shows us that there is an urgent need to get the fundamentals of medicines use right. Medicines optimization represents that step change. It is a patient-focused approach to getting then best from investment in and use of medicines that requires a holistic approach, an enhanced level of patient centered professionalism, and partnership between clinical professionals and a patient. Medicines optimization is about ensuring that the right patients get the right choice of medicine, at the right time. By focusing on patients and their experiences, the goal is to help patients to improve their outcomes; take their medicines correctly; avoid taking unnecessary medicines; reduce wastage of medicines; and improve medicines safety. Ultimately medicines optimization can help encourage patients to take ownership of their treatment. However, the medicines optimization approach will require multidisciplinary team working to an extent that has not been seen previously. Healthcare professionals will need to work together to individual care, monitor outcomes more carefully, review medicines more frequently and support patients when needed. The pharmaceutical industry also has a key role to play in medicines optimization through transparent.

WHO clearly states Primary care facility to be in reach of every person all around the globe regardless of area on the globe. One lives in every country has devised its own health care system in accordance with World Health Organization's Alma Ata 1978 declaration (Hall and Taylor, 2003). Primary health care facilities in various countries have different ways to diagnose and treat patients and refer the undiagnosed and critical patients to Secondary and Tertiary care centers. In Pakistan, Primary Care facilities comprise of 6000 Basic Health Units and more than 600 Rural Health Centers Country wide for rural areas and 968 General hospitals for urban/town areas which refer to Tehsil head quarter hospitals (THQ's) and District headquarter hospitals (DHQ's) for referrals. These refer to tertiary care hospitals or teaching hospitals for higher forms of treatments. People lately have lost their trust on the Government led BHU's and RHC's and instead of getting primary care facilities from these lower health care centers seek direct health facilitation from tertiary and secondary centers which is tertiary in most cases. Many people living in urban areas come directly to a teaching hospital which has led to an increased burden in these higher centers. Outdoor departments has seen a robust increase in number of slightly ailed patients as compared to earlier when tertiary care outdoor patient departments were used for higher health services.

2. OBJECTIVE OF THE STUDY

- To identify how many people's taking medicine for specialist.
- To identify how many people's relief disease within two days.

3. LITERATURE REVIEW

Brian Haynes and David Sackett (1979) One of the first studies on compliance. Melnikow and Kiefe (1994) McMaster Workshop/ Symposium on compliance with healing regimen were held, resulting in a book on compliance and compliance research. Many centuries ago Hippocrates was aware of the fact that patients pretended to have taken their medication.

Sackett et al. (1975) became interested in compliance around 1972 when it dawned on him that in hypertension, unpredictable or disappointing responses to treatment were probably due to low compliance. Anon (1997) explored factors doctors and takes their medicine, or follows their advice, a phrase that also means accepting punishment. Patwari et al. (2003) Improvement in the hematological values was also observed As a significant increase has been found in the present study, the data succeeds to prove that hemoglobin and PCV may further increase after 6 months of regular follow up and repeated diet intervention. Sitzia and Wood (1997) Exploring "patient satisfaction" is intuitively appealing as one-way to understand the patient experience and to help guide healthcare providers to improve healthcare. Experts however differ in the methodology of satisfaction surveys in terms of means of communication (like telephonic interview e-mails and questionnaires) consideration of demographic factors and focusing on “satisfaction” or “dissatisfaction”

Hall and Taylor (2003) and Health for all beyond (2000) Primary care according to WHO’s 1978 Alma Ata declaration is better social, mental and physical health for all regardless of the ethnicity and gender etcetera. Saaiq and Zaman (2006), Borras et al. (2001) and Salisbury (1997). A lot of studies have been carried out in the recent past, auditing satisfaction amongst indoor as well as outdoor patients. Saaiq and Zaman (2006). The average of these responses to the questions in each of the three areas was taken as the original Area Score (FAS) and the average of all these individual area scores was taken as the Patient Satisfaction Score (PSS). Hongoro et al. (1998); and Steinmann et al., (2012) Tertiary care centers generally have better diagnostic and treatment facilities and so attract the patients more towards themselves as compared to lower healthcare centers. Roussille and Deschamps (2013) The patients demands efficiently in their daily clinical settings by learning advanced clinical skills social norms, behavioral phenomenon, cultural differences, religious values, attitudes toward health and physical activity. Lobachova et al. (2013). Few patients consider it necessary to have a referral slip before coming to tertiary centers for conditions they think are complicated enough for primary centers clinicians to understand and they have generally decided that on their own. It has increased burden on teaching centers or hospitals. Denecke (2014) this is the basic reason for increasing awareness and attention on professional ethics in physical therapy practice.

4. RESEARCH METHODOLOGY

This study is about the preference of taking medical interventions by different professionals in district Hafizabad. For this purpose questionnaire was designed to collect the data. Peoples were selected as population and simple random sampling technique was used to collect the data. Total respondent were 300.

5. HYPOTHESIS TESTING

For data analysis χ^2 test is applied to check the association between certain attributes. First of all we check the association between age group and preference of health provider. The results were insignificant which shows that there was no association between age group and preference of health provider that the two attributes are independent Now check the association between age group and take medical treatment in six month,. Concerning, Social class and preference for health provider we have constructed the following hypotheses for the possible rejection or acceptance of H_0:

H_0: Social class is independent of the preference for health provider
H_1: Social class is associated with the preference for health provider

We observe that p-value is (0.285) is statistically insignificant indicating that the two attributes are independent for the said purpose we applied the chi-square test of independence and the results are given in Table 2. We can see that the calculated value of $\chi^2 = 22.109$ and the corresponding p-value is (0.036). Both of these values indicate that "Social class" are associated with the "preference for health provider" This show that the patients significant importance to the preference for health provider because they know that own social class on your budget. While checking the association between social class and Take medical treatment in six month p-value (0.001) show that the results are significant. While checking the association between Social class and Specific reason for selecting p-value is (0.104) shows that the results are Insignificant which mean that the two attributes are independent. When we check the association between dwelling you belong and preference for health provider, we find the p-value is (0.920) show that the results are Insignificant which means that there is no association between dwelling you belong and preference for health provider. Now checking the association between preference for health provider and Specific reason the p-value (0.000) so the results are significant. At the order we checking the association between preferences for health provider and Suffering side effects the p-value (0.03) show that the result is significant. By checking the association between preference for health provider and price per visit p-value is (0.000) concluded that the two attributes are highly associated. Check the association between Take medical treatment in six month and Specific reason the p-value (0.000) the significant results. While checking the association between Take medical treatment in six month and Relief the p-value (0.000) the results are significant. While checking the association between Specific reason for selecting and price per visit the p- value is (0.002) show that the results are higher association. Check the association between what kind of treatment like and Relief from disease, p-value (0.096) show that is results are Insignificant which means that there is no association between what kind of treatment like and Relief from disease. While checking the association between Relief from disease and price per visit the p-value is (0.001) show that the results are higher association between Relief from disease and price per visit.

6. CONCLUSION

From this study we find that majority (60%) of the respondents were male while rest of the respondents (40%) were female. The distribution of age group (22.7%) of the respondents have 20 -25. Majority (24.3%) of the respondents has age 25-30 while (16.7%)

have 30-35 while (16%) have 30-40 while (20.3%) above 40 of the respondents has age group. The which social class (6.3%) of the respondents have upper class while (19%) of the respondents have upper middle. Majority (33.1%) of the respondents have middle while (30.3%) have lower middle while(10.7%) have lower of the respondents have social class. Majority (54.3%) of the respondents were belong dwelling rural while (45.7%) of the respondents were urban. Majority (39%) of the respondents have preference of specialist while (32.7%) have dispenser while (7.7%) have herbalist while (20.7%) have medical store keeper of the respondents have preference of health provider. The satisfaction level (34.3%) of the respondents have all time satisfied. Majority (37.3%) of the respondents have most of time while (16.7%) have half time while (11.7%) have occasionally of the respondents satisfaction level in number of cases. Majority (30.7%) of the respondents have one time take medical treatment while (18%) have two time while (18.7%) have three time while (8.3%) have four time while (24.3%) have more than four time of the respondents have taken medical treatments during six months. (15.7%) of the respondents have economic reason to selected health care. Majority (54.7%) of the respondents have personal satisfaction to selected while (13.7%) have method of treatment while(16%) have nearness of the respondents have specific reason for selecting health care provider. (2.7%) of the respondents have all time side effects while (4%) have most time while (5.7%) have half time and Majority (87.7%) of the respondents have occasionally suffering in side effects. Majority (71%) of the respondents have like tablets as a treatment while (20%) have injections while (9%) have drip of the respondents like the treatments. Relief from disease (20.7%) of the respondents have one day. Majority (34.3%) of the respondents have relief in two days while (20%) have three days while (6%) have four days while (19%) have more than four days of the respondents have relief from disease. Majority (47.7%) of the respondents have up to 200 price per visit while (23.7%) have up to 400 while (10.3%) have up to 600 while (5%) have up to 800 while (13.3%) have up to 1000 of the respondents have gave price of medication per visit.

Table 1
Univariate Analysis of Data

Variable	Categories	Number	Percent
Gender	Male	180	60
	Female	120	40
Age group	20 – 25	68	22.7
	25 – 30	73	24.3
	30 – 35	50	16.7
	30 – 40	48	16.0
	above 40	61	20.3
Which Social Class	Upper	19	6.3
	Upper Middle	57	19.0
	Middle Middle	101	33.1
	Lower Middle	91	30.3
	Lower	32	10.7

Variable	Categories	Number	Percent
Dwelling you Belong	Rural	163	54.3
	Urban	137	45.7
Preference for Health Provider	Specialist	117	39.0
	Dispenser	98	32.7
	Herbalist	23	7.7
	Medical Store keeper	62	20.7
Satisfaction Level	All Time	103	34.3
	Most Time	112	37.3
	Half Time	50	16.7
	Occasionally	35	11.7
Take Medical Treatment in Six Month	One	92	30.7
	Two	54	18.0
	Three	56	18.7
	Four	27	8.3
	More than four	73	24.3
Specific reason for selecting health provider	Economic	47	15.7
	Personal satisfaction	164	54.7
	Method of treatment	41	13.7
	Nearness	48	16.0
Side Effects	All time	8	2.7
	Most time	12	4.0
	Half time	17	5.7
	Occasionally	263	87.7
What Kind Treatment Like	Tablets	213	71.0
	Injection	60	20.0
	Drip	27	9.0
Relief from Disease	One day	62	20.7
	Two days	103	34.3
	Three days	60	20.0
	Four days	18	6.0
	More than four days	57	19.0
Price per Visit	Up to 200	143	47.7
	Up to 400	71	23.7
	Up to 600	31	10.3
	Up to 800	15	5.0
	Up to 1000	40	13.3

Table 2
Association between Attributes

Attributes	Chi-square Value	p-value	Conclusion
Age group and preference for health provider	19.708	0.073	Insignificant
Age group and Take medical treatment in six month	18.694	0.285	Insignificant
Social class and preference for health provide	22.109	0.036	Significant
Social class and Take medical treatment in six month	39.484	0.001	Significant
Social class and Specific reason for selecting	18.390	0.104	Insignificant
dwelling you belong and preference for health provider	0.497	0.920	Insignificant
preference for health provider and Specific reason	31.751	0.000	Significant
preference for health provider and Suffering side effects	24.686	0.03	Significant
preference for health provider and price per visit	62.215	0.000	Significant
Take medical treatment in six month and Specific reason	39.421	0.000	Significant
Take medical treatment in six month and Relief	40.852	0.000	Significant
Specific reason for selecting and price per visit	30.714	0.002	Significant
What kind of treatment like and Relief from disease	13.485	0.096	Insignificant
Relief from disease and price per visit	38.78	0.001	Significant

REFERENCES

1. Anon (1997) Editor's Choice. Compliance - A broken concept. *British Medical Journal*, 314(7082), 1.
2. Borras, J.M., Sanchez-Hernandez, A., Navarro, M., Martinez, M., Mendez, E., Ponton, J.L.L., Espinas, J.A. and Germa, J.R. (2001). Compliance, satisfaction, and quality of life of patients with colorectal cancer receiving home chemotherapy or outpatient treatment: A randomised controlled trial. *BMJ*, 322(7290), 826-828.
3. Denecke, K. (2014). Ethical aspects of using medical social media in healthcare applications. *Health Technol. Inform*, 198, 55-62.
4. Hall J.J. and Taylor, R. (2003). Health for all beyond 2000, the demise of the Alma-Ata Declaration and primary health care in developing countries Access to basic health services was affirmed as fundamental human right in the Declaration of Alma-Attain 1978. *Global Health*. "*The Medical Journal of Australia"*, 178, 17-20.
5. Hongoro, C., Musonza, T. G., Macq, J. and Anozie, A. (1998). A qualitative assessment of the referral system at district level in Zimbabwe: implications on efficiency and effective delivery of health services. *The Central African Journal of Medicine*, 44(4), 93-97.
6. Lobachova, L., Brown, D.F., Sinclair, J., Chang, Y., Thielker, K.Z. and Nagurney, J.T. (2014). Patient and provider perceptions of why patients seek care in emergency departments. *The Journal of Emergency Medicine*, 46(1), 104-112.
7. Melnikow, J. and Kiefe, C. (1994) Patient compliance and medical research: issues in methodology. *Journal of General Internal Medicine*, 9, 96-105.
8. Patwari, A.K., Anand, V.K., Kapur, G. and Narayan, S. (2003). Clinical and nutritional profile of children with celiac disease. *Indian Pediatrics*, 40(4), 337-342.
9. Roussille, B. and Deschamps, J. P. (2013). The Ethical Dimensions of Health Education, or the Limits of Benevolence. *Sante Publique*, (HS), 85-91.
10. Saaiq, M. and Zaman, K.U. (2006). Pattern of satisfaction among neurosurgical inpatients. *Journal of the College of Physicians and Surgeons–Pakistan: JCPSP*, 16(7), 455-459.
11. Sackett, D., Gibson, E., Taylor, D.W., Haynes, R.B., Hackett, B., Roberts, R. and Johnson, A. (1975). Randomised clinical trial of strategies for improving medication compliance in primary hypertension. *The Lancet*, 305(7918), 1205-1207.
12. Sackett, D.L. (1979). Introduction. In: Haynes B, Taylor, DW, Sackett DL, (Eds.) *Compliance in Health Care*. Baltimore, MD: The John Hopkins University Press; 1-7.
13. Salisbury, C. (1997). Postal survey of patients' satisfaction with a general practice out of hours cooperative. *BMJ*, 314(7094), 1594-8.
14. Sitzia, J. and Wood, N. (1997). Patient satisfaction: a review of issues and concepts. *Social Science & Medicine*, 45(12), 1829-1843.
15. Steinmann, P., Baimatova, M. and Wyss, K. (2012). Patient referral patterns by family doctors and to selected specialists in Tajikistan. *International Health*, 4(4), 268-276.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 169-180

THE EMPIRICAL ANALYSIS OF GROSS DOMESTIC PRODUCT OF PAKISTAN

Muhammad Waqar Hussain[1], **Wajiha Nasir**[2],
Adeela Shahzadi[1] and **Maria**[1]

[1] Department of Statistics, University of Lahore, Lahore, Pakistan
Email: m.waqarhussain90@gmail.com
adeela.shahzadi@math.uol.edu.pk
maria.afzal71@yahoo.com

[2] Department of Statistics, Govt. College Women University,
Sialkot, Pakistan. Email: wajiha.nasir@gcwus.edu.pk

ABSTRACT

This paper about the study of Annual data Gross Domestic Product of Pakistan. We used data for future forecasting of Gross Domestic Product of Pakistan. Pakistan is counted in developing country in the world. To apply Box-Jenkins methodology to identify the order of ARIMA model and forecast that model. The study evaluates the sample for next 10 years forecasting data. First check the stationary and unit root test is used to check the stationarity of data. It is found that a slight trend in the data and remove trend at 1st difference is applied which makes the data set is stationary. For identification model ACF and PACF are plotted. Selecting the best model used Box Jenkins methodology on basis of smaller AIC, SBC and MSE. Since, ARIMA (0,1,1) has the lowest value of AIC, SBC and MSE. So, this model is recommended as best for forecasting.

1. INTRODUCTION

Time series is a significant area of predicting based on previous observation. Time series is a collection of data which are generally used for forecast and predicting. Time series forecasting refers to the process by which the future values of a system is forecasted based on the information achieved from the past and current data. Generally, predefined mathematical model is used to make accurate predictions. Time series prediction model are mostly used in economical area. In time series two main techniques are used for prediction. These are Auto Regressive (AR) and Moving Average (MA). With the help of these technique we can establish Auto Regressive Moving Average (ARMA), Auto Regressive Integrated Moving Average (ARIMA), Seasonal Regressive Integrated Moving Average (SARIMA) and Box Jenkins models. Many professional and financial time series exhibition seasonal and trend variation. Rabia, et al. (2020) actions to study the Australia percentage time that parts for industrial projects available when needed. Box-Jenkins methodology is used to forecast next 20 observations. To apply BoxJenkins methodology data should be stationary and unit root test is used to check stationarity of data. It is found that a minor trend is found in data set and to remove trend 1st difference is applied which makes the data stationary. For model identification ACF and PACF are plotted. Also, using Box-Jenkins methodology best model is selected on the basis of smaller AIC, SBC and MSE. Since, ARIMA (0,1,1) has the lowest value of AIC, SBC and MSE. So, this model is recommended as best for forecasting. In addition, to check the accuracy of forecasted values MAPE, MAE, RMSE are also computed. It can be concluded that the percentage time that parts for industrial projects in Australia will increase gradually. Iqbal, et al. (2005) used yearly data of wheat area and production time period of 1999-2000 and applied the Box-Jenkins methodology to forecast wheat area and production in Pakistan. They used ARIMA (1,1,1) and ARIMA (2,1,2) models to forecast wheat and production area correspondingly from 2002 to 2022. ARIMA model are quite flexible in that they can denoted many different types of time series, i.e. pure auto-regressive, pure moving average and combined AR and MA (ARMA) series, their main restriction is the pre-assumed linear form of the model and therefore, no nonlinear patters can be captured by the ARIMA. Badar et al. (2015) find that the predicting of the production and yield of food crops in Pakistan. ARIMA model is utilized to predict the wheat, rice and maize production to yield estimate for 2012-30. It was decided that the production growth rate of wheat, rice and maize was higher in 2012. Iqbal et al., (2016) studied that a time series model for forecasting annual wheat production of Pakistan and India. They selected the best ARIMA model by using Box-Jenkins methodology It was developed that best model for wheat production for forecasting of Pakistan and India are ARIMA (1,1,1) and ARIMA (0,1,1) respectively because these models have lower AIC and SBC as associated to other fitted time series models.

2. LITERATURE REVIEW

Atique et al. (2004) struggled to calculate the impression of foreign direct investment on economic growth of Pakistan. The variables they selected for study are FDI, gross capital formation, GDP, ratio of total merchandise trade to GDP, taken as percentage of GDP and education expenditure. They used Eangle Granger and Hansen Methods for data analysis. They selected 31 year data from 1970 to 2001. He analyzed that foreign direct investment make more effect on budget than export of the state. These studied determined that the significant role of FDI in economic growth of Pakistan. Ilhan & Huseyin (2007) investigated the impact of FDI on economic growth of Turkey and Pakistan in time frame of 1975-2004 using Engle-Granger co-integration and Granger causality techniques. They found FDI causality in GDP in the case of Pakistan; so, there was strong indication of a bi-directional causality between FDI and GDP of Turkey. Mortaza et al. (2007) carried out a study to examine the five Asian countries over the period of 1980-2004 using panel data and estimated the relationship between the FDI, trade liberalization and economic growth. They explored positive relationship between FDI and economic growth. The study also examined the direction of causality among FDI, trade liberalization and economic growth using country specific data over the same period. They stated that FDI makes large impact on local investment and trade liberalization along with FDI makes country’s economic growth upward for Bangladesh and Pakistan.

Aqeel et al. (2004) studied the significant impact of FDI on GDP, exchange rate, wages and trade is found in Pakistan in both short run and long run relationship. Aqueel and Nishat (2005) analyzed that the Pakistan government must be changing their economic policy to compete globally, In 1980s, government choice market based economic policy and kept it until 1988. After that, government became very substantial to the FDI by providing fair trade policy, economic indicators and tariff facilities to the foreign investors to make Pakistan the most attractive investment region.

In decomposing foreign capital inflows into its many components, Aurangzeb and Ul Haq, (2012) is examined that the influence of foreign capital inflows on financial development of Pakistan for the period of 1981-2010. A multiple regression examination method was used to categorize the significance of different factors. Their results designated that the three independent variables (remittances, external debt and foreign direct investment) are positive and have an important relationship with commercial progress (GDP). They eminent that the Granger-causality test showed a bidirectional relationship between remittances and external debt, GDP and external debt, foreign direct investment and external obligation, FDI and remittances. On the opponent the outcomes exposed a Uni-directional association from gross domestic production to foreign direct investment. Energy ingesting existence an energetic input in manufacture process, disturbs the GDP directly. Availability of energy at reasonable cost improves competitive of home products in international market, increases exports and affects GDP indirectly. In addition, request for heavy equipment and electrical equipment basic machineries for industrialized growth also depends on appropriate source of energy (Majeed Tariq et al. 2013).

3. METHODOLOGY

In this study, Annual data Gross Domestic Product of Pakistan (Appendix 1) has been forecasted through using ARIMA methodology. Data of these years has been removed and then we applied the Box-Jenkins methodology to forecast the future observation.

3.1 Box-Jenkins Approach

Box-Jenkins approach denotes to an orderly technique of Identification Estimation and testing Application are using for integrated autoregressive moving average (ARIMA) time series models suggested by Box and Jenkins (1976). This method is suitable for time series that have minimum 50 observations. There are three phases of Box-Jenkins approach.

3.2 Model Identification

In time series forming through Box-Jenkins method, an important theory that must be checked beforehand choosing the order of ARIMA model is that the series must be stationary. But it is understand that the typically data is not stationary. Therefore in order to create a series stationary first difference or log difference is used normally. In this study we used first difference and in order to make series stationary. We create ACF and PACF graph of stationary series to study the order of ARIMA model.

3.3 Model Selection

When we make series stationary then next step in the Box-Jenkins method is to classify the order (p, q) of the autoregressive and moving average positions. Plot of ACF and PACF perform the main role is to categorize the order of AR (p) and MA (q). In ACF diagram we select the order of MA (q) model on the foundation of significance spikes cut off after lag q while in PACF charts we select the order of AR (p) model when spikes cuts off after lag p. For model selection AIC quantity is used to observe and estimate the goodness of fit on models. Therefore, a model which has lowest AIC values are measured as best fitted model.

3.4 Diagnostic checking

In the last step, if the model is acceptable then it is used for forecasting otherwise repeat the process again until the required model is obtained. Estimates resulted from this model is specifically for short-term predictions and in most cases. It is more consistent than traditional modeling method of econometrics. Of course, it is necessary to separately judge about each special case.

4. STATIONARY TEST

4.1 Graphical Method

In graphical method we use correlogram and line graph

4.2 Empirical Method

Generally, in this method two tests are used which are Augmented Dickey Fuller Test (ADF Test) and Phillips Perron Test (PP Test).

5. GENERAL PROCEDURE TO CHECK STATIONARITY OF DATA

To analyse the data first of all we check data is stationary or not. If not then make it stationary and make its possible model. The procedure is given below:

Step 1:

Hypothesis:

H_0: Data is not stationary.

H_1: Data is stationary.

Step 2:

Level of significance α=0.05

Step 3:

Test statistic

1. Unit root test or Augmented Dickey Fuller test (ADF Test)
2. Phillip Perron test (PP Test)

Step 4:

Calculation

On E-views

Step 5:

Critical region On the basis of p-value. If the p-value less than level of significance reject Ho otherwise don’t reject Ho.

Step 6:

Decision If reject we conclude series is stationary otherwise we say series is non-stationary.

5.1 To Diagnose Stationary

When the variable has no change in mean and variance for a long time, it said to be stationary. For applying Box Jenkins methodology, variable must be stationary.

5.2 Unit Root Test

When the proof that the series is stationary in the series different unit root tests are applied in works. The most regularly used unit root tests are Augmented Dickey Fuller (ADF) and Phillips Perron (PP) which varies from mostly in how they deal with serial correlation and heteroskedasticity in the errors. Hence ADF and PP test are used to find the reality of unit root in a time series data. Therefore, stationarity tests also support us to identify order of the integration.

5.3 Diagnostic Checking

In Box-Jenkins Methodology, the model is adequate if there is no existence of correlation between residuals and the residuals are independently identical normally distributed (IID) i.e. random. To test whether estimated results of residuals are white noise or not (when residuals shows white noise it means that the model is just right) ACF and PACF of results are plotted.

6. COMMENTS AND CONCLUSION

In this study, time series model utilized to forecast Gross Domestic Product of Pakistan. Firstly data should be stationary and by observing plot of data, a slight trend is found in data set and to remove trend at 1st difference is applied which makes the data stationary. Secondly, for model identification ACF and PACF are plotted. Since, ARIMA (0,1,1) has the lowest value of AIC, SBC and MSE. So, this model is suggested as best for forecasting. Moreover, on the basis of MAPE, MAE, RMSE the accuracy of the forecasted values are also examined and ARIMA (0,1,1) gives the more accurate results as compare to other suggested models. It can be concluded that the Gross Domestic Product of Pakistan will increase regularly in the upcoming years.

REFERENCES

1. Aqeel, A., Nishat, M. and Bilquees, F. (2004). The determinants of foreign direct investment in Pakistan [with comments]. *The Pakistan Development Review*, 651-664.
2. Aqeel, A. and Nishat, M. (2005). The Determinants of Foreign Direct Investment in Pakistan. *Journal of Economics and Finance*, 109(1), 149-82.
3. Arısoy, I. (2012). The impact of foreign direct investment on total factor productivity and economic growth in Turkey. *The Journal of Developing Areas*, 46(1), 17-29.
4. Atique, Z., Ahmad, M.H. and Azhae, U. (2004). The impact of FDI on Economic Growth under Foreign Trade Regimes: A Case Study of Pakistan. *The Pakistan Development Review*, 43(4), 707-718.
5. Badar, N., Fatima, H., Jabbar, A. and Asif, M. (2015). Major food crops production and yield forecast in Pakistan. *Pakistan Journal of Agricultural Research*, 28(3), 295-303.
6. Ilhan, O. and Huseyin, K. (2007). (2007). Foreign Direct Investment and Growth: An Empirical Investigation Based on Cross-Country Comparison. *Economia Internazionale*, 60(1), 75-82.

7. Iqbal, N., Khuda, B., Asif, M. and Ahmad, A.S. (2005). Use of the ARIMA model for forecasting wheat area and production in Pakistan. *Journal of Agriculture and Social Sciences*, 1(2), 120-122.
8. Iqbal, M.C., Jamshaid, T.M. and Rashid, A.Q.A. (2016). Forecasting of wheat production: a comparative study of Pakistan and India. *International Journal of Advanced Research (IJAR)*, 4(12), 698-709.
9. Mortaza, M.G. and Narayan, C.D. (2007). Foreign direct investment, trade liberalization and economic growth: empirical evidence from South Asia and implications for Bangladesh. *Policy Analysis Unit Working Paper Series No. WP0712*, Bangladesh Bank.
10. Rabia, M., Irshad, R., Nawaz, H., Khalid, M. and Raana, A. (2020); forecasting of percentage time that parts for industrial projects in Australia. *International Journal of Advanced Research*, 8(1), 795-802.
11. Shakeel, M., Iqbal, M.M. and Majeed, M.T. (2014). Energy consumption, trade and GDP: A case study of South Asian countries. *The Pakistan Development Review*, 53(4), 461-476.

Table 1.1
Hypothesis and Test Statistic of Unit Root Tests

Unit Root Tests	ADF	PP
Null Hypothesis	Series is not stationary	Series is not stationary
Alternative Hypothesis	Series is stationary	Series is stationary
Test Statistic	$t = \frac{\hat{\theta}}{SE(\hat{\theta})}$	$t = t_{\theta}\sqrt{\frac{\gamma_0}{f_0}} - \frac{T(f_0-\gamma_0)(SE(\hat{\theta}))}{2s\sqrt{f_0}}$

Table 1.2
Unit Root Test to check Stationarity at Level.

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		2.120398	0.9998
Test critical values:	1% level	-3.670170	
	5% level	-2.963972	
	10% level	-2.621007	

From the provided evidence as our p-value (0.9998) so we have strong evidence against Ho and conclude that the results are not significant. So the data is not stationary at 1st level (Table 1.2).

Table 1.3
Unit Root Test to check Stationarity at first difference.

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.067014	0.0039
Test critical values:	1% level	-3.679322	
	5% level	-2.967767	
	10% level	-2.622989	

From the provided evidence as our p-value (0.0039) so we have weak evidence against Ho and conclude that the results are significant. So the data is stationary at 1st difference (Table 1.3)

Model Identification

Table 1.4
AIC and SBC of Different ARIMA Models

Model	AIC	SBC	MSE
0,1,1	10.89805*	11.03817*	0.048*
1,1,0	10.90821	11.04833	0.056
1,1,1	10.93147	11.11829	0.129
4,1,4	10.90232	11.08914	0.123

Table 1.4 determines some possible ARIMA models. The results revealed that ARIMA (0,1,1) has the lowest value of AIC, SBC and MSE. That's why suggested model is ARIMA (0,1,1) because it has least value of AIC and SBC. So, this model is recommended for forecasting.

The general model is:

$$D\ (GDP) = C + \alpha AR\ (p) + \beta MA\ (q) + \mu i$$

where D (GDP) is 1st differenced series of data, C is an intercept, α is coefficient of autoregressive lag values i.e. AR (p), β is a coefficient of moving average lag values i.e. MA (q) and μ shows the residuals of model where residuals should be independently identically and normally distributed.

In ARIMA (0, 1, 1), we use MA (1) model so its estimated equation is:

$$GDP = 36.72087 - 0.259683\ MA\ (1)$$

where GDP denotes the Gross Domestic Product of Pakistan

For checking the accuracy of forecasting we apply forecasting checks.

1. RMSE (Root Mean Square Error)
2. MAE (Mean Absolute Error)
3. MAPE (Mean Absolute Percentage Error)

We select the model which has minimum RMSE, MAE, and MAPE

Table 1.5
MAPE, MAE and RMSE of Different ARIMA Models
Models MAPE MAE

Model	MAPE	MAE	RMSE
0,1,1	30.863*	171.028*	210.265*
1,1,0	38.455	212.852	248.480
1,1,1	31.666	174.942	213.621
4,1,4	32.724	179.24	244.174

Forecasting:

Table 1.6
Forecasting for next 10 years

YEAR	Forecast
2016	1476
2017	1513
2018	1550
2019	1586
2020	1623
2021	1660
2022	1696
2023	1733
2024	1770
2025	1807

Table 1.6 shows the forecasting for next 10 years. It shows that Gross Domestic Product of Pakistan will increase progressively.

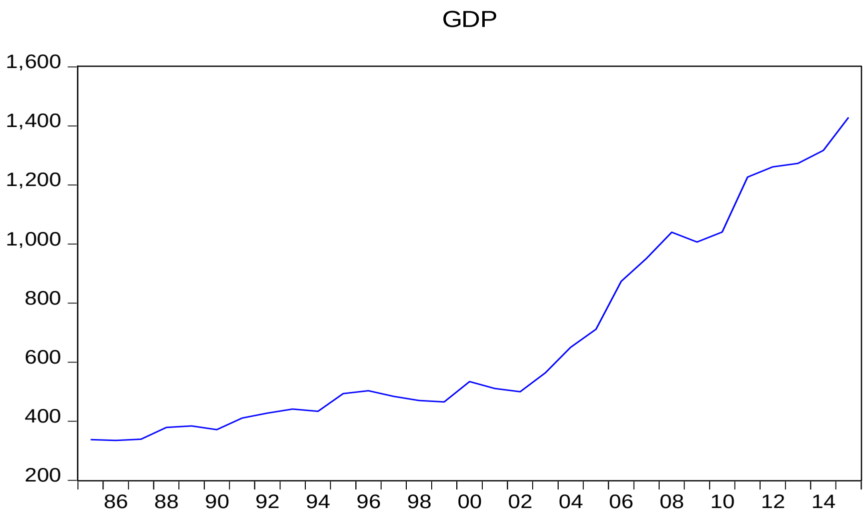
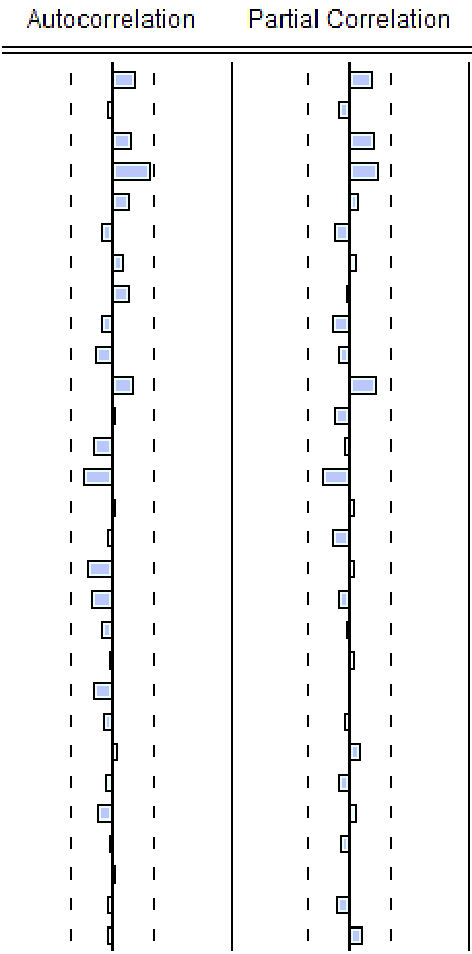


Figure 1.1: Time Series Plot of Data

Figure 1.1 shows the time series plot for data (Gross Domestic Product of Pakistan). The graph shows that there is oscillations or an increasing trend in the data.

Correlogram of GDP

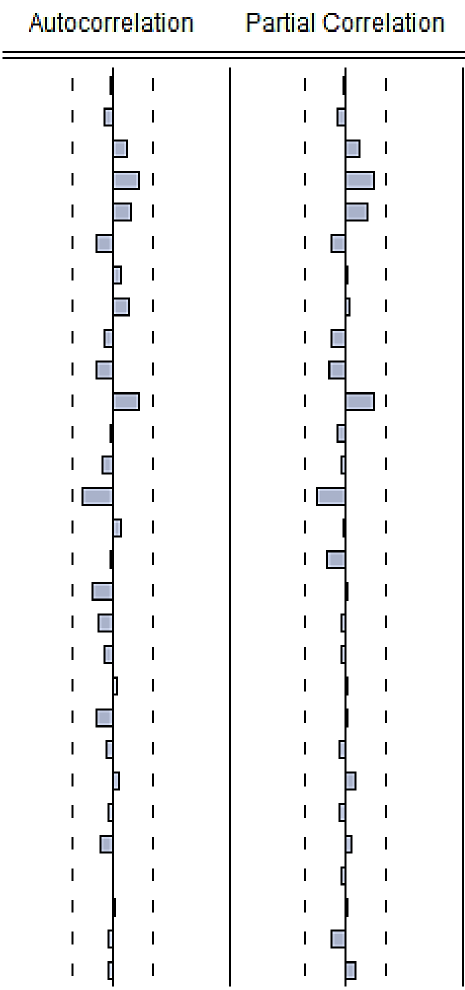


	AC	PAC	Q-Stat	Prob
1	0.197	0.197	1.2795	0.258
2	-0.045	-0.087	1.3484	0.510
3	0.170	0.206	2.3787	0.498
4	0.318	0.254	6.1166	0.191
5	0.149	0.077	6.9747	0.223
6	-0.097	-0.137	7.3500	0.290
7	0.089	0.065	7.6809	0.362
8	0.136	-0.021	8.4835	0.388
9	-0.085	-0.145	8.8145	0.455
10	-0.144	-0.086	9.8110	0.457
11	0.186	0.229	11.568	0.397
12	0.012	-0.126	11.576	0.480
13	-0.164	-0.048	13.102	0.440
14	-0.262	-0.233	17.208	0.245
15	0.022	0.029	17.240	0.305
16	-0.043	-0.143	17.365	0.362
17	-0.209	0.040	20.590	0.245
18	-0.179	-0.084	23.148	0.185
19	-0.083	-0.021	23.744	0.206
20	-0.016	0.030	23.769	0.253
21	-0.168	-0.000	26.785	0.178
22	-0.079	-0.044	27.527	0.192
23	0.039	0.083	27.736	0.226
24	-0.050	-0.083	28.141	0.254
25	-0.120	0.055	30.924	0.192
26	-0.018	-0.070	31.001	0.228
27	0.012	-0.008	31.049	0.269
28	-0.033	-0.108	31.586	0.292
29	-0.036	0.112	32.836	0.284

Figure 1.2: ACF and PACF at 1st difference

From figure of ACF graph, we observed that all spikes are inside the confidence limits. So, same as the PACF plot show’s that all the spikes are inside the confidence limits, In this case we mostly used general models. Here selected (0, 1, and 1). Mean that AR = 0 and MA = 1 is used.

Correlogram for Residuals



	AC	PAC	Q-Stat	Prob
1	-0.016	-0.016	0.0081	
2	-0.067	-0.068	0.1637	0.686
3	0.119	0.117	0.6668	0.716
4	0.246	0.249	2.8922	0.409
5	0.160	0.203	3.8697	0.424
6	-0.153	-0.131	4.8069	0.440
7	0.074	0.019	5.0357	0.539
8	0.145	0.033	5.9517	0.545
9	-0.073	-0.124	6.1958	0.625
10	-0.152	-0.151	7.2981	0.606
11	0.230	0.250	9.9700	0.443
12	-0.011	-0.066	9.9761	0.533
13	-0.089	-0.031	10.427	0.579
14	-0.265	-0.255	14.635	0.331
15	0.074	-0.013	14.988	0.379
16	-0.009	-0.161	14.994	0.452
17	-0.184	0.012	17.488	0.355
18	-0.126	-0.045	18.760	0.342
19	-0.073	-0.040	19.226	0.378
20	0.034	0.022	19.335	0.436
21	-0.153	0.022	21.841	0.349
22	-0.064	-0.065	22.327	0.381
23	0.059	0.087	22.799	0.413
24	-0.044	-0.063	23.110	0.454
25	-0.104	0.056	25.205	0.395
26	-0.006	-0.047	25.214	0.450
27	0.013	0.008	25.265	0.504
28	-0.027	-0.132	25.615	0.540
29	-0.036	0.085	26.832	0.527

Figure 1.3: ACF and PACF of Residuals

Based on the Figure 1.3 it appears that none of the correlations for the autocorrelation and partial autocorrelation function of the residuals are significant. So, the model meet the assumption that the residuals are independent, random and white noise.

APPENDIX 1

http://data.worldbank.org/

YEAR	GDP
1985	337.726
1986	334.8476
1987	339.0848
1988	379.1358
1989	384.0864
1990	371.5726
1991	410.474
1992	427.5728
1993	440.9293
1994	433.2886
1995	493.6615
1996	502.7867
1997	483.6529
1998	470.2489
1999	465.0758
2000	533.8624
2001	510.6568
2002	499.86
2003	563.5943
2004	649.8048
2005	711.4699
2006	873.7703
2007	950.4328
2008	1039.312
2009	1006.604
2010	1040.142
2011	1226.215
2012	1261.209
2013	1272.441
2014	1316.981
2015	1428.638

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 181-194

COMPARISON OF RIDGE M-ESTIMATORS FOR LINEAR REGRESSION MODELS VIOLATING THE USUAL ASSUMPTIONS

Muhammad Suhail[1§] and **Iqra Babar**[2]

[1] College of Statistical and Actuarial Sciences (CSAS),
University of the Punjab, Lahore, Pakistan.
Email: msuhail88@aup.edu.pk

[2] Department of Statistics, The University of Agriculture
Peshawar, Amir Muhammad Khan Campus, Mardan,
Pakistan. Email: iqra.babar@ucp.edu.pk

[§] Corresponding author

ABSTRACT

The ordinary least squares and ridge regression estimators lead to inefficient results when the joint problem of multicollinearity and y-direction outliers are present in the multiple linear regression models. In order to get precise estimates, ridge M-estimation method is usually used in such situations. The ridge parameter, k, plays a key role in achieving the optimal results. The estimation of ridge parameter is an important problem for many researchers. In this article, we considered some existing ridge M-estimators and developed some new estimators for k. Extensive Monte Carlo simulations are used to compare the performance of estimators through mean squared error criterion. The factors we choose to vary are multicollinearity, y-direction outliers, sample size, predictors and error distributions. Some of the new ridge M-estimators performed well compared to the ordinary least square, ridge regression and some existing popular ridge M-estimators. Finally, a numerical example is presented to illustrate the benefits of the new estimators.

KEYWORDS

Multicollinearity; M-estimator; MSE; Outliers; Ridge regression; OLS.

MSC 2010 Classification: 62J05; 62J07.

1. INTRODUCTION

Consider a multiple linear regression model:

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is the random response vector and ε is the vector of random errors and both are of order $n \times 1$. X is a full rank matrix of order $n \times p$ whose column contains centered and standardized explanatory variables. Vector β is a vector of unknown parameters of order $p \times 1$. The ordinary least square (OLS) estimator of vector β is given as follows:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{2}$$

OLS estimators give imprecise results in the case of multicollinearity (Ertaş 2018). Ridge regression (RR) estimator proposed by (Hoerl and Kennard 1970) is used to deal with multicollinearity problem and can be defined as:

$$\begin{aligned}\hat{\beta}(k) &= (X'X + kI)^{-1} X'Y \\ &= R_k \hat{\beta} \end{aligned} \tag{3}$$

where $R_k = (X'X + kI)^{-1} X'X$ and $k(>0)$ is known as the ridge or biasing parameter. OLS and RR estimators are sensitive to typical unusual observations (Maronna 2011). M-estimator (ME) proposed by (Huber 1981) is used to deal with outliers. ME can be obtained by the solution of M-estimating equations $\sum \psi(e_j / s) = 0$ and $\sum \psi(e_j / s) x_j = 0$ where $e_j = Y_j - x_j' \hat{\beta}_m$, s is the scale estimator for errors, $\hat{\beta}_m$ is the ME and $\psi(.)$ is some suitably chosen function (Hampel et al. 1986). (Silvapulle 1991) first suggested the use of ridge M-estimator (RM) to deal with multicollinearity and outliers in y-direction and is defined as:

$$\hat{\beta}_m(k) = R_k \hat{\beta}_m \tag{4}$$

Another approach based on prior information b_0 proposed by (Swindel 1976) as follows:

$$\tilde{\beta}(k, b_0) = (X'X + kI)^{-1}(X'Y + kb_0) \tag{5}$$

Later, (Jahufer and Jianbao 2009) and (Gültay and Kaçıranlar 2015) proposed the use of prior information $b_0 = \hat{\beta}(k)$, therefore the modified RR (MRR) estimator becomes:

$$\begin{aligned}\tilde{\beta}(k) &= (X'X + kI)^{-1}(X'Y + k\hat{\beta}(k)) \\ &= (I + kH_k) H_k X'X \hat{\beta} \\ &= S_k \hat{\beta}, \end{aligned} \tag{6}$$

where $H_k = (X'X + kI)^{-1}$ and $S_k = (I + kH_k) H_k X'X$. To make MRR robust towards outliers in y-direction, (Ertaş 2018) proposed the modified ridge M-estimator (MRM) by replacing the OLS estimator with ME in Eq. (6):

$$\tilde{\beta}_m(k) = S_k \hat{\beta}_m \tag{7}$$

The RR and RM estimators mainly depend on the good choice of ridge parameter k. Several approaches have been proposed for estimating k, see e.g., (Algamal 2018a), (Hoerl and Kennard 1970), (Kibria 2003), (Khalaf, Månsson, and Shukur 2013), (Lawless and Wang 1976), (Gültay and Kaçıranlar 2015), (Kibria and Banik 2016), (Ertaş 2018), (Ali et al. 2019), (Suhail, Chand, and Kibria 2019a, 2019b), (Suhail and Chand 2019), and very recently (Roozbeh, Arashi, and Hamzah 2020) and (Suhail, Ilyas, and Ayanullah 2020) among others. Inspired by (Ertaş 2018), we have reviewed the relevant existing estimators and proposed a new quantile based MRM estimator for estimating k and is defined in Section 2. The criterion for the comparison of estimators is also presented in Section 2. The Monte Carlo simulation study and its results are discussed in Section 3. A real life application is considered in Section 4. Some concluding remarks are outlined in Section 5.

2. STATISTICAL METHODOLOGY

Consider the canonical form of model in Eq. (1):

$$Y = Z\alpha + \varepsilon, \tag{8}$$

where $Z = XD$ and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_p)' = D'\beta$, D is an orthogonal matrix such that $D'D = I$ and $Z'Z = D'X'XD = \Lambda$, $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ contains the eigen values of the $X'X$ matrix. The estimators defined in Eq. (2) to (4) can be written in canonical form as follows:

$$OLS: \quad \hat{\alpha} = \Lambda^{-1} Z'Y, \tag{9}$$

$$RR: \quad \hat{\alpha}(k) = R_k^* \hat{\alpha}, \tag{10}$$

$$RM: \quad \hat{\alpha}_m(k) = R_k^* \hat{\alpha}_m, \tag{11}$$

where $R_k^* = (\Lambda + kI)^{-1}\Lambda$, $k > 0$, $\hat{\alpha}$ is the OLS and $\hat{\alpha}_m$ is the m-estimator of canonical coefficient α. The MSE of estimators OLS, RR, ME and RM are given below:

$$MSE(\hat{\alpha}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \tag{12}$$

$$MSE(\hat{\alpha}(k)) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p} \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}, \tag{13}$$

$$MSE(\hat{\alpha}_m) = \sum_{j=1}^{p} \Omega_{jj}, \tag{14}$$

$$MSE(\hat{\alpha}_m(k)) = \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} \Omega_{jj} + \sum_{j=1}^{p} \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}, \tag{15}$$

where σ^2 is the error variance of model (1), α_j $(j = 1, 2, ..., p)$ is the jth element of α and Ω_{jj} are the diagonal elements of the matrix $\Omega = Cov(\hat{\alpha}_m)$. Note that $\hat{\alpha} = D'\hat{\beta}$ such that $MSE(\hat{\alpha}) = MSE(\hat{\beta})$. Therefore, it suffices to consider the canonical form only. The estimators in Eq. (6) and (7) can be written as:

$$MRR: \quad \tilde{\alpha}(k) = S_k^* \hat{\alpha} \tag{16}$$

$$MRM: \quad \tilde{\alpha}_m(k) = S_k^* \hat{\alpha}_m \tag{17}$$

where $S_k^* = (I + kH_k^*)H_k^*\Lambda$ and $H_k^* = (\Lambda + kI)^{-1}$. The MSE of estimators MRR and MRM suggested by (Ertaş 2018) are given below:

$$MSE(\tilde{\alpha}(k)) = \sigma^2 \sum_{j=1}^{p} \frac{(\lambda_j + 2k)^2 \lambda_j}{(\lambda_j + k)^4} + \sum_{j=1}^{p} \frac{k^4 \alpha_j^2}{(\lambda_j + k)^4}, \tag{18}$$

$$MSE(\tilde{\alpha}_m(k)) = \sum_{j=1}^{p} \frac{(\lambda_j + 2k)^2 \lambda_j^2}{(\lambda_j + k)^4} \Omega_{jj} + \sum_{j=1}^{p} \frac{k^4 \alpha_j^2}{(\lambda_j + k)^4}, \tag{19}$$

Ridge parameter k plays an important role in reducing the MSE of these estimators when the combined problem of multicollinearity and outliers in y-direction is present. The first estimator for k was proposed by (Hoerl and Kennard 1970) as:

$$\hat{k}_j = \frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}, \qquad j = 1, 2, ..., p \tag{20}$$

where $\hat{\sigma}^2 = (\sum_{i=1}^{n} \hat{\varepsilon}_i^2)/(n-p)$ is the unbiased estimator of error variance σ^2 and $\hat{\varepsilon} = Y - \hat{Y}$ is the residual. Furthermore, (Hoerl and Kennard 1970) also suggested to use:

$$\text{RR:} \quad \hat{k}_{RR} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\max}^2}, \tag{21}$$

RM estimator for $\hat{k}_j$, suggested by (Silvapulle 1991), is given by:

$$\text{RM:} \quad \hat{k}_{RM} = \frac{p\hat{A}^2}{\sum_{j=1}^{p} (\hat{\alpha}_m^2)_j}, \tag{22}$$

where $\hat{A}^2$ is given by (Huber 1981):

$$\hat{A}^2 = \frac{s^2 (n-p)^{-1} \sum_{j=1}^{p} (\psi(e_j / s))^2}{(n^{-1} \sum_{j=1}^{p} \psi'(e_j / s))^2}, \tag{23}$$

where $\psi(.)$ is some suitably chosen function (Hampel et al. 1986). (Gültay and Kaçıranlar 2015) proposed MRR estimator is given by:

$$\tilde{k}_j = \frac{\sqrt{\lambda_j \hat{\alpha}_j^2 \hat{\sigma}^2 + \hat{\sigma}^4} + \hat{\sigma}^2}{\hat{\alpha}_j^2} \tag{24}$$

MRM estimator of $\tilde{k}_j$ suggested by (Ertaş 2018) is given by:

$$\tilde{k}_{mj} = \frac{\sqrt{\lambda_j (\hat{\alpha}_m^2)_j \hat{A}^2 + \hat{A}^4} + \hat{A}^2}{(\hat{\alpha}_m^2)_j} \tag{25}$$

(Ertaş 2018) further suggested using harmonic mean of Eq. (25) as follows:

$$\text{MRM: } \tilde{k}_{MRM} = \frac{p}{\sum_{j=1}^{p} \frac{(\hat{\alpha}_m^2)_j}{\sqrt{\lambda_j (\hat{\alpha}_m^2)_j \hat{A}^2 + \hat{A}^4} + \hat{A}^2}} \tag{26}$$

The performance of considered estimators deteriorates in terms of MSE when the degrees of multicollinearity increases from moderately high to severe, outliers in y-direction are present and for some non-normal errors. Therefore, an automated choice was needed to select the optimal value of ridge parameter in such a way that these factors could not adversely affect the performance of estimators. To overcome this problem, we propose a new estimator for *k*. The next section narrates this proposal.

2.1 New Estimator for Ridge Parameter

Let $(\tilde{k}_{m1}, \tilde{k}_{m2}, ..., \tilde{k}_{mp})$ be the realizations of $\tilde{k}_{mj}$ defined in Eq. (25). We write $(\tilde{k}_{m1}, \tilde{k}_{m2}, ..., \tilde{k}_{mp})$ in ascending order of magnitude so that:

$$\tilde{k}_{m(1)} \le \tilde{k}_{m(2)} \le ... \le \tilde{k}_{m(p)} \tag{27}$$

where $\tilde{k}_{m(1)} = \min(\tilde{k}_{m1}, \tilde{k}_{m2}, ..., \tilde{k}_{mp})$ and $\tilde{k}_{m(p)} = \max(\tilde{k}_{m1}, \tilde{k}_{m2}, ..., \tilde{k}_{mp})$. Then the set $\{\tilde{k}_{m(1)}, \tilde{k}_{m(2)}, ..., \tilde{k}_{m(p)}\}$ is the order statistics for $(\tilde{k}_{m1}, \tilde{k}_{m2}, ..., \tilde{k}_{mp})$ and $\tilde{k}_{m(j)}$, $j = 1, 2, ..., p$, is the j^{th} ordered observation. Now let QRM_γ, $0 < \gamma < 1$, be the $100\gamma^{th}$ quantile of $\{\tilde{k}_{m(1)}, \tilde{k}_{m(2)}, ..., \tilde{k}_{m(p)}\}$, then the new proposed quantile based MRM (QRM) estimator is:

$$QRM_\gamma = \left\{\tilde{k}_{m(j)}\right\}_\gamma = \left\{\tilde{k}_{m(1)}, \tilde{k}_{m(2)}, ..., \tilde{k}_{m(p)}\right\}_\gamma , \tag{28}$$

such that

$$P(\tilde{k}_m < QRM_\gamma) = \gamma. \tag{29}$$

The quantile probability γ mainly depends on degrees of multicollinearity and outliers present in y-direction to obtain minimum MSE. Our newly developed estimator QRM_γ is an automated choice of ridge parameter i.e., it depends on quantile probability ‘ γ ’ whose value is selected according to the degrees of multicollinearity and outliers. So the different varying degrees of multicollinearity and outliers do not adversely affect the performance of our new QRM_γ estimator. Here we consider four values for the quantile level, i.e., $\gamma = 0.25, 0.50, 0.75$ and 0.95. The resulting estimators are then denoted respectively by QRM1, QRM2, QRM3 and QRM4.

2.2 Performance Criteria

We use MSE criteria to measure the goodness of an estimator. The estimated MSE of $\hat{\alpha}$ is defined as:

$$MSE(\hat{\alpha}) = E(\hat{\alpha} - \alpha)'(\hat{\alpha} - \alpha) = \frac{1}{p}\sum_{j=1}^{p}(\hat{\alpha}_j - \alpha_j)^2, \tag{30}$$

Following (Kibria 2003), (Khalaf, Månsson, and Shukur 2013) and (Suhail, Chand, and Kibria 2019a), we will compare the estimators through a Monte Carlo simulation study in the following section.

3. SIMULATION STUDY

Since a theoretical comparison among the estimators is not possible, a simulation study will be conducted in this section. It consists of three components: (i) Simulation design (ii) Algorithm of MSE computation and (iii) Simulation results with discussion.

3.1 Simulation Design

Following (McDonald and Galarneau 1975), (Huang and Yang 2014) and (Algamal 2018b), the predictors are generated by using the following equation:

$$x_{ij} = (1 - r^2)^{1/2} z_{ij} + r z_{ip},\ i = 1, 2, ..., n,\ j = 1, 2, ..., p, \tag{31}$$

where r^2 is the degree of multicollinearity between predictors and z_{ij} are the independent pseudo random numbers generated from standard normal distribution. Here we consider three values of r and these are 0.85, 0.95 and 0.99. The dependent variable can be obtained by:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, 2, ..., n, \tag{32}$$

where ε_i is the random error generated from the following distributions.

1) Normal distribution [$N(0,1)$],
2) Student’s t distribution [$t(3)$] and
3) Cauchy distribution [$Cauchy(0,1)$].

Following (Kibria 2003), we consider $\beta_0 = 0$ and choose the coefficients $\beta_1, \beta_2, ..., \beta_p$ as the normalized eigen vectors corresponding to largest eigen value of the $X'X$ matrix so that $\beta'\beta = 1$. The other factors we choose to vary are the sample size (*n=20* and *50*) and the number of predictors (*P=4* and *10*). We consider two different cases for outliers in y-direction, i.e. 10% and 20% outliers in y-direction. Same cases of outliers was also considered by (Ertaş 2018). The above discussion and computation of estimated MSE is explained through the following algorithm.

3.2 Algorithm-1: Computation of MSE of Considered Estimators

1) Generate the standardized data matrix X using Eq. (31).
2) Calculate the eigen values $\lambda_1, \lambda_2, ..., \lambda_p$ and eigen vectors $e_1, e_2, ..., e_p$ of the $X'X$ matrix, such that $\sum_{j=1}^{p} \lambda_j = p$.

1) Obtain the canonical regression coefficients vector, α, using $\alpha = e_{\max} D$ where $e_{\max}$ is the eigen vector corresponding to the largest eigen value and $D = [e_1, e_2, ..., e_p]$.
2) Generate random errors from different distributions considered above from (I)-(III).
3) Generate the dependent variable, Y, using the model in canonical form given in Eq. (8).
4) Obtain the estimates of α using Eq. (9)-(11) and (17).
5) Repeat the above steps (4)–(6) for each of the 5000 Monte Carlo runs.
6) Estimate the MSE for each of the considered estimators using the equation given below:

$$MSE(\hat{\alpha}_j) = \frac{1}{5000} \sum_{k=1}^{5000} (\hat{\alpha}_{jk} - \alpha_j)'(\hat{\alpha}_{jk} - \alpha_j),\ j = 1, 2, ..., p. \tag{33}$$

The estimated MSE of the estimators are presented in Tables 3.1-3.4. R-programming is used for simulation study and code ‘rlm’ is used for the computation of M-estimators.

3.3 Results and Discussions

Tables 3.1-3.4 show the estimated MSE of all considered estimators for different degrees of multicollinearity, sample sizes, number of predictors, percentage of outliers and error distributions. Bold value in the tables show the minimum estimated MSE of an estimator. We can draw some conclusions from these tables as follows.

It is observed that increase in the degree of multicollinearity substantially increases the estimated MSE of all the considered estimators but the RM-type estimators are least effected. Thus, in the presence of multicollinearity when the outliers are also present, the estimators RM, MRM and QRM outperform. Among the RM-type estimators, the proposed estimators showed better results. Increase in the sample size (n), as expected, decreases the estimated MSE of all the considered estimators. When the number of predictors increases from 4 to 10, the estimated MSE of estimators also increases. But the proposed estimators showed efficient results then other estimators. Similar increasing pattern of MSE of estimators is observed when the percentage of y-direction outliers is increased from 10 to

20. The estimated MSE of all the estimators increases except the proposed estimators when the errors of the model are generated from non-normal distributions.

Therefore, based on simulation results, our new quantile based ridge M-estimators (QRM) outperform in many evaluated instances and we recommend the use of estimator QRM for all the considered error distributions. Moreover, among proposed estimators, QRM4 outperform in most of the considered cases. Estimator QRM3 remains a close competitor to QRM4 in many situations. However, in some cases, when the degree of multicollinearity is not very high and n=50, then estimators QRM1 and QRM2 perform better than other considered estimators. An application for the proposed estimators is given in the next section.

4. APPLICATION

To illustrate the benefit of the proposed estimators, we use water data published by Pakistan council of research in water resources (PCRWR) in annual report 2014-15 (PCRWR 2016) to illustrate the usefulness of proposed estimators in presence of both multicollinearity and outliers. The data consist of 25 observations on water mineral contents (mg/l) obtained from district Islamabad, Pakistan. In our study, we consider the following four predictors and one dependent variable as:

Y: Total dissolved solids (TDS),
X_1: Electric conductivity (EC),
X_2: Bicarbonate (HCO_3),
X_3: Hardness (Hard.),
X_4: Power of hydrogen (pH).

We consider the following multiple linear regression model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \tag{33}$$

The eigen values are $\lambda_1 = 3.6811$, $\lambda_2 = 0.1804$, $\lambda_3 = 0.1336$, $\lambda_4 = 0.0490$. The condition number is CN=75.1245 which shows moderately high multicollinearity among predictors. Table 4.1 shows the correlation among predictors and dependent variable. In addition, estimated error variance is 0.1996. Furthermore, observations 7, 12 and 16 have high leverage values than others, i.e., 0.42, 0.77 and 0.61. This shows that data contain 12% outliers in y-direction. Moreover, Shapiro-Wilk normality test is used to test the normality of residuals. We obtain P-value=0.02009 which shows that the residuals are non-normal at 5% level of significance. Regression coefficients are estimated using Eq. (9)-(13) and estimated MSE of all considered estimators are obtained using Eq. (14)-(19). The estimated regression coefficients and MSE results are given in Table 4.2. Estimated regression coefficients of ridge M-estimators are different from other ridge and OLS estimators. In general, estimated MSE of the class of M-estimators are smaller than other ridge type and OLS estimators. Moreover, our new estimators QRM1 and QRM2 outperform all other considered estimators.

5. COMMENTS AND CONCLUSION

In this article, we proposed new quantile based modified ridge M-estimator (QRM_{γ}) in situation when multicollinearity and outliers are present in the data. A simulation study has been conducted to compare the performance of the estimators. Results from simulation study and real life application, using MSE criterion, reveals that our new estimator outperforms other considered estimators in many evaluated instances. Hope the findings of this paper will be helpful for practitioners.

6. ACKNOWLEDGEMENT

We are thankful to reviewers and the editor for their valuable and constructive suggestions which certainly have improved the presentation and quality of the paper.

Table 3.1
Estimated MSE from *N (0, 1)* and 10% Outliers in y-direction

n	**20**			**50**		
r	**0.85**	**0.95**	**0.99**	**0.85**	**0.95**	**0.99**
	P=4					
OLS	55.965	160.506	755.299	5.713	16.441	76.218
RR	5.559	9.283	35.552	4.811	5.264	12.154
M	4.985	8.951	33.562	3.640	4.851	11.904
RM	3.208	4.721	15.986	3.075	3.655	6.882
MRM	3.052	4.383	14.437	2.993	3.516	6.426
QRM1	3.021	4.493	16.018	3.038	3.558	6.662
QRM2	2.460	3.265	9.698	2.788	3.204	5.370
QRM3	1.887	2.242	5.654	2.395	2.740	4.274
QRM4	1.700	1.973	5.016	2.212	2.529	4.004
	P=10					
OLS	133.925	400.469	1983.262	13.782	50.169	282.776
RR	40.366	212.606	1224.030	4.663	14.080	163.807
M	35.651	99.512	472.534	3.051	7.586	28.887
RM	20.656	53.890	249.457	1.795	3.936	10.437
MRM	19.100	49.507	227.089	1.702	3.772	9.714
QRM1	17.685	45.645	211.110	1.810	3.763	8.869
QRM2	14.150	33.885	145.546	1.417	2.923	5.635
QRM3	11.085	25.772	101.087	0.991	2.182	3.445
QRM4	6.120	15.946	69.652	0.808	1.592	2.388

Table 3.2
Estimated MSE from *N (0, 1)* and 20% Outliers in y-direction

n	**20**			**50**		
r	**0.85**	**0.95**	**0.99**	**0.85**	**0.95**	**0.99**
	P=4					
OLS	25.620	95.216	576.876	43.532	111.596	473.421
RR	8.103	34.059	215.016	32.204	79.273	322.172
M	4.128	6.460	21.888	4.390	6.230	15.872
RM	2.927	3.830	10.224	3.849	4.927	10.180
MRM	2.816	3.599	9.226	3.752	4.740	9.540
QRM1	2.921	3.740	10.439	3.805	4.832	9.821
QRM2	2.346	2.989	6.691	3.577	4.418	8.088
QRM3	1.831	2.135	3.784	3.128	3.807	6.600
QRM4	1.656	1.899	3.323	2.892	3.506	6.129
	P=10					
OLS	143.323	387.862	1761.763	19.476	58.631	296.381
RR	10.160	31.434	32.413	6.574	18.176	103.929
M	9.503	21.538	90.206	4.468	8.518	32.109
RM	5.853	11.838	44.038	2.862	4.098	11.132
MRM	5.482	10.847	39.556	2.760	3.890	10.350
QRM1	5.769	11.380	40.944	2.818	3.866	9.904
QRM2	4.507	8.172	27.059	2.371	2.923	6.021
QRM3	3.417	5.664	16.482	1.982	2.277	3.726
QRM4	2.127	3.460	9.667	1.476	1.698	2.676

Table 3.3
Estimated MSE from t *(3)*

n	**20**			**50**		
r	**0.85**	**0.95**	**0.99**	**0.85**	**0.95**	**0.99**
	P=4					
OLS	79.069	198.926	866.386	37.020	90.864	383.865
RR	14.748	29.912	9.661	16.380	36.181	138.353
M	5.001	9.065	32.809	3.943	5.489	14.570
RM	3.420	5.751	17.691	3.235	3.974	8.234
MRM	3.291	5.427	16.029	3.157	3.833	7.695
QRM1	3.291	5.603	16.986	3.176	3.844	7.798
QRM2	2.751	4.621	12.502	2.915	3.460	6.256
QRM3	2.106	3.554	9.843	2.452	2.917	4.921
QRM4	1.871	3.144	8.949	2.234	2.672	4.526
	P=10					
OLS	2543.857	9036.014	48542.406	115.366	327.931	1567.875
RR	11.988	12.051	281.819	62.376	190.272	947.688
M	67.701	225.216	1219.951	6.325	13.550	56.549
RM	27.599	77.648	419.299	3.825	6.493	22.160
MRM	24.764	67.672	361.076	3.650	6.069	20.315
QRM1	20.276	50.812	266.552	3.778	6.158	20.302
QRM2	16.129	38.559	201.552	3.003	4.356	12.366
QRM3	9.508	19.660	99.044	2.324	3.022	6.851
QRM4	4.203	7.351	29.597	1.624	2.057	4.300

Table 3.4
Estimated MSE from *Cauchy (0, 1)*

n	**20**			**50**		
r	**0.85**	**0.95**	**0.99**	**0.85**	**0.95**	**0.99**
	P=4					
OLS	3878575	304993.8	13183118	116668.5	51822.14	774052.5
RR	708889.5	93553.86	4923865	23012.5	17088.27	149818
M	27.926	73.690	425.239	5.844	8.693	42.427
RM	13.941	35.232	234.049	3.958	3.770	19.747
MRM	12.034	30.189	202.795	3.817	3.370	18.142
QRM1	10.104	26.072	187.343	3.734	3.171	18.228
QRM2	6.847	16.680	120.031	3.260	2.204	12.282
QRM3	4.367	11.120	88.954	2.618	1.547	8.740
QRM4	3.552	9.062	78.989	2.345	1.347	7.883
	P=10					
OLS	1142168	1332293	11679104	391586	612076.3	96246902
RR	463872	480089.7	1668250	107287.2	93685.38	7285684
M	190.463	688.329	3051.887	10.203	26.814	127.275
RM	104.742	378.246	1634.928	4.698	10.541	46.121
MRM	96.789	350.141	1491.912	4.331	9.502	41.359
QRM1	103.261	370.138	1559.619	4.623	9.850	42.123
QRM2	69.916	257.664	1033.322	3.099	5.872	23.933
QRM3	39.493	147.292	538.107	1.957	3.288	12.550
QRM4	15.490	55.132	208.088	1.239	1.821	7.201

Table 4.1
Correlations among Predictors and Dependent Variable of Water Data from [24]

Correlation Table	**x1**	**x2**	**x3**	**x4**	**Y**
x1	1.000	0.947	0.900	-0.901	0.930
x2	0.947	1.000	0.874	-0.874	0.840
x3	0.900	0.874	1.000	-0.866	0.953
x4	-0.901	-0.874	-0.866	1.000	-0.846
y	0.930	0.840	0.953	-0.846	1.000

Table 4.2
Estimated MSE and Regression Coefficients of the Estimators of Water Data from [24]

Estimators	MSE	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	1.416	0.485	-0.381	0.538	0.941
RR	0.794	0.485	-0.375	0.530	0.906
M	0.407	0.954	-0.242	0.317	0.047
RM	0.177	0.953	-0.239	0.313	0.045
MRM	0.119	0.954	-0.225	0.294	0.035
QRM1	**0.111**	0.954	-0.229	0.301	0.038
QRM2	0.115	0.954	-0.228	0.298	0.037
QRM3	0.195	0.952	-0.176	0.230	0.020
QRM4	0.399	0.945	-0.108	0.140	0.010

REFERENCES

1. Ertaş, H. (2018). A modified ridge m-estimator for linear regression model with multicollinearity and outliers. *Communications in Statistics-Simulation and Computation*, 47(4), 1240-1250.
2. Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
3. Maronna, R.A. (2011). Robust ridge regression for high-dimensional data. *Technometrics* 53, 44-53.
4. Huber, P.J. (1981). *Robust Statistics*. *New Jersey* (John Wiley & Sons).
5. Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986). *Robust Statis tics*. (J. Wiley & Sons). doi:10.1186/1742-4755-7-12.
6. Silvapulle, M.J. (1991). Robust ridge regression based on an m-estimator. *Australian Journal of Statistics*, 33(3), 319-333.
7. Swindel, B.F. (1976). Good ridge estimators based on prior information. *Commun. Stat. Methods* 5, 1065–1075.
8. Jahufer, A. and Jianbao, C. (2009). Assessing global influential observations in modified ridge regression. *Stat. Probab. Lett.* 79, 513-518.
9. Gültay, B. and Kaçıranlar, S. (2015). Mean square error comparisons of the alternative estimators for the distributed lag models. *Hacettepe J. Math. Stat.* 44, 1215-1233.
10. Algamal, Z.Y. (2018). A new method for choosing the biasing parameter in ridge estimator for generalized linear model. *Chemom. Intell. Lab. Syst.* 183, 96-101.
11. Kibria, B.M.G. (2003). Performance of some New Ridge regression estimators. *Commun. Stat. Part B Simul. Comput.* 32, 419-435.
12. Khalaf, G., Månsson, K. and Shukur, G. (2013). Modified ridge regression estimators. *Commun. Stat. Methods*, 42, 1476-1487.
13. Lawless, J.F. and Wang, P. (1976). Simulation Study of Ridge and other Regression

Estimators. *Commun. Stat. Part A-theory methods* 307-323.
14. Kibria, B.M.G. and Banik, S. (2017). Some Ridge Regression Estimators and Their Performances. *J. Mod. Appl. Stat. Methods,* 15, 206-238.
15. Ali, S., Khan, H., Shah, I., Butt, M.M. and Suhail, M.A (2019). Comparison of some new and old robust ridge regression estimators. *Commun. Stat. Simul. Comput.* 1-19. https://doi.org/10.1080/03610918.2019.1597119
16. Suhail, M., Chand, S. and Kibria, B.G. (2020). Quantile based estimation of biasing parameters in ridge regression model. *Communications in Statistics-Simulation and Computation*, 49(10), 2732-2744.
17. Suhail, M., Chand, S. and Kibria, B.G. (2019). Quantile-based robust ridge m-estimator for linear regression model in presence of multicollinearity and outliers. *Communications in Statistics-Simulation and Computation*, 1-13. https://doi.org/10.1080/03610918.2019.1621339
18. Suhail, M. and Chand, S. (2019). Performance of some new ridge regression estimators. In 2019 *13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (pp. 1-4). IEEE.
19. Roozbeh, M., Arashi, M. and Hamzah, N.A. (2020). Generalized cross-validation for simultaneous optimization of tuning parameters in ridge regression. *Iranian Journal of Science and Technology, Transactions A: Science*, 44(2), 473-485.
20. Suhail, M., Ilyas, M. and Ullah, A. (2019). Modified Ridge Regression Estimators: A Simulation Study. *In 15th Islamic Countries Conference on Statistical Sciences (ICCS-15).* (119-144).
21. McDonald, G.C. and Galarneau, D.I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *J. Am. Stat. Assoc.* 70, 407-416.
22. Huang, J. and Yang, H.A. (2014). Two-parameter estimator in the negative binomial regression model. *J. Stat. Comput. Simul.* 84, 124-134.
23. Algamal, Z.Y. (2018). Developing a ridge estimator for the gamma regression model. *J. Chemom.* 32, e3054.
24. PCRWR (2016). *Annual 2014-15 Report*. Pakistan Council of Research in Water Resources (PCRWR), Islamabad, Pakistan.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 195-207

TEACHER EVALUATIONS AND QUALITY EDUCATION

Ammara Nawaz Cheema and **Arooj Zaheer**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Student feedback is the output provided by the students related to their perception of teachings. This is offered to the students in every semester at the end in response to the various question asked about their activities, tasks and the environment they are engaged in the whole semester. Student's achievements are a big contribution by the teachers. Teacher evaluation is focused these days to determine the teaching impact on student outcomes and is also use for their professional development. Different researches show different results for evaluation of teachers. This overview provides the information about the teachers practice and using it to improve their system. With existing empirical evidence mostly restricted to short-run outcomes. This study examines long-term academic outcomes of Teacher Evaluation for teachers by bachelor students at Air University. On the basis of statistical analysis the effectiveness of said dataset and relation towards teacher evaluation policies have been formulated.

1. INTRODUCTION

The aim of this study is to discover the course and teacher characteristics that may affect the teacher rating in university or college. Characteristics of course and lecturer are correlated with the teaching performance. These days' teachers are held responsible for how well they are teaching and how well the students are learning as it is becoming a common practice to evaluate teachers and grading them in colleges and universities which will help the administration in the promotion of the teachers.

Generally in whole semester students observe the teachers activity in class, how they interact with the students on the basis of which students evaluate teachers performance and course characteristics well. Various studies proposed different forums for evaluation systems. They vary from institution to institution in objectives. Mainly three objectives are highlighted by every institution i.e. teacher characteristics, course characteristics and tutorials. Teacher performance in the class is directly proportional to student positive response for the teachers. If the teacher's nature is friendly, he is concerned about his class environment and the student feels free to ask any question he will evaluate the teacher positively course difficulty is directly related to the negative feedback of the teacher. If a student is able to learn the course, his understanding level is good and he will evaluate the teacher favorably. Student ratings for teachers are effective means of teacher's evaluation and can provide: a) well defined feedback to improve the quality of teaching performance in university b) helps administration for fairer promotions. c) Improvement in course content d) information about the course for students which will help in teacher selection. Student evaluation system used in many universities and has been used as assurance of their quality education. This input by the students is used by universities worldwide to evaluate faculty and to observe their teaching style. With the help of these surveys and feedback systems students provide quality input for the improvement of course content and teaching performance. These days institutions need the quality input for the promotion of faculty, increment in salary and tenure. Student evaluation play a great role in improving the rating of institutions as teachers are observed by the students keenly in the class and feedback system is an anonymous message from students to the administration for the betterment in their system. But student evaluation may be wrong for the teacher if he is not having any idea about the course content, not attending the class properly so his feedback for the teacher may not be enough to provide valuable information. So, the main aim of this study is to look over the role of students and to set some criteria on which the student may be able to evaluate the teacher in higher institutions.

Evaluating teaching ability of a faculty member is one of the most unfavorable and controversial work faced by accounting administrators. While teaching ability is considered as one of the basic factor included in promotion and tenure decisions, there is agreement on how teacher effectiveness should be measured. The conclusive result of effective teaching is student learning, where students master the content of specific courses. Many faculty members often declare the impossibility of quantitatively measuring effective teaching, yet they feel certain that they are effective teachers.

H.H Rammers were the first researcher who takes the initiated toward the research programmed in this field and is considers as the father of students' evaluation. In 1927 Rammers published the three basic principles for designing these types of instruments such as

- Traits should be short to avoid the effects and carelessness due to the boredom of students.
- Traits must be chosen those on which the experts are agreed and are considered as important.

Traits chosen must be susceptible to the student evaluation and judgment in his research study he examined many issues such as reliability, relation between students' grades and ratings, discrimination and relative value of the various traits.

For the students who belongs to college and the universities it seems common practice to evaluate their teachers and grade them as they are detained accountable for how fit they behave and help the students (Germain & Scendura, 2005)

In this regard of evaluation students noticed the teacher's activity in class and observe the characteristics which may affect the student learning process. Students tend to like their lecture in view whether they have a friendly nature or not and whether they are worried for their students or not (Magno & Sambrano, 2008).

The course content plays an important role in student evaluation. Course difficulty is directly proportional to the students' negative behavior and the bad evaluation for the teachers. Students who are capable to understand the course evaluate their teacher more favorably (Chang, 2000).

Later on (Rindermann & Schofield; 2001) studied the effective instructions given by the teachers totally depend on the behavior and the response of the students.

Teachers as scholars are now a portion of educational background. In 1980s the teachers research has been discriminating, teachers need to be active researchers with their students in the classroom (Lytle, 1999).

Classroom testing is described as organized data collection for problem research and to expand specialized exercise (Hatch, Greer, & Bailey, 2006).

While focusing on teachers' practice, classroom environment has been measured dominant to student success. Though, it is standard that what is teacher's activity outside the classroom, mainly complete their own investigation and assessment of informative methods, and is important in terms of excellence teaching and effect for students. In current years there have been several national calls in both the United Kingdom and Australia (White et al., 2018) for an amplified attention on ability building for the teaching occupation in relation to research. By motivating above articles, we decide to enhance research on teacher evaluations. The rest of the article is planned as: in section 3 hypothesis and goal of study are described, sec 4 and 5 contains methodology and future work and in sec 6 we conclusion of study appears.

In 1997 Timpson and Andrews argue that; to evaluate effective teaching there is powerful compatibility that student evaluations are a valid and reliable tool.

The study of Flores and Clark (2004) found that personality (character), activities, teaching style, and moral worth using multivariate regression analysis showed effective teaching efficacy.

In 1997 Calderon, Green, and Reider concluded it will lower the standard of the course if the faculty of accounting ratings will be improved on students by lowering the regour.

Mason et al. in 1996 concluded that ratings will not be adjust by the administration for difference in class size and difficulty in course, and to improve their results for evaluation as their promotion and tenure depends on it teachers change their behaviors.

(Yunker & sterner 1988) conclude that student evaluation for teachers is one of the primary tools for checking the teaching effectiveness in the university. After further research they found that many faculties believe that there is no relation between student evaluation and teaching effectiveness. Faculty also do objection that factors such as class size and difficulty of course effect their evaluation.

Faculty do believe that their evaluations can be improved by giving students high grades and by lowering the standards of their course which will weaken student learning and teaching effectiveness. faculty requested administrators should realize that evaluation by student is not simple a primary tool to measure teaching effectiveness but it also results in behavior change of teachers to improve their ratings by students (Stratton et al., 1994)

(Simpson, 1995) for promotion decisions, administrators must be able to validly measure effective teaching by using the teaching ability as a primary factor in evaluation. Faculty argue most often that it is difficult to identify the effectiveness of teaching although it's almost impossible. So, every faculty member should be allowed to use their subjective judgment to determine how to conduct their classes. He also found consistency in the student evaluation and most disputable source of information used to evaluate the teaching effectiveness.

Shatter in 1998 said that student value the well-organized lectures but their presence in class and attention is mandatory for evaluating teacher well.

In 2010 Beuningen Restricted feedback system can divert learner's attention from broader view of accuracy which possible hinders their domains

Savita and Paul in 2012 concluded the teaching qualities, as approachable, knowledgeable, enthusiastic, logical, and reliable, role model, learner and a mentor. Becoming an effective teacher is a continual refinement in teaching.

Many studies argue that effective information about many factors related to effectiveness of teaching can provide properly designed student. (Cohen 1981)

2. METHODOLOGY AND ANALYSIS

The variables used and analysed are Face preparation (X_1), Subject demo (X_2), Course completion (X_3), Additional material (X_4), Citations (X_5, Communication (X_6), Behaviour (X_7), environment (X_8), Arrives on time (X_9), Leaves on time (X_{10}), Fairer examination (X_{11}), Grade returns in time (X_{12}), Office hours (X_{13}), Knowledge driven (X_{14}), Syllabus clarity (X_{15}), Integrate with real world (X_{16}), Exam relevant study (X_{17}).

S.No.	Department
1	Computer Science
2	Business Administration

Data is obtained from Air University Islamabad. The set of observations consists of evaluation from eight semesters of different teachers for prediction of an appropriate model and to identify the variables that are most important. This data set has thousands of observations depending on 16 variables. This evaluation is done between two departments as mentioned in above table.

Statistical Analysis (BBA Department)

The representation of 1, 2, 3 and 4 in the graph is (1 for strongly agree, 2 for agree, 3 for disagree and 4 for strongly disagree)

Table 1
Descriptive Analysis (Face preparation, Subject Demo, Course Completion, Additional Material)

Description	Face Preparation	Subject Demo	Course Completion	Additional Material
Mean	1.56	1.62	1.62	1.56
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.599	.655	.635	.599
Skewness	1.458	1.284	1.284	1.458
Std. Error of Skewness	.011	.011	.011	.011
Kurtosis	1.786	1.142	1.179	1.786
Std. Error of Kurtosis	.023	.023	.023	.023
Range	3	3	3	3

The above Table 1 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables.

Table 2
Descriptive Analysis (Integrate Real world, Citations, Communication, Behaviour)

Description	Integrate Real World	Citations	Communication	Behaviour
Mean	1.66	1.62	1.63	1.57
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.733	.655	.684	.598
Skewness	1.261	1.284	1.315	1.427
Std. Error of Skewness	.012	.011	.011	.011
Kurtosis	.905	1.142	1.164	1.747
Std. Error of Kurtosis	.025	.023	.023	.023
Range	3	3	3	3

The above Table 2 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. This descriptive analysis shows the results of comparison between the 4 headers mentioned.

Table 3
Descriptive Analysis (Environment, Arrives on Time, Leaves on Time, Fair Examination)

Description	Environment	Arrives on time	Leaves on time	Fair Examination
Mean	1.63	1.54	1.59	1.60
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.684	.581	.617	.642
Skewness	1.315	1.498	1.376	1.342
Std. Error of Skewness	.011	.011	.011	.011
Kurtosis	1.164	2.001	1.507	1.331
Std. Error of Kurtosis	.023	.023	.023	.023
Range	3	3	3	3

The above Table 3 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. Which means individual performance of the teacher is very good as per point of view of students. This descriptive analysis shows the results of comparison between the 4 headers mentioned.

Table 4
Descriptive Analysis (Grade Returns, Available in Office Hours, Knowledge Driven, Syllabus Clarity)

Description	Grade Returns	Available in Office Hours	Knowledge Driven	Syllabus Clarity
Mean	1.62	1.62	1.60	1.65
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.648	.657	.621	.691
Skewness	1.280	1.296	1.343	1.245
Std. Error of Skewness	.011	.011	.011	.011
Kurtosis	1.153	1.169	1.437	.940
Std. Error of Kurtosis	.023	.023	.023	.023
Range	3	3	3	3

The above Table 4 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. Which means the teacher is very fair as per point of view of students regarding their studies and grading them. This descriptive analysis shows the results of comparison between the 4 headers mentioned.

These tables are showing the statistical analysis of BBA department and we have found that use of technology and lecture preparation and communication between the teacher and students are the most valuable factors and only these variables vary from student to student and helps to evaluate the teachers.

Pie Charts

The representation of A, B, C and D in the graphs is as:

A = Strongly Agree
B = Agree
C = Disagree
D = Strongly Disagree

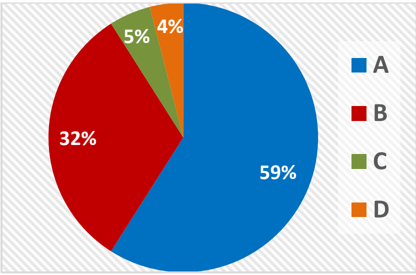


Figure 1: Use of Technology and Lecture Preparations Chart

As we clearly see that pie chart showing the percentages for various responses given by the students. For this factor "use of technology and lecture preparation" 59% students are strongly agreed, 32% are agreed, 5% for disagree and rest 4% are strongly disagree. Strongly agree has the majority votes which means the teacher was able to teach the students very well and also satisfy them with proper lecture deliverance and grading their work.

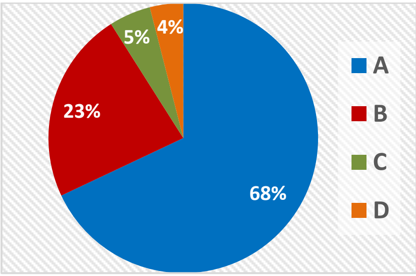


Figure 2: Communication Chart

As we clearly see that pie chart showing the percentages for various responses given by the students. For this factor "teacher student connection" 68% students are strongly agreed, 23% are agreed, 5% for disagree and 4% strongly disagree. This means that the teachers has communicated with students properly and was able to resolve the students various questions and problems regarding the course.

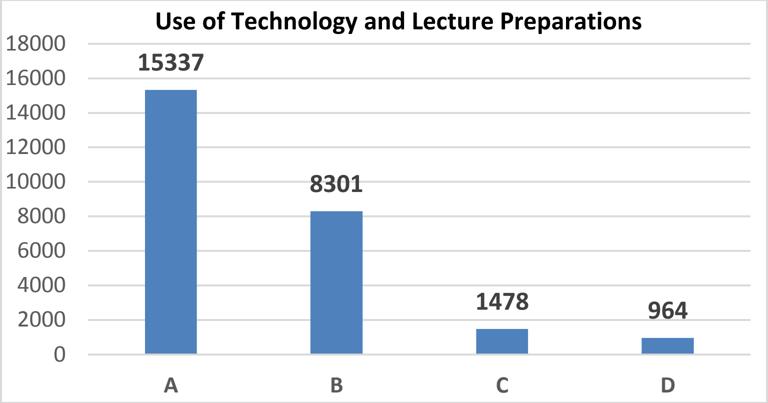


Figure 3: Bar Graphs

In this graph we can see 15337 students are strongly agreed, 8301 are agreed, 1478 are disagreed and 964 are strongly disagreed for the use of technology and lecture preparations. As per defined in above pi chart the data we collected is displayed in bar chart to distinguish and more properly display the results of our observation.

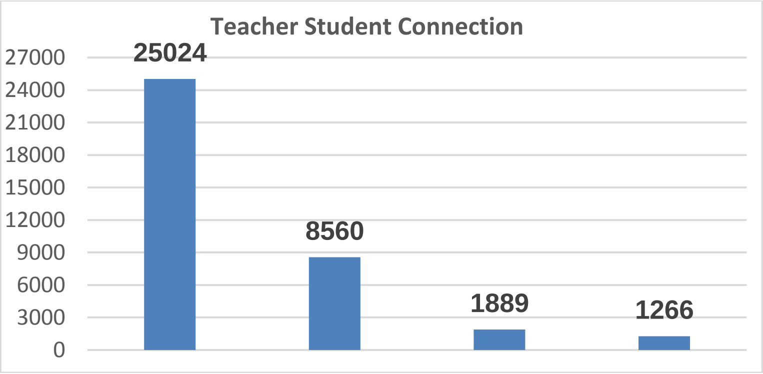


Figure 4:

As we can see that bar graphs are clearly showing the number of students for each response given by students. 8560 students are agreed, 25024 are strongly agreed, 1889 are disagreed and 1266 are strongly disagreed for the connection between the teacher and student. As per defined in above pi chart the data we collected is displayed in bar chart to distinguish and more properly display the results of our observation. We can see that students are satisfied from their teachers.

Statistical Analysis (CS Department)

The representation of 1,2,3 and 4 in the graph is (1 for strongly agree, 2 for agree, 3 for disagree and 4 for strongly disagree)

Table 5
Descriptive Analysis (Face preparation, Subject Demo, Course Completion, Additional Material)

Description	Face Preparation	Subject Demo	Course Completion	Additional Material
Mean	1.59	1.64	1.66	1.66
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.700	.739	.756	.763
Skewness	1.444	1.335	1.250	1.295
Std. Error of Skewness	.023	.023	.023	.023
Kurtosis	1.459	1.063	.774	.917
Std. Error of Kurtosis	.046	.046	.046	.046
Range	3	3	3	3

The above Table 5 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. Which means the teacher is very fair as per point of view of students regarding their studies and grading them. This descriptive analysis shows the results of comparison between the 4 headers mentioned

Table 6
Descriptive Analysis (Integrate Real World, Citations, Communication, Behaviour)

Description	Integrate Real World	Citations	Communication	Behaviour
Mean	1.59	1.62	1.60	1.62
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.694	.725	.702	.725
Skewness	1.445	1.393	1.426	1.393
Std. Error of Skewness	.023	.023	.023	.023
Kurtosis	1.472	1.260	1.379	1.260
Std. Error of Kurtosis	.046	.046	.046	.046
Range	3	3	3	3

The above Table 6 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. This descriptive analysis shows the results of comparison between the 4 headers mentioned.

Table 7
Descriptive Analysis (Environment, Arrives on Time, Leaves on Time, Fair examination)

Description	Environment	Arrives on Time	Leaves on Time	Fair Examination
Mean	1.57	1.60	1.65	1.67
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.653	.676	.768	.772
Skewness	1.486	1.400	1.292	1.266
Std. Error of Skewness	.023	.023	.023	.023
Kurtosis	1.708	1.391	.859	.821
Std. Error of Kurtosis	.046	.046	.046	.046
Range	3	3	3	3

The above Table 7 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. Which means individual performance of the teacher is very good as per point of view of students. This descriptive analysis shows the results of comparison between the 4 headers mentioned

Table 8
Descriptive Analysis (Grade Returns, Available in Office Hours, Knowledge Driven, Syllabus clarity)

Description	Grade Returns	Available in Office Hours	Knowledge Driven	Syllabus Clarity
Mean	1.65	1.62	1.68	1.66
Median	1.00	1.00	1.00	1.00
Mode	1	1	1	1
Variance	.757	.702	.791	.785
Skewness	1.301	1.369	1.253	1.311
Std. Error of Skewness	.023	.023	.023	.023
Kurtosis	.926	1.285	.732	.921
Std. Error of Kurtosis	.046	.046	.046	.046
Range	3	3	3	3

The above Table 8 shows the results in which we can see that the number 1 representing the strongly agree variable is selected most often, which means that its variance is lower as compared to the other 3 variables. Which means the teacher is very fair as per point of view of students regarding their studies and grading them. This descriptive analysis shows the results of comparison between the 4 headers mentioned.

These tables are showing the statistical analysis of CS department and we have found that use of technology and lecture preparation and communication between the teacher and students are the most valuable factors and only these variables vary from student to student and helps to evaluate the teachers

Pie Charts

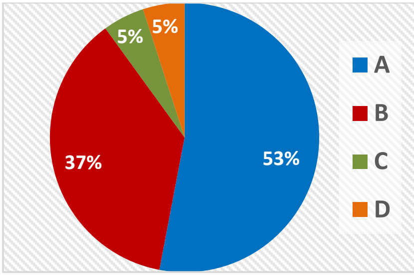


Figure 5: Use of Technology and Lecture Preparations

As we clearly see that pie chart showing the percentages for various responses given by the students. For this factor "use of technology and lecture preparation" 53% students are strongly agreed, 37% are agreed and 5% for both disagree and strongly disagree.

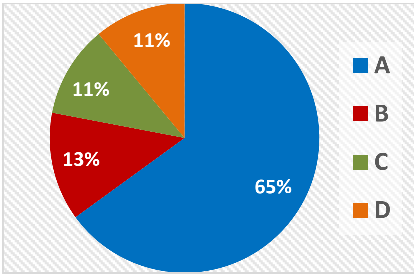


Figure 6: Communication

As we clearly see that pie chart showing the percentages for various responses given by the students. For this factor "teacher student connection" 65% students are strongly agreed, 13% are agreed and 11% for both disagree and strongly disagree.

Bar Graphs (CS Department)

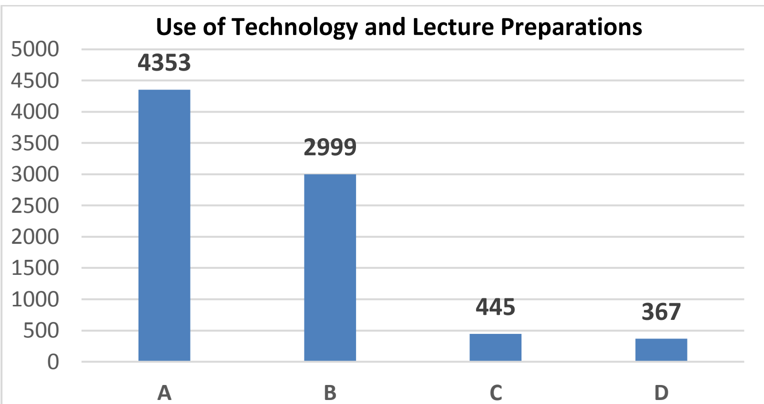


Figure 7

In this graph we can see 4353 students are strongly agreed, 2999 are agreed, 445 are disagreed and 367 are strongly disagreed for the use of technology and lecture preparations.

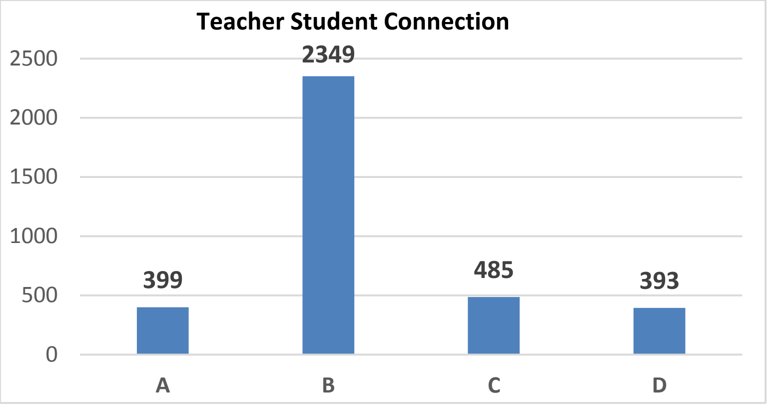


Figure 8

As we can see that bar graphs are clearly showing the number of students for each response given by students.

4. COMMENTS AND CONCLUSION

- The recommendation made in this study is helpful for university management and for administration to analyse the large amount of data.
- The selection of variables that contributes more towards performance of teachers and to select best prediction model for developing more policies that helps teachers to perform better in future.
- This research work also facilitates the administration to work with large amount of data.
- This study brings more ease in interpreting the teacher evaluation which helps in fairer promotions, tenure and pay increments for the staff, job satisfaction criteria is improved and basically it will create a state of justice for every teacher.

REFERENCES

1. Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The Utility of Student Ratings of Instruction for Students, Faculty, and Administrators: A “Consequential Validity” Study. *Canadian Journal of Higher Education*, 35, 49-70.
2. Chang, T.S. (2000). Student Ratings: What Are Teacher College Students Telling Us about Them? Paper presented at the *Annual Meeting of the American Educational Research Association* (New Orleans, LA, April 24-28, 2000).
3. Emery, C.R., Kramer, T.R. and Tian, R.G. (2003). Return to Academic Standards: A Critique of Student Evaluations of Teaching Effectiveness. *Quality Assurance in Education*, 11(1), 37-46.

4. Germain, M.L. and Scandura, T.A. (2005). Grade Inflation and Student Individual Differences as Systematic Bias in Faculty Evaluations. *Journal of Instructional Psychology*, 32(1), 58-67.
5. Hattie, J.C. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge, London & New York.
6. Kupermintz, H. (2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25(3), 287-298.
7. Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151.
8. MacCallum, R. C. (2009). Factor analysis. In: R. E. Millsap, & A. Maydeu-Olivares (Eds.), *The Sage handbook of Quantitative Methods in Psychology* (pp. 123-177) Sage Publications, London;
9. Morgan, D.A., Sneed, J. and Swinney, L. (2003). Are Student Evaluations A Valid Measure of Teaching Effectiveness: Perceptions of Accounting Faculty Members and Administrators. *Management Research News*, 26(7), 17-32.
10. Mukherji, S. and Rustagi, N. (2008). Teaching Evaluations: Perception of Students and Faculty. *Journal of College Teaching & Learning*, 5(9), 45-54.
11. Rindermann, H. and Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education*, 42(4), 377-399.
12. Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
13. Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute Press.
14. Conti, G., Frühwirth-Schnatter, S., Heckman, J.J. and Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, 183, 31-57.
15. Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.
16. Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187-200.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 209-216

STATISTICAL ANALYSIS OF CEMENT STUDY

Umm-e-Habiba and **Ammara Nawaz Cheema**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Cement is broadly utilized all over for development purposes and quality comparison of diverse ordinary Portland cement (OPC) is essential. This study is related to analysis of cement study in Pakistan and for this study distinctive brands of cement are tried for their chemical constituents such as silica, alumina, iron, calcium, magnesium, sulfates, insoluble residue, free lime and loss on ignition with ASTM. We applied statistical tools for our study like to conclude results. This study found that these methods have a major role in collecting, evaluating data for the identification and resolution of production process experiments, in edict to promote the accomplishment of the fineness in performance of organizations.

1. INTRODUCTION

Different companies of OPC specifically Askari, DG, Best way, Pioneer and Flying cement had been chosen for this comparative study. For these cements chemical parts such as Silica, Alumina, Iron Oxide , Calcium Oxide , Magnesium Oxide , Sulfur trioxide ,Insoluble Residue (*IR*), free lime and Loss on Ignition had been decided the usage of well-known methods.

A massive number of OPC is formed at specific cement plants in Pakistan and is used for the construction of building, and dual carriageway and extra home purposes. This is particularly widespread for Pakistan the place tremors are every day and alongside these lines entirely constructions must have the choice to face up to excessive shock. It is used to enhance constructing shape effectively. The cement overall performance will be applicable if the stated residences loosen up interior indicated limits .To make certain the great of cement a variety of exams carried out that conform the requirement of the applicable necessities. In Pakistan, the cement sector is divided into two zones, the North and South Zones. Askari cement, DG khan cement, Best way cement, Pioneer cement, flying cement, Fauji cement, and maple leaf cement are part of the North zone. Attock / falcon cement and lucky cement are part of the South zone.

A brief review of the literature related to this study is given below:

Zahid (2018) stated that Pakistan's cement sector is divided into two regions: North and South, Lucky Cement and Attock Cement are included in the South zone, while North zones include DG Khan Cement, Maple leaf cement, Best way cement and Askari cement etc. Lucky Cement has a large market share of 16% followed by Best Way Cement 12%, DG Khan Cement 13%, Askari cement 13%, Maple Leaf 7%, Fauji Cement 6% and 35 % others respectively.

Pakistan Cement Manufacture Association will use it to produce cement soon. Limestone and clay are the main materials but there are thirty different materials that can be used. Clay, shale, laterite, gypsum etc are the secondary raw material. Testing for the suitability of materials is important to identify their mineralogical and chemical properties (Shah M T 2007).

Shahnaz et al. (2013) stated that Portland cement is a type of hydraulic binder made by grinding cement clinker with one or more types of calcium sulfate. It is a mixture of silits, aluminum and calcium ferrites, with some negligible pollutants of low significance.

Edward examined the structure of natural resources like clinkers and concretes which is communicated in term of oxides of different components existing in the structures of clinker and concrete is portrayed by compound composition as opposed to oxides structure (2013).

Physical properties of Portland cement are controlled by chemical constituents that are existed in Portland cement. Tricalcium silicate (C3S), dicalciumsilicate (C2S), tricalcium aluminate (C3A), and tetra calcium aluminoferite (C4AF) are essential components of Portland cement. The reinforcement of hydrated cement paste is the main feature of silicates. C3A in cement accused of attacking sulfate. The presence of C4AF in minor quantities does not greatly affect the action of the cement (Pakistan Journal of Chemistry 2013).

For the edifice of buildings, highways and other local purposes, ordinary Portland Cement (OPC) is produced and used. (2019).

Existing literature review depicts exports and imports. These imported cements have not been evaluated according to scientific and economic standards dealing with production quality in order to assess their suitability for use in local conditions, which has led to a negative effect on the users, especially in the construction of buildings (Turkey Journal of African Earth Sciences).

Bogue proposed that catastrophic loss of life and property damage may be caused by the use of low quality cement in structural and building works. PC quality assurance is one of the essential and important factors. There are several PC brands available on the market, but they are made of identical chemical compositions. Changes in physical properties occur as a result of changes in the relative amounts of chemical elements. A chemical analysis of cement is approved for the purpose of testing whether or not the product supplied confirms the standard specifications. Usually every oxide is represented in percentages after the examination.

2. METHODOLOGY

The main focus of this proposed study is in the statistical analysis of the cement study in Pakistan. To illustrate the proposed methodology, we took data from different companies of Pakistan. We apply basic statistical analysis. Descriptive statistics are used to describe the basic features of the data in a study. For carrying out statistical analysis, descriptive analysis is considered as first principal. It provides understanding of data distribution, assistance to notice typos and outliers, and help you to recognize relation between variables; therefore you are enabling to deal with statistical Analysis, Descriptive techniques frequently involves table’s construction of mean, median, measures of dispersion, variance, and standard deviation which can differentiate relation between variables and applied to study many diverse hypotheses.

It is also known as average or representative score or measure central location. It is used to describe a distribution of scores by distributing a single value that is the center of the distribution. There are three main types of central tendency: Mean, Median, Mode. The variance is defined as average of the squared differences from the mean. It contains measuring of how data is distributed itself about mean or expected value in given set.

Graphical representation of data set is considered as essential task for interactive approach in graphics of statistics. This particular method permits us to display results in form of pictures or graphs. They are represented in form of control charts, scatter charts, pie charts and histogram charts.

3. RESULT AND DISCUSSION

Table-1 indicates that the average composition of all elements is between 10 to 12 for different companies. The composition of all elements is not same. Some of them use more while some are less. Since there is positive skew so most of the values lie above the average values i.e. mean. In addition similarly kurtosis values shown that our distribution of element is not normal.

Table 1
Chemical Composition of Various Brands of OPC

Descriptive	Askari	DG	Best-way	Pioneer	Flying	Falcon
Mean	11.062	11.023	11.137	11.195	11.482	10.978
Median	2.45	2.35	2.41	2.65	3.1	2.26
Std. Deviation	20.481	20.261	20.377	20.665	20.698	20.651
Standard Error	6.827	6.754	6.792	6.888	6.899	6.884
Variance	419.463	410.503	415.242	427.032	428.443	426.448
Skewness	2.573	2.563	2.588	2.601	2.612	2.552
Kurtosis	6.769	6.708	6.851	6.926	6.992	6.683

Table 2 indicates that the average composition of all elements is 14 to 16 for normal, sulphate and white cement. The composition of all elements is not same. Some of them use more while some are less. Since there is positive skew so most of the values lie above the average values i.e. mean. In addition similarly kurtosis values shown that our distribution of element is not normal.

Table 2
Composition of Normal, Sulphate-Resisting and white

Descriptive	Normal	Sulphate Resisting	White
Mean	14.456	14.448	15.316
Median	2.155	2.15	2.00
Std. Deviation	29.193	29.296	29.483
Standard Error	7.802	7.829	7.879
Variance	852.217	858.277	869.249
Skewness	2.378	2.393	2.221
Kurtosis	4.963	5.074	4.110

Table 3 indicates that the average composition of all elements is from 22 to 24 of mineral composition usually. The composition of all elements is not same. Some of them use more while some are less because the deviation of 23 to 30. Since there is positive skew so most of the values lie above the average values i.e. mean. In addition similarly kurtosis values shown that our distribution of element is not normal.

Table 3
Mineral Composition of Normal, Sulphate and white

Descriptive	Normal	Sulphate Resisting	White
Mean	22.75	22.725	23.1
Median	12.5	14	12.25
Std. Deviation	23.071	26.951	30.417
Standard Error	11.535	13.476	15.208
Variance	532.25	726.369	925.207
Skewness	1.882	1.657	1.592
Kurtosis	3.557	3.062	2.375

We also perform the regression analysis of the model by considering coefficients, SE coefficients, t static coefficient, p value, variance inflation factor (VIF) and results are reported in following **Table 4.**

The regression equation of y_1 versus $x_1 \ldots x_6$ Is:

$$y_{1i} = 0.184 - 0.040x_{1i} + 0.189\,x_{2i} + 2.33x_{3i} - 0.777x_{4i} - 0.614x_{5i} - 0.116x_{6i}$$

Table 4
Regression Results of Y_1 versus $X_1, X_2 \ldots X_6$

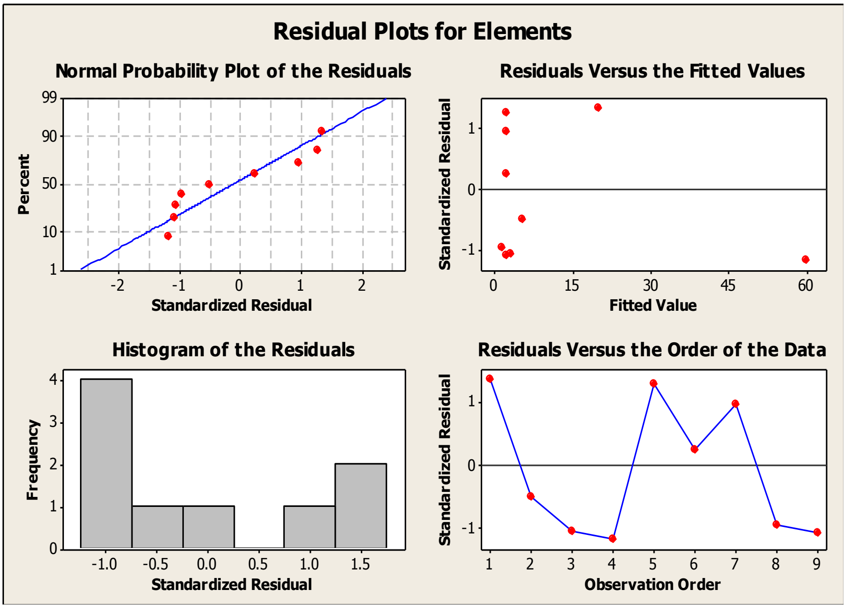
Predictor	Coef	SE Coef	T	P	VIF
Constant	0.18431	0.07592	2.43	0.136	
X_1	-0.0395	0.4544	-0.09	0.939	48780.9
X_2	0.1887	0.3067	0.62	0.601	21743.0
X_3	2.3335	0.3208	7.27	0.018	24063.1
X_4	-0.7775	0.2571	-3.02	0.094	15894.3
X_5	-0.6144	0.2030	-3.03	0.094	9941.7
X_6	-0.11658	0.9174	-1.27	0.332	2021.5

The sign of correlation coefficient guides us about direction of association.

Here y_{1i} is different chemical constituent average whereas $X_1 \ldots X_6$ are six different companies of Pakistan. Positive sign indicates that y_{1i} and x_{2i}, x_{3i},are directly related to each other. On the other hand, y_{1i} and, x_{1i}, x_{4i}, x_{5i}, x_{6i} have inverse relation.

Standard error (S)	0.119177	Obs*R-square (adj)	100.0%
Obs*R-square	100.0%	Durbin-Watson statistic	2.00645

We can calculate more precise view of correlation with the help of Adjusted R-squared, where independent variables are added to specific model. This modification is performed because independent variables increase the credibility of model and influence the additional variables that effect the R-squared observations. Durbin-Watson statistic results are noteworthy when predictor's variables appear to be insignificant. Durbin-Watson statistic indicates that our data is negative autocorrelation. Smallest standard error means that given data set has values which are closer to mean and larger standard error means values are not closer to mean.



One-way ANOVA: North Market Share versus South Market Share

Source	DF	SS	MS	F	P
x_7	3	50.3	16.8	0.87	0.637
Error	1	19.2	19.2		
Total	4	69.5			

In these results, the null hypothesis states the p-value is 0.637, which is greater than the significance level of 0.05; we cannot reject the null hypothesis and conclude that mean is same.

Standard error (S)	4.384	Obs*R-square (adj)	0.00%
Obs*R-square	72.36%	Durbin-Watson statistic	

Here tabulated value of χ^2 is 72.36%, which are greater than adjusted value. 0.00% is more précised. The values in data set are closer to mean.

Pie Chart

The following pie charts show the proportion of north and south marketing share of different companies.

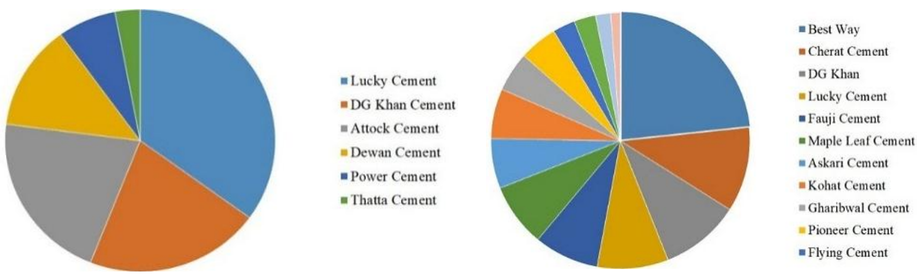


Figure 1: North and south market share

Following figures show the chemical composition percentage of normal, sulphate-resisting and white Portland cement. SiO_2, Cao and LSF are the main constituents.

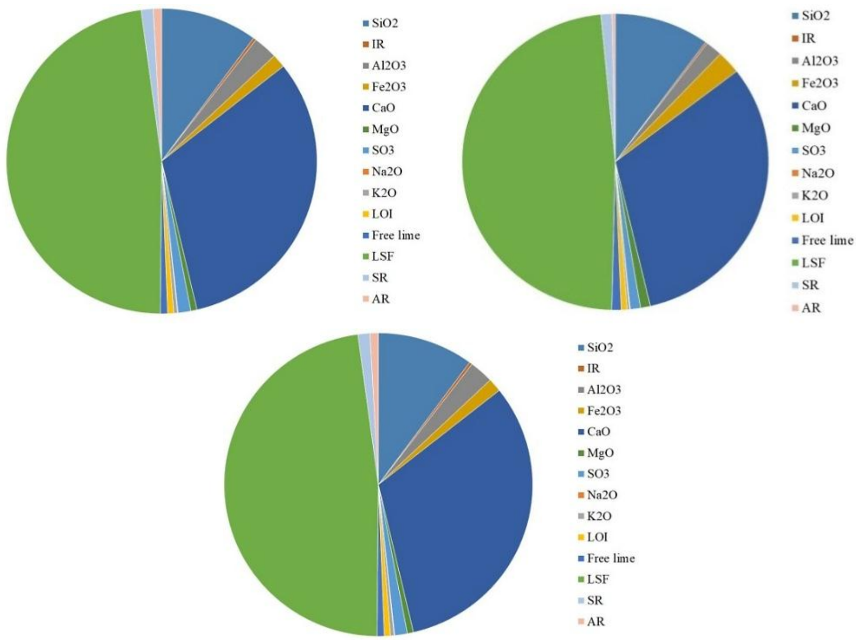


Figure 2: Normal, Sulphate Resisting and White Cement

Following figures show the mineral composition of percentage of normal, Sulphate-resisting and white cement. Four main mineral compounds of OPC. C3S is the main constituent of cement. 57% C3S in normal, 62% in sulphate resisting and 67% in white cement.

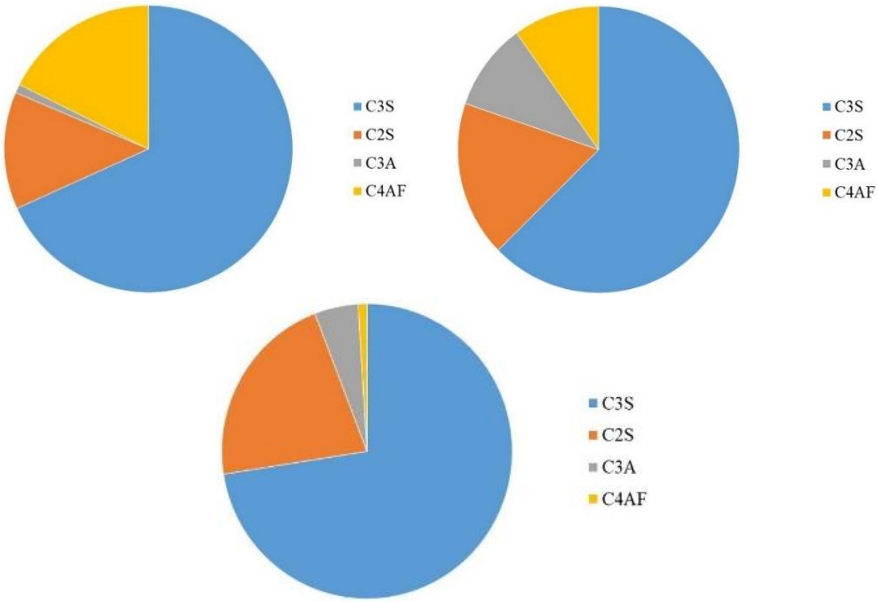


Figure 3: Normal, Sulphate Resisting, White Cement

4. COMMENTS AND CONCLUSION

Statistical relation towards OPC has been formulated. This study is all about chemical composition of cement study in Pakistan. So for this study we took data from various brands of cements available throughout the Pakistan. Chemical composition data and normal, sulphate resisting and white cement data of various brands was taken. Then we applied various statistical techniques to determine whether the cement produced in Pakistan is up to a standard quality or not. Pie charts and various other tools were used for the sake of this study and results obtained from our analysis were as expected. Most of the cement brands working in Pakistan have similar chemical composition, there are minor differences in the chemical compositions, and not a single brand has something special or different from the rest, so the result we got through our analysis shows that all the data of different companies use nearly same chemical composition. So we can say that the quality of cement produced by various companies in Pakistan is up to standard and that's appreciated.

REFERENCES

1. Taylor, H.F.W. (1980). *Cement chemistry, 2nd Ed. Thomas Telford*, London.
2. Ali, M.S., Khan, I.A. and Hossain, M.I. (2008). Chemical analysis of ordinary Portland cement of Bangladesh. *Chemical Engineering Research Bulletin*, 12, 7-10.
3. Hwang, Y.C., Lee, S.H., Hwang, I.N., Kang, I.C., Kim, M.S., Kim, S.H., Son, H.H. and Oh, W.M. (2009). Chemical composition, radiopacity, and biocompatibility of Portland cement with bismuth oxide. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 107(3), e96-e102.

4. Huntzinger, D.N. and Eatmon, T.D. (2009). A life-cycle assessment of Portland cement manufacturing: comparing the traditional process with alternative technologies. *Journal of Cleaner Production*, 17(7), 668-675.
5. Jaffar, A., Ali, N., Anwer, M., Naeem, M.A., Riaz, M.R., Jamal, M.H., Hussain, A. and Ming, X. (2014). The impact of Cement Manufacturing on Economic Development of Pakistan: The Analysis of Cement Industry. *International Journal of Research*, 2, 205-220.
6. Hussain, Z., Khan, N.M., Shah, S.M.A., Nawaz, F., Khan, M.N., Hussain, S. and Ahmad, I. (2018). Investigation of raw materials for cement industry of Upper Hunza, Gilgit-Baltistan, Pakistan. In *IOP Conference Series: Materials Science and Engineering*, 414(1), 012012. IOP Publishing.
7. Reyes, R.S.A., Díaz, A.A., Rodríguez, S.B., Rodríguez, C.A.L. and Hernández, J.F.M. (2018). Assessment of Cuban Kaolinitic Clays as Source of Supplementary Cementitious Materials to Production of Cement Based on Clinker–Calcined Clay–Limestone. *In Calcined clays for sustainable concrete* (pp. 21-28). Springer, Dordrecht.
8. Shah, M.T., Fayez, A. and Ali, L. (2007). Chemical study of the raw material in Gandghar range, district Haripur, NWFP, for Portland cement manufacturing. *Journal of the Chemical Society of Pakistan*, 29(2), 103-110.
9. Shahnaz, A., Shahzadi, P., Khan, M. S., Rana, B. B., Chaudry, M. Y., & Yasmeen, S. (2013). A Short Study to Test the Compliance of Various Pakistani Ordinary Portland Cements with ASTM Composition Standards. *Pak. J. Chem*., 3(4), 1-7.
10. Hewlett, P. (2003). *Lea's chemistry of cement and concrete*. (4^{th} Ed.) Butterworth-Heinemann.
11. Austin, G.T. (1984). *Shreve's chemical process industries*. McGraw-Hill Companies.
12. Australian Standard 2350. *2/80-Chemical Composition of Portland Cement*, 2006.
13. Özgüner, A.M. (2014). Prospection of Portland cement raw material: A case study in the Marmara region of Turkey. *Journal of African Earth Sciences*, 97, 230-243.
14. Bogue, R.H. (1955). *The Chemistry of Portland Cement*, 2nd Ed. Reinhold Publishing Corporation.
15. Bogue, R.H. (1929). Calculation of the compounds in Portland cement. *Industrial & Engineering Chemistry Analytical Edition*, 1(4), 192-197.
16. Pandey, G.N. and Shukla, S.D. (1980). *A Text Book of Chemical Technology*. Vol. 1. Vikas Publishing House, New Delhi.
17. Duda, W.H. (1984). *Cement Data Book: Automation, Storage, Transportation, Dispatch*, Vol. 2. Bauverlag GmbH, Berlin.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 217-227

STUDY OF ONLINE EDUCATION

Ammara Nawaz Cheema and **Haseeba Sajjad**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Education is an important part of people's lives. A person with a good education can do good in his life. If we talk about online education, it is a very fresh way to teach students. In the early history there was nothing such like online classes. Overview the history we collect data on the thinking of students that how they take studies as online or onsite. Our findings conclude that online education is very beneficial for the students who work due to additional family responsibilities. Online learning is distinct learning and can help a lot of people who want to learn new things staying home, but on the same time it can be really bad for some people who do not have a good internet connection in their back yard.

1. INTRODUCTION

Education is an important part of people's lives. A person with a good education can do well in his life. If we talk about online education, it is a very fresh way to teach students. In the early history there was nothing such like online classes. To be honest there were very less number of education institute, which were only afforded by the lead class of people. According to the given history in the research paper, pervious history of 1920's up to 1990's it was observed that there was nothing difference between learning outcomes which were achieved by the students of different age which were engaged in online classes as well as when they were having onsite education. This shows that there was less interest towards technology in the past. As soon as technology was welcomed the world started to prefer the online learning as well. In 2000's several studies were conducted online and about 2010 it was easy for every individual to conduct the online classes and as well as some institute were able to conduct online exams as well. In 1993, online education became legal. Watching the history we suggested to collect data on the thinking of students that how they take studies as online or onsite. Online education is very beneficial for the students who work due to additional family responsibilities. Online learning is distinct learning and can help a lot of people who want to learn new things staying home, but on the same time it can be really bad for some people who do not have a good internet connection in their back yard.

[1] As nicely taken into consideration doctrine and culture in learner on line classes. Even though their findings supported the motive that on line and normal college students completed likewise suitably at the pattern tests. Further currently, [2] deliberate education results and apprentice pleasure in an online man or women occurrence rate circulate and in a routine person to person pathway. On the other hand, [3-6] entirely second hand a little bigger phase sizes over longer intervals of time, however the situation regions listening cautiously on each thing from Shakespeare to laptops to element and have been [7] Inquiry

turned into one of a small quantity of examine articles that listening carefully on a regulate desire in education, however for a second time with a diminutive mock-up measurement of 32 person to person learner and 29 on-line learners. Evaluation of Schulte's toil extra supported those findings. Online understanding is a device of nearness know-how or separate training, which has gradual been a duty of the American edification gadget, and it has turn out to be the top sector of expanse lifestyle in latest being [8-9]. Innocently online publications are publications delivered complete through web, and mix or combine tradition merge recurring person to person training, information over the Internet, and scholarship supported by different technology [10-11]. One basis why in attendance is consequently a big amount communicate around understanding is that to hand are ratings of self-styled earnings and makes use of on-line mastering. Many of the specifically key points : success in teaching learners , its gain from as certified improvement, its successfulness to action the growing rate tag of higher studies training, believe equality at the higher learning stage, [12 -14].

[15] Examine the barriers to online teaching from a distribution point of view and also listed a number of online education benefits and drawbacks.[16] presented a list of 10 guidelines for teaching a course online. Furthermore, several volumes h written by scientist that include records on promoting onscreen information gain (For example, Collison, Elbaum, Haavind, Tinker, 2000; Salmon, 2004).

[17]looked at matters of online meeting/commitment.[18]was focused on how online postpones can be encourage[19] discussed online learning constructivist while [20] talked about building a organized online research environment.[21]talks about 37 blanketed undergraduate students who had been unplanned give on the web or in class gathering.

2. METHODOLOGY AND ANALYSIS

Table 1
Descriptive Analysis of Variables

	Gender	First Language	Age	Student/ Employ	Institute
Mean	1.68	1.45	20.92	1.04	2.38
Median	2.00	1.00	21.00	1.00	2.00
Mode	2	1	21	1	1
Std. Deviation	.520	.762	2.328	.184	1.119
Variance	.271	.580	5.421	.034	1.252
Skewness	-.202	1.651	2.366	5.099	.163
Kurtosis	-.831	1.944	16.064	24.239	-1.335
Range	2	3	24	1	3

(Gender) on average response is 1.68. Median and mode tells us that mostly people responses are female. Maximum responses are female. Variance is 0.271 and standard deviation is 0.520. Skewness shows negative and kurtosis platykurtic for this variable.

(First language) on average response is 1.45. Median and mode tells us that mostly people responses are in urdu. Maximum responses are in urdu. Variance is 0.580 and standard deviation is 0.762 Skewness shows positive and kurtosis leptokurtic for this variable.

(Age) on average response is 20.92. Median and mode tells us that mostly people responses are 21 year old. Maximum responses are in 24. Variance is 5.421 and standard deviation is 2.328. Skewness shows positive and kurtosis leptokurtic for this variable.

(Student/employee) on average response is 1.04. Median and mode tells us that mostly people responses are students. Maximum responses are students. Variance is 0.034 and standard deviation is 0.184 Skewness shows positive and kurtosis leptokurtic for this variable.

Table 2
Descriptive Analysis of Variables

	Credit Hour	Online GPA	Onsite GPA	Accessibility of internet	Online Commu nication
Mean	5.13	3.01	3.06	2.63	2.39
Median	5.00	3.00	3.00	3.00	2.00
Mode	6	3	3	3	3
Std. Deviation	1.221	.716	.699	.740	.774
Variance	1.491	.512	.489	.547	.600
Skewness	-.871	-.596	-.617	-.401	-.134
Kurtosis	1.290	.655	.809	-.049	-.492
Range	7	3	3	3	3

(Credit hour) on average response is 5.13. Median and mode tells us that mostly people responses are 5 and 6. Maximum responses are 7. Variance is 1.491 and standard deviation is 1.221. Skewness shows negative and kurtosis leptokurtic for this variable.

(Online GPA) on average response is 3.01. Median and mode tells us that mostly people responses are in 2.50-3.50. Maximum responses are in 2.50-3.50. Variance is 0.512 and standard deviation is 0.716 Skewness shows negative and kurtosis leptokurtic for this variable.

(Onsite GPA) on average response is 3.06. Median and mode tells us that mostly people responses are 2.50-3.50. Maximum responses are in 2.50-3.50. Variance is .489 and standard deviation is .699 Skewness shows negative and kurtosis leptokurtic for this variable.

(Accessibility of internet) on average response is 2.63. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.547 and standard deviation is 0.740 Skewness shows negative and kurtosis platykurtic for this variable.

(Online communication) on average response is 2.39. Median tell us that mostly peoples responses are disagree and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.600 and standard deviation is 0.774 Skewness shows negative and kurtosis platykurtic for this variable.

Table 3
Descriptive Analysis of Variables

	Written Communication	Time Management	Communicate Electronically	Group Working	Interaction with Instructor
Mean	2.69	2.65	2.53	2.90	2.71
Median	3.00	3.00	3.00	3.00	3.00
Mode	3	3	3	3	3
Std. Deviation	.829	.729	.814	.796	.782
Variance	.687	.532	.663	.633	.611
Skewness	-.275	-.359	-.166	-.544	-.258
Kurtosis	-.403	-.023	-.466	.091	-.252
Range	3	3	3	3	3

(Written communication) on average response is 2.69. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .687 and standard deviation is .829 Skewness shows negative and kurtosis leptokurtic for this variable.

(Time management) on average response is 2.65. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is 0.532 and standard deviation is 0.729 Skewness shows negative and kurtosis leptokurtic for this variable.

(Communicate electronically) on average response is 2.53. Median and mode tells us that mostly people responses are agree. Maximum responses are in agree. Variance is .663 and standard deviation is .814 Skewness shows negative and kurtosis leptokurtic for this variable.

(Group working) on average response is 2.90. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.633 and standard deviation is 0.796 Skewness shows negative and kurtosis leptokurtic for this variable.

(Interaction with instructor) on average response is 2.71. Median tell us that mostly peoples responses are agree and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.611 and standard deviation is 0.782 Skewness shows negative and kurtosis platykurtic for this variable.

Table 4
Descriptive Analysis of Variables

	Communication in English	Face to Face Interaction	Quick Response	Learning	Interact Activity
Mean	2.45	3.05	2.38	2.03	2.53
Median	3.00	3.00	2.00	2.00	3.00
Mode	3	3	2	2	3
Std. Deviation	.748	.841	.824	.888	.795
Variance	.560	.707	.679	.788	.633
Skewness	-.283	-.598	.009	.430	-.446
Kurtosis	-.396	-.225	-.560	-.692	-.374
Range	3	3	3	3	3

(Communication in English) on average response is 2.45. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .560 and standard deviation is .748 Skewness shows negative and kurtosis platykurtic for this variable.

(Face to face interaction) on average response is 3.05. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is 0.707 and standard deviation is 0.841 Skewness shows negative and kurtosis platykurtic for this variable.

(Quick response) on average response is 2.38. Median and mode tells us that mostly people responses are disagree. Maximum responses are in agree. Variance is .679 and standard deviation is .824 Skewness shows positive and kurtosis platykurtic for this variable.

(Learning) on average response is 2.03. Median and mode tells us that mostly people responses are disagree. Maximum responses are agree. Variance is 0.788 and standard deviation is 0.888 Skewness shows negative and kurtosis platykurtic for this variable.

(Internet activity) on average response is 2.53. Median and mode tell us that mostly peoples responses are agree. Maximum responses are agree. Variance is 0.633 and standard deviation is 0.795 Skewness shows negative and kurtosis platykurtic for this variable.

Table 5
Descriptive Analysis of Variables

	Learning on Internet	Motivation	Classroom Environment	Personal Feedback	Understanding
Mean	2.28	2.47	2.77	2.72	2.62
Median	2.00	3.00	3.00	3.00	3.00
Mode	2	3	3	3	3
Std. Deviation	.852	.801	2.297	.710	.799
Variance	.725	.642	5.276	.504	.639
Skewness	.168	-.151	11.544	-.728	-.339
Kurtosis	-.603	-.477	152.470	.583	-.281
Range	3	3	3	3	3

(Learning on internet) on average response is 2.28. Median and mode tells us that mostly people responses are disagree. Maximum responses are agree. Variance is .725 and standard deviation is .852 Skewness shows positive and kurtosis platykurtic for this variable.

(Motivation) on average response is 2.47. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is .642 and standard deviation is 2.801 Skewness shows negative and kurtosis platykurtic for this variable.

(classroom-Environment) on average response is 2.77. Median and mode tells us that mostly people responses are agree. Maximum responses are in agree. Variance is 5.276 and standard deviation is 2.297. Skewness shows positive and kurtosis leptokurtic for this variable.

(Personal-feedback) on average response is 2.72. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.504 and standard deviation is 0.710 Skewness shows negative and kurtosis leptokurtic for this variable.

(Understanding) on average response is 2.62. Median and mode tell us that mostly peoples responses are agree. Maximum responses are agree. Variance is 0.639 and standard deviation is 0.799 Skewness shows negative and kurtosis platykurtic for this variable.

Table 6
Descriptive Analysis of Variables

	Grades	**Computer Timing**	**Face to Face Instruction**	**Instructor Understanding**	**Technology**
Mean	2.59	2.29	2.92	2.70	2.68
Median	3.00	2.00	3.00	3.00	3.00
Mode	3	2	3	3	3
Std.Deviation	.752	.865	.867	.868	.770
Variance	.566	.748	.752	.754	.592
Skewness	-.365	.025	-.489	-.307	-.303
Kurtosis	-.164	-.788	-.376	-.513	-.172
Range	3	3	3	3	3

(Grades) on average response is 2.59. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .566 and standard deviation is .752. Skewness shows negative and kurtosis platykurtic for this variable.

(Computer timing) on average response is 2.29. Median and mode tells us that mostly people responses are disagree. Maximum responses are in agree. Variance is 0.748 and standard deviation is 0.865. Skewness shows positive and kurtosis platykurtic for this variable.

(Face to face instruction) on average response is 2.38. Median and mode tells us that mostly people responses are disagree. Maximum responses are in agree. Variance is .679 and standard deviation is .824 Skewness shows positive and kurtosis platykurtic for this variable.

(Instructor-understanding) on average response is 2.70. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.754 and standard deviation is 0.868 Skewness shows negative and kurtosis platykurtic for this variable.

(Technology) on average response is 2.68. Median and mode tell us that mostly peoples responses are agree. Maximum responses are agree. Variance is 0.592 and standard deviation is 0.770 Skewness shows negative and kurtosis platykurtic for this variable.

Table 7
Descriptive Analysis of Variables

	Expectations	**Problem Solving**	**High Quality Learning**	**Teaching Environment**	**Onsite Classes GPA**
Mean	2.42	2.89	2.46	2.80	2.74
Median	3.00	3.00	3.00	3.00	3.00
Mode	3	3	3	3	3
Std.Deviation	.828	2.206	.912	.846	.909
Variance	.686	4.866	.832	.717	.827
Skewness	-.262	11.632	-.066	-.448	-.190
Kurtosis	-.663	154.165	-.812	-.274	-.794
Range	3	31	3	3	3

(Expectations) on average response is 2.42. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .686 and standard deviation is .828. Skewness shows negative and kurtosis platykurtic for this variable.

(Problem-solving) on average response is 2.89. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is 4.866 and standard deviation is 0.912 Skewness shows positive and kurtosis leptokurtic for this variable.

(High-quality-learning) on average response is 2.46. Median and mode tells us that mostly people responses are agree. Maximum responses are in agree. Variance is .832 and standard deviation is .912. Skewness shows negative and kurtosis platykurtic for this variable.

(Teaching – environment) on average response is 2.80. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.717 and standard deviation is 0.846 Skewness shows negative and kurtosis platykurtic for this variable.

(Onsite-classes GPA) on average response is 2.74. Median and mode tell us that mostly peoples responses are agree. Maximum responses are agree. Variance is 0.827 and standard deviation is .909. Skewness shows negative and kurtosis platykurtic for this variable.

Table 8
Descriptive Analysis of Variables

	Technical Skills	Assignments	Interpersonal Skills	Assignments During Online Classes	Quizzes During Online Classes
Mean	2.72	2.90	2.78	2.67	2.65
Median	3.00	3.00	3.00	3.00	3.00
Mode	3	3	3	3	3
Std.Deviation	.853	2.222	.805	.828	.819
Variance	.727	4.939	.648	.686	.671
Skewness	-.447	11.350	-.562	-.377	-.270
Kurtosis	-.316	149.386	.058	-.319	-.377
Range	3	31	3	3	3

(Technical skills) on average response is 2.72. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .727 and standard deviation is .853. Skewness shows negative and kurtosis platykurtic for this variable.

(Assignments) on average response is 2.90. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is 4.939 and standard deviation is 0.805 Skewness shows positive and kurtosis leptokurtic for this variable.

(Interpersonal-skills) on average response is 2.78. Median and mode tells us that mostly people responses are agree. Maximum responses are in agree. Variance is .648 and standard deviation is .805. Skewness shows negative and kurtosis platykurtic for this variable.

(Assignments during online classes) on average response is 2.67. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is 0.686 and standard deviation is 0.828. Skewness shows negative and kurtosis platykurtic for this variable.

(Quizzes during online classes) on average response is 2.65. Median and mode tell us that mostly peoples responses are agree. Maximum responses are agree. Variance is 0.671 and standard deviation is 0.819 Skewness shows negative and kurtosis platykurtic for this variable.

	Onsite Assignment Percentage	Onsite Quizzes Percentage
Mean	3.00	2.66
Median	3.00	3.00
Mode	3	3
Std.Deviation	.805	.780
Variance	.648	.609
Skewness	-.574	-.207
Kurtosis	.008	-.296
Range	3	3

(Onsite-assignment percentage) on average response is 3.00. Median and mode tells us that mostly people responses are agree. Maximum responses are agree. Variance is .648

and standard deviation is .805. Skewness shows negative and kurtosis leptokurtic for this variable.

(Onsite-quizzes – percentage) on average response is 2.66. Median and mode tells us that mostly people responses are in agree. Maximum responses are in agree. Variance is .609 and standard deviation is 0.780 Skewness shows negative and kurtosis platykurtic for this variable.

Variable of this graph is face to face interaction. According to this graph 38.2% are strongly agree and 44.4% people are agree moreover 3.2% are strongly disagree and 13.5 % are disagree with question. The question i feel that face-to-face contact with my instructor is necessary to learn.

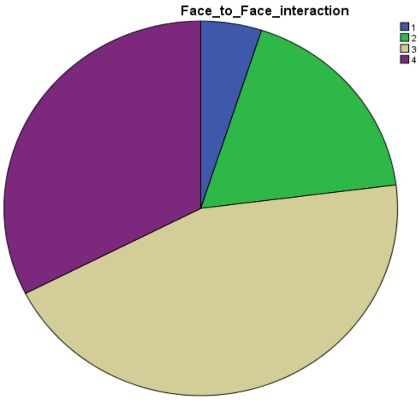


Fig 1: Pie Chart of Face to Face Interaction

Variable of this graph is internet activity. According to this graph 9.7% are strongly agree and 35.3% people are agree moreover 13.5% are strongly disagree and 41.5 % are disagree with question. The question I can ask my teacher questions and receive a quick response during Internet activities outside of class.

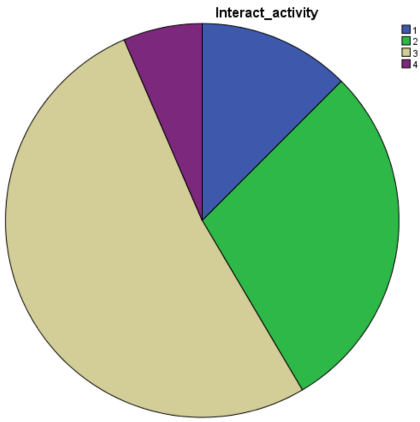


Fig 2: Pie Chart of Learning on Internet

Variable of this graph is motivation. According to this graph 7.7% are strongly agree and 35.3% people are agree moreover 20.3% are strongly disagree and 42.5 % are disagree with question. The question I believe that learning on the Internet outside of class is more motivating than a regular course.

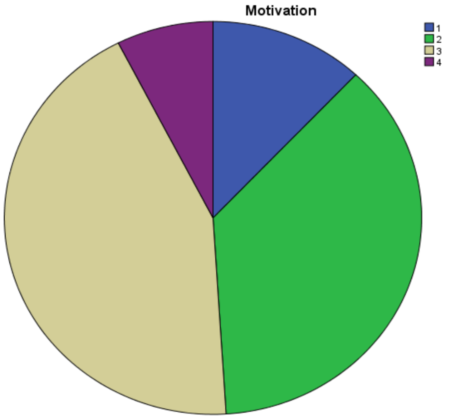


Fig 3: Pie Chart of Motivation

4. COMMENTS AND CONCLUSION

- The recommendation made in this study is helpful for educational institutes to analyses the quality of education.
- The selection of variables that contributes more towards online classes to improve quality of education and make more policies that helps teachers to perform better in future.
- This research work also facilitates the administration to judge the enrollment of online classes in future.
- This study brings more ease in interpreting the online and insight teaching benefits which helps in to improve the process, and tools used to improve the teaching online.

REFERENCES

1. Davies, R.S. and Mendenhall, R. (1998). *Evaluation Comparison of Online and Classroom Instruction for HEPE 129--Fitness and Lifestyle Management Course*. Evaluation Program, Department of Instructional Psychology and Technology. Provo, UT: Brigham Young University. (ERIC Document Reproduction. Service No. ED 427752).
2. Johnson, S.D., Aragon, S.R. and Shaik, N. (2000). Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments. *Journal of Interactive Learning Research*, 11(1), 29-49.
3. Connolly, T.M., MacArthur, E., Standsfield, M. and McLellan, E. (2005). A quasi experimental study of three online learning courses in computing. *Computers and Education*, 49(2), 345-359.

4. McCollum, K. (1997). A professor divides his class in two to test value of on-line instruction. *Chronicle of Higher Education*, 43(24), 23.
5. Kartha, C.P. (2006). Learning business statistics: Online vs traditional. *The Business Review*, 5(1), 27-32.
6. Koory, M.A. (2003). Differences in learning outcomes for the online and F27 versions of “An introduction to Shakespeare”. *Journal of Asynchronous Learning Networks*, 7(2), 18-35.
7. Legutko, R.S. (2007). Face-to-face or cyberspace: Analysis of course delivery in a graduate educational research course. *MERLOT Journal of Online Learning and Teaching*, 3(3), 288-294.
8. Bartley, S.J. and Golek, J.H. (2004). Evaluating the cost effectiveness of online and face-to-face instruction. *Journal of Educational Technology & Society*, 7(4), 167-175.
9. Evans, J.R. and Haase, I.M. (2001). Online business education in the twenty-first century: an analysis of potential target markets. *Internet Research,* 11(3), 246-260.
10. Hoic-Bozic, N., Mornar, V. and Boticki, I. (2009). A blended learning approach to course design and implementation. *IEEE Transactions on Education*, 52(1), 19-30.
11. Osguthorpe, R.T. and Graham, C.R. (2003). Blended Learning Environments: Definitions and Directions. *Quarterly Review of Distance Education*, 4(3), 227-233.
12. de la Varre, C., Keane, J. and Irvin, M.J. (2011). Enhancing Online Distance Education in Small Rural US Schools: A Hybrid, Learner-Centred Model. *Journal of Asynchronous Learning Networks*, 15(4), 35-46.
13. Gratton-Lavoie, C. and Stanley, D. (2009). Teaching and learning principles of Microeconomics online: An empirical assessment. *The Journal of Economic Education*, 40(1), 3-25.
14. Lorenzetti, J. (2013). *Academic administration-running a MOOC: Secrets of the world’s largest distance education classes*. Magna Publication, Wisconsin.
15. Berge, Z. (1998). Barriers to online teaching in post-secondary institutions: Can policy changes fix it? *Online Journal of Distance Learning Administration,* 2(1), 1-22.
16. D'Orsie, S.M. and Day, K. (2006). Ten tips for teaching a web course. *Tech. Directions*, 65(7), 18-20.
17. Oliver, R. (1999). Exploring strategies for online teaching and learning. *Distance Education*, 20(2), 240-254.
18. Tuckman, B.W. (2005). The effect of motivational scaffolding on procrastinators’ distance learning outcomes. *Computers and Education*, 49(2), 414-442.
19. Taylor, P. and Maor, D. (2000). Assessing the efficacy of online teaching with the constructivist on-line learning environment survey. In: *9th Annual Teaching Learning Forum*, 2-4 February 2000, Perth, W.A.
20. Waltonen-Moore, S., Stuart, D., Newton, E., Oswald, R. and Varonis, E. (2006). From virtual strangers to a cohesive learning community: The evolution of online group development in a professional development course. *Journal of Technology and Teacher Education*, 14(2), 287-311.
21. Schutte, J.G. (1997). *Virtual Teaching in Higher Education: The New Intellectual Superhighway or Just Another Traffic Jam*? Online available at http://www.csun.edu/~egodard/dta/docs/P-VirtualTeaching.pdf

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 229-236

SUICIDE ATTEMPTS – A CASE STUDY OF PAKISTAN

Ammara Nawaz Cheema and **Iqra Farooq**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Suicide is an illogical hunger to die. We use the word "illogical" here because no matter what happens how bad a person's life is, suicide is an everlasting solution to what is absolutely always an impermanent problem. Suicide is an indication and sign of severe depression. Depression is a recoverable disorder, but often the therapy takes time, energy, and determination on the part of the person who is feeling discouraged. Much considerations, work, and research is needed to properly understand the suicidal behavior of young generation and why they indulge in such self-harming actions. We apply basic statistical analysis; descriptive techniques, graphical methods, to draw inference about the suicide attempts in Pakistan.

1. INTRODUCTION

Suicide is an illogical hunger to die. We use the word "illogical" here because no matter what happens, how bad a person's life is, suicide is an everlasting solution to what is absolutely always an impermanent problem. Suicide is an indication and sign of severe depression. Depression is a recoverable disorder, but often the therapy takes time, energy, and determination on the part of the person who is feeling discouraged. Suicide's effects are tragic and felt long after the individual has taken his or her own life. It is usually the second or third leading cause of death among teenagers, and remains one of the top ten leading causes of death well into middle age. A person who dies by suicide leaves behind them a tangled confusion of family members and friends who try to make sense of a senseless and purposeless act. Suicide is a serious public health problem; however, suicides are preventable with timely, evidence-based and often low-cost interventions. For national responses to be effective, a comprehensive multisectoral suicide prevention strategy is needed. It is estimated that around 20% of global suicides are due to pesticide self-poisoning, most of which occur in rural agricultural areas in low- and middle-income countries. Other common methods of suicide are hanging and firearms.

1. According to "Chaleby", Suicide is the most disgraced act in the mainstream religions of the globe. All the well-known religions like Judaism, Buddhism, Christianity, etc. ventured to publically declare suicide as wrongdoing and evil. Islam, in particular, imparts significant importance to the matter by stating that whoever commits suicide, he/ she will be damned to Hell. In the light of the Islamic teachings, the Islamic scholars have held their stance that performing suicide is undoubtedly an unjustifiable sin.

2. The difference between conducted and tried suicide must be understood since the difference between the two is a subject of great confusion among the population

guided by strict Islamic stature. For the police inquiry and proper trial of a suicide attempter to take place, it is mandatory that proper laws and restrictions pertaining the act of suicide attempt must be put in place.

3. According to “Soni Raleigh”, in most of the non-Muslim countries of the world, the rate of suicide attempts carried out by the Muslims is marginally less than the non-Muslim general population. It is estimated that 26 percent of the male and 16 percent of the female population of Muslims commit suicide in England/ Whales. On the other hand, the Hindus/ Sikhs account for 69 percent male and 83 percent female suicide attempters among the immigrant people of India.

4. According to a study among the Indian foreigners in Fiji, “Haynes”, in 1984 said, although Hindus were 5 times more abundant in the region than Muslims at that time, the suicide rate of Hindu male population exceeded that of Muslims by 14 times, whereas the Hindu female suicide cases were 8 times more frequent than Muslim women respectively.

5. Other studies related to the statistical study of suicide and suicide attempts indicate that among the Indian settlers in the region of South Africa, Hindus comprised of almost 67 percent of the total population, while they engaged in the act of taking their own life by as much as 79 percent when it comes to the investigation of 50 continuous suicide cases. Unfortunately, very few studies and analysis have been done regarding the suicide/ chanced suicides in Pakistan and few studies have been made public.

6. Ali et al and Reza et al indicate that women are engaging in suicide more than men in Pakistan and most of them lie under the age of thirty. However, the aspect that really sets Pakistan apart from the Western Nations is the increased number of married women partaking in this act. Marriage, instead of serving as a source of emotional satisfaction and companionship, is becoming a cause of great stress and anxiety among the Pakistani women which is inducing severe suicidal thoughts in their minds.

7. According to “Bukhari”, the major contributor to the young married women committing in suicidal behavior is the domestic pressure, household violence, and disputes among the family (In-laws). This is a huge problem in Pakistan and is very dominant. Another contributor is the readily available solutions that are poisonous to the human body and are extremely fatal in most cases. The ease of availability of these drugs make them the most commonly abused material for the commitment of suicide. Such commonly used poisonous solutions include benzodiazepines and organophosphates. Hospital and clinical records within Pakistan reveal that there are 36,232 cases of deaths/ illnesses in Pakistan due to toxic effects. Extrapolating from the already available data and statistics, a suggestive measure would indicate that 25,000 out of 36,232 (70 percent) are in fact, cases of self-infliction of poison.

8. This data was compiled by “Jamil” encompassing only Karachi. The figure is bound to multiply many times over if the region of study was extended all across Pakistan.

9. According to the 2 reports by “Ahmed” and “Zuberi” encompassing a period of five years each, from 1959 to 1963 and from 1973 to 1977 respectively, from the Karachi

Police documentations, it is revealed that there were 65 men and 24 women, and 14 men and 11 women in both, respectively. It is found that there are more cases of attempted suicide by newlywed women than the actual mortality rate of men due to suicide. Out of these female suicide attempters, 75 percent are under the age of twenty-five. The causes include domestic violence, unhappy love-affairs, and disputes with the new family (in-laws). The methods utilized, in the order of decreasing priority included poisoning, hanging, using lethal or sharp weapons, and by burning.

10. In an unpublished work by “Khan”, suicide study between 1996 and 1997 enlisted 306 suicides. The study encompassed a surplus of thirty-five cities and their vicinity. The ratio of suicides committed by men to the women came out to be 2:1. Twenty-nine percent men were married as compared to fifty-nine percent married women. Household issues were the most prominent reason of suicide commitment and the mode of suicide was almost always poison ingestion. Most of these suicide committers were under thirty years old.

11. Diekstra (1993) predicts that dramatic growth in the suicide casualties will be seen in the next decade, especially in the under-developed countries and countries like Pakistan. The under-developed countries are already suffering severely from intense socio-economic deficiencies and they are only subject to increase as compared to most industrialized countries. Sanctions imposed on Pakistan relating to its recent nuclear tests also put additional financial/ economic pressure on it. How would these issues affect the behavioral response of the citizens regarding suicide commitment and will Islam be able to defend its followers from the ever-increasing curse of suicide? The answer to both these questions is subject to new form of research.

12. Pakistan is a developing country of South Asia. It is a predominantly Muslim country in which advancement has continued slowly. The difference on the basis of past, caste, creed, social background, and gender has led to broad deprivation.

13. Mental ailment has been very common in Pakistan with almost 34% of the population suffering from common mental disorders and one percent suffer from mental illness, a publication by “Mirza & Jenkins”, 2004.

14. Child health assistance in Pakistan has suffered from poverty severely. There are next to no institutions that may work on the emotional, social, and behavioral aspects of children. An estimate by “Javad, Kundi & Khan”, in 1992 concluded that about 9.3 percent completely detrimental problems have been made an unavoidable part of the system. These problems can only be resolved by means of proper distribution of resources and employment, which is hindered by the lack of exploration in these fields and the absence of statistical data.

15. Suicide remains as one of the most common cause of expiration (From age 10-64). It ranks between 2 to 4 (From age 10-45) (McKeown, Cuffe, and Shultz, 2006). Risk for suicide mainly arises from the following factors: 1. Extreme depression, 2. Bipolarity, 3. Drug use/ abuse, 4. Past experience of suicide attempt, 5. Aggressive/

destructive ideation. Anxiety also plays a major role in the attempt of suicide. Stress, anxiety, pressure, failures, and frauds induce suicidal thoughts in a person.

16. Disruptive disorder is often induced by material abuse. Academic pressure asserted by the educational institutions play a key role in increasing stress and anxiety among the youth (Nickel et al., 2006).
17. Taiwanese have been prone to extreme depression, according to "Lin & Lu" in 2006, which led to suicides in the country.
18. India had been suffering severely from the curse of suicide, mainly in females of age 15 to 19. Negligence by parents/ Guardians, Physical and mental abuse, toxic love affairs, and running away from school, etcetera, account for the major reasons of suicide in India (DSH; Sidhartha & Jena, 2006).
19. Pakistan lacks in statistical data regarding suicides. Of what little studies have been published, the suicides are dominantly committed by married women under the age of 24 and the common risk factors are household stressors, family dispute, etcetera (Khan & Reza, 1998).
20. This study on common risk factors on children and adolescents is purely based upon diagnosis given in case notes. Suicides in Pakistan is led by female gender and just like India, the females are the ones that openly expressed their desire to end their life, however males outnumber females for suicides in India (60%). The difference between gender population for both India and Pakistan is not significant enough for us to draw any important conclusions. Pakistan is a developing country where orthodox thinking still survives to this day. As a result, the female population of Pakistan suffers from deprivation of rights, injustice, false accusations, and overall enjoy less sovereignty than their male counterparts. This serves as a reason for sorrow and there is lack of self-disclosure.
21. Females also avoid seeking help from mental health services due to the risk of negatively affecting their future marriage prospects (Durrant, 2000).
22. Our findings point out that academic pressure doesn't necessarily cause someone to commit suicide. The academic institutions exert immense pressure on youngsters to perform well in school and failure to do so may result in deliberate self-harm (DSH) but it doesn't cause someone to intentionally end their life. Someone causing DSH may indicate that the person needs help, care, and much needed attention to their problems. It rarely means that the person is willing to end their life.
23. The study done by Sidhartha and Jena (2006) indicate a direct relation of tendency to avoid school, and lack of interest, motivation, and enthusiasm with committing suicide however our findings dictate that the only thing that is most prominently related with committing suicide are the interpersonal interactions between two or more personnel and theoretical problems that makes someone especially susceptible to suicide.
24. Another factor that plays a vital role in the lack of record of suicidal behavior in Pakistan is that the people that undergo DSH don't go to the applicable headquarters to receive any mental health service, yet they prefer to go to personal hospital to

seek medication, which neither diagnose the reason/ cause, nor report it to the respective organization (Police and other special services).

25. All of this is driven by the fear of humiliation, risk of detrimentally affecting their social stature and a possible chance of harassment by the police itself. This results in the minimization of reporting and documentation of such cases (Khan MM).

26. Information on suicide in Pakistan emerge solely from police, non-governmental organizations (NGOs), newspapers, and organizations working on the protection of human rights (Khan MM, Prince M).

2. METHODOLOGY AND ANALYSIS

We have taken 100 observations (cases) of a suicide. The variable of study includes ‘Gender’ ‘City Province’ ‘Method of suicide’ ‘Reason of suicide’.

The descriptive analysis of gender shows maximum suiciders are belongs to male population.

The mean (average) of the reason of suicide is 14.35 which indicates failed to relationship. The standard deviation is 10.28 and variance is 105.68 and skewness is 0.34

(Positively Skewed) and kurtosis is -0.70 and the range is 37.

The mean (average) of the method of suicide is 5.12 which is wrist cutting and Median is 4. Mode (maximum value) Is 4 jumped in front of train. The standard deviation is 3.09 and variance is 9.55 and skewness is 0.68 (positively Skewed) and kurtosis is 0.14 and the range is 13.

The mean (average) of the city/province is Tando/Sindh and Faisalabad/Punjab i.e. 7.88 and Median is 5. Mode (maximum value) is 5 that is Lahore. The standard deviation is 7.40 and variance is 54.88 and skewness is 1.14 (positively skewed) and kurtosis is 0.14 and the range is 26.

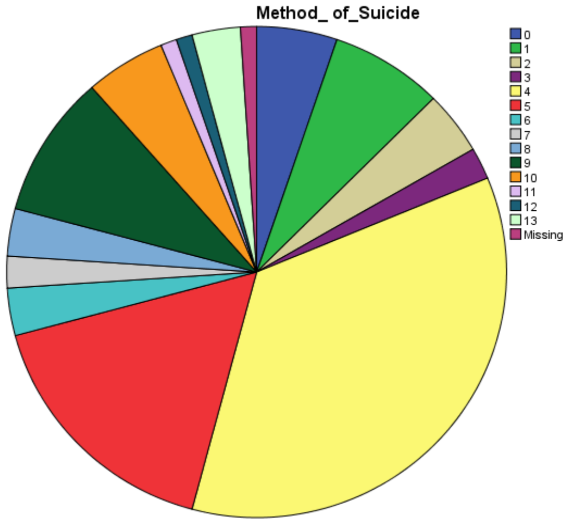


Fig 1: Pie Chart of Methods of Suicide

The method of suicide. Firearm (1). Poisoning (2). hanging (4). Shot him/herself (4). She jumped In front of a train (5). Wrist cutting (6). The black stone (7). Jumped from the roof (8). Taking pills (9). Jumped in the water and due to drowning (10). Jumped into well (11). Drank pesticide (12). It is clear from the above figure maximum suicides occur in front of train and then wrist cutting and least are hanging.

3. COMMENTS AND CONCLUSION

In this study, the statistical data on the suicide cases in Pakistan is formulated. The statistical data is collected and compiled, preferably on a regional basis. The estimated results are used to infer the main causes of suicide attempts in different regions of Pakistan. The data is used to compare the suicide cases and the circumstances which lead to such societal disorders, to better understand the underlying causes, and to propose an in-depth judgment on the Situation. the data we have collected through this data we can see the suicide rate is getting higher day by day and we see it is going to have a very bad situation in future as our mostly youth is indulging in this act if we don't control and we don't guide them we are going to see many more suicides in future. Young people are more likely to disclose their suicidal thoughts to a peer rather than an adult. Youth Suicide Prevention Education and Gatekeeper Training are Mental Health Literacy projects aimed at skill development, capacity building and youth engagement. Gatekeepers are trained to assess and respond appropriately to youth who are contemplating suicide and also have a vital role in reducing stigma associated with suicide and mental illness through education.

REFERENCES

1. Chaleby, K.S. (1996). Issues in forensic psychiatry in Islamic jurisprudence. *Bulletin of the American Academy of Psychiatry & the Law*, 24(1), 117-124.
2. Raleigh, V.S., Bulusu, L. and Balarajan, R. (1990). Suicides among immigrants from the Indian subcontinent. *The British Journal of Psychiatry*, 156(1), 46-50.
3. Haynes, R.H. (1984). Suicide in Fiji: a preliminary study. *The British Journal of Psychiatry*, 145(4), 433-438.
4. Gangat, A.E., Naidoo, L.R. and Wessels, W.H. (1987). Suicide in South African Indians. *South African Medical Journal*, 71(3), 169-171.
5. Ahmed, S.H. and Zuberi, H. (1981). Changing pattern of suicide and parasuicide in Karachi. *J. Pak. Med. Assoc.*, 31(4), 76-78.
6. Headley, L.A. (1983). *Suicide in Asia and the near East*. Berkeley: University of California Press.
7. Ali, B., Saud, A., Mohammad, S.N., Lobo, M., Midhet, F., Ali, S.A. and Saud M. (1993). Psychiatric morbidity: prevalence, associated factors and significance. *Journal of Pakistan Medical Association*, 43(4), 69-70.
8. Bhutta, Z.A. (1998). Staring into the abyss: walking the nuclear tightrope in south Asia: Sanctions can only make things worse for the people of India and Pakistan. *BMJ*, 317:363.
9. Bukhari, S. (1998). *Burning incidents: Women getting little justice*. The Dawn.
10. Diekstra, R.F. (1993). The epidemiology of suicide and parasuicide. *Acta Psychiatrica Scandinavica*, 87(S371), 9-20.
11. Jamil, H. (1990). Acute poisoning: A review of 1900 cases. *J. Pak. Med. Assoc.*, 40(6), 131-3.
12. Khan, M.M., Islam, S. and Kundi, A.K. (1996). Parasuicide in Pakistan: experience at a university hospital. *Acta Psychiatrica Scandinavica*, 93(4), 264-267.
13. Khan, M.M. (1998). Suicide and attempted suicide in Pakistan. *Crisis*, 19(4), 172-176.
14. Khan, M.M. and Reza, H. (2000). The pattern of suicide in Pakistan. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 21(1), 31-35.
15. Reza, H., Choudhry, S.M. and Khan, M.M. (1993). Emergency psychiatric referrals to a university hospital in Pakistan. *Psychiatric Bulletin*, 17(8), 471-472.
16. Simpson, M.E. and Conklin, G.H. (1989). Socioeconomic development, suicide and religion: a test of Durkheim's theory of religion and suicide. *Social Forces*, 67(4), 945-964.
17. Durrant, V. (2000). *Adolescent girls and boys in Pakistan: Opportunities and constraints in the transition to adulthood*. Population Council, Islamabad.
18. Federal Bureau of Statistic (2007). *Pakistan Statistical Year Book 2007*. Federal Bureau of Statistic Government of Pakistan, Islamabad.
19. Foley, D.L., Goldston, D.B., Costello, E.J. and Angold, A. (2006). Proximal psychiatric risk factors for suicidality in youth: The Great Smoky Mountains Study. *Archives of General Psychiatry*, 63(9), 1017-1024.
20. Horesh, N. and Apter, A. (2006). Self-disclosure, depression, anxiety, and suicidal behavior in adolescent psychiatric inpatients. *Crisis*, 27(2), 66-71.
21. Javed, M.A., Kundi, M.Z. and Khan, P.A. (1992). Emotional and behavioral problems among school children in Pakistan. *J. Pak. Med. Assoc.*, 42(8), 181-183.

22. Khan, M.M. and Reza, H. (1998). Gender differences in nonfatal suicidal behavior in Pakistan: Significance of sociocultural factors. *Suicide and Life-Threatening Behavior*, 28(1), 62-68.
23. Lin, J.J. and Lu, T.H. (2006). Association between the accessibility to lethal methods and method-specific suicide rates: An ecological study in Taiwan. *Journal of Clinical Psychiatry*, 67(7), 1074-1079.
24. McKeown, R.E., Cuffe, S.P. and Schulz, R.M. (2006). US suicide rates by age group, 1970–2002: An examination of recent trends. *American Journal of Public Health*, 96(10), 1744-1751.
25. Mirza, I. and Jenkins, R. (2004). Risk factors, prevalence, and treatment of anxiety and depressive disorders in Pakistan: systematic review. *BMJ*, 328(7443), 794.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 237-246

COMPARATIVE STUDY OF ONLINE VS TRADITIONAL SHOPPING

Ammara Nawaz Cheema and **Huma Akram**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

There are two methods of shopping. One is online shopping and other is traditional or conventional shopping. The word online is related with a framework which for this situation is the web. One of the numerous investigations accessible online has finished up some significant focuses. This examination was directed in a specific setting and time span in association with comprehend the truth of internet shopping experience. The variables that decides the purchaser's mentality towards internet shopping incorporates, sexual orientation, age, pay, and guidelines. For the purpose, the basic statistical analysis has been carried out to check the effectiveness of said dataset. The estimated results are compared for online vs traditional shopping.

1. INTRODUCTION

There are two methods of shopping i.e. online and traditional shopping. Online shopping is the drive of purchasing items or products over the web. It entails going web based. The whole process of buying items via website is comprised of arriving on a seller's site, choosing desired item, and finally arranging its delivery.

Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the internet using a web browser. Consumers find a product of interest by visiting the website of the retailer directly or by searching among alternative vendors using a shopping search engine, which displays the same product's availability and pricing at different e-retailers.

On contrary, traditional shopping is a term used to refer to the method of doing shopping where you visit stores and purchase products. In this way of shopping one needs to risk out from the comfort of home. To reach mall or store one must walk or take a ride and then purchase needed thing.

Traditional shopping may be a method of shopping for a product by getting to a store. Traditional shopping is also a kind of entertainment especially with friends and family. After long days of hard work they group up and go out shopping into malls purchasing things of their choice after enjoying window shopping; however, online shopping is deprived of this fun time. Moreover, after shopping they even go to the food courts and enjoy a meal together.

Ratings and reviews or feedbacks provided by customers help the other purchasers enormously. By taking these feedbacks into account the purchasers can decide whether they wanted to get that item or not (13). It is to be needed to state that refinement in the selling pace helps in reducing the cost of the items and ultimately increases the efficiency level. The factors that determines the buyer's attitude towards online shopping includes, gender, age, income, and instructions (15). The fundamental factors that have significant effect are happiness, value, and ease (1). The shopping experience is not solely dependent on environment's impact the personal experience is itself a change (2). Keeping purchased products and then further selling items from a well-known brand can also improve the trust (11). Online shopping frequency as well as experiences vary greatly among different regions of world. Information quality as well as buy quality are visibly huge in association of consumer loyalty in Malaysia (5). As well as brand image, value of items, people's liberality also greatly impacts buying of online items (16). The factors that hinders people from opting e-retailing. Primarily, it includes the lack of physical association. Secondarily, it is the concern over protection of individual personal data and security of banking details used while making money exchanges over the internet (4). The difficult solvers opt to shop online to obtain specifically desired items, wherein case shopping is usually observed as a work (10).The functional sensible as well as luxurious measurements or offers are foremost inspiration for customers to opt for online shopping (9). Description, goals, facts, data, worth, attributes if retailer as well as of consumer, information, conduct, and consumers preferences are main consideration of online shopping (8). Client's satisfaction depends on the item's durability as well as the customer services provided by the suppliers or retailers (14). Reveals that the most web based shopping is done by persons, which own a turbulent schedule and are not ready to waste their power and time in searching of a product on markets (6). Another strategy utilized by retailers is to advance their items online. For this a driving force is given by the suppliers to the shoppers showing some interest in their products for inspiring them to purchase these items (7). Customer's living the experience depends on the surrounding conditions, the attributes of purchased products and services that are related to the conditions (3). Risks associated with e-retailing equally affects all. It does not care whether the client went through hours looking for an item over internet. There is no work relationship between the buyer and seller (12). The purchasing and utilization experience plays an important part in shaping buyer's trust, corresponding behavior, and loyalty on the web and even in informal gatherings. Building online trust is a psychological approach that takes time and developed as a result of repeated stopovers to a website (17, 18). The clients only revisits sites if they feel that their desires or requirements are being fulfilled and that particular brand suits them The elements or factors that hold clients stick to online shopping in function of cost (19). The mentality of buyers is changing with the time. Considering numerous factors including easy payment methods, customized products, site personalization, home delivery and more buyers are considering e-shopping more convenient (20).

2. METHODOLOGY AND ANALYSIS

In this section we apply some basic analysis on our questionnaire to conclude findings.

Table 1
Percentage Distribution of Variables of study by their Options

Variable Name	Strongest Agree	Agree	Disagree	Strongly Disagree
Safe during COVID19	23.3%	46.9%	16.3%	13.9%
Saves time	20.2%	59.1%	13.8%	6.9%
Shop at any time	26.2%	55.9%	10.9%	6.9%
Prefer traditional shopping	18.3%	41.1%	31.7%	8.9%
Is online shopping risky	24.9%	47.3%	20.9%	7%
supersede traditional shopping	8.9%	50.2%	31%	9.9%
Delivery of products	15.8%	51.5%	26.2%	6.4%
Availability	17.8%	53%	22.3%	6.9%
Websites accurate	7%	36.8%	43.8%	12.4%
Sufficient information	8.9%	45.5%	33.2%	12.4%
Secure	10.4%	28.2%	42.6%	18.85
Reduces monetary cost	10%	49.8%	32.2%	8%
credit card creates difficulty	18.3%	48.5%	24.8%	8.4%
Prefer cash on delivery	30.7%	49.5%	13.4%	6.4%
Provision of home delivery	19.3%	59.9%	10.9%	9.9%
infrastructure underdeveloped	24.3%	51%	16.8%	7.9%
Hesitation	22.8%	55.4%	14.9%	6.9%
More convenient	13.9%	59.7%	19.4%	7%
Out your personal information	14.9%	55%	20.3%	9.9%
No Boundaries between countries	27.9%	47.3%	12.9%	11.9%
Time consuming	17.3%	53.3%	22.3%	6.9%

In the above table the options of questions are define as Strongly disagree (numeric form 1), Disagree (numeric form 2), Agree (numeric form 3) and Strongly agree (numeric form 4) and also variables name which are used in questionnaire such as safe during COVID19, saves time, shop at any time, prefer traditional shopping etc.

Table 2
Descriptive Analysis

Variable	Mean	Median	Mode	Range	Variance	Std. Deviation	Skewness	Kurtosis
Safe during COVID-19	2.78	3	3	3	0.90	0.95	-0.52	-0.57
Saves time	2.94	3	3	3	0.59	0.76	-0.75	0.71
Shop at any time	3.00	3	3	3	0.65	0.80	-0.81	0.58
Prefer traditional shopping	2.67	3	3	3	0.74	0.86	-0.16	-0.60
Is online shopping risky	2.88	3	3	3	0.72	0.85	-0.43	-0.36
supersede traditional shopping	2.58	3	3	3	0.62	0.79	-0.33	-0.29
Delivery of products	2.76	3	3	3	0.63	0.79	-0.32	-0.22
Availability	2.82	3	3	3	0.65	0.80	-0.46	-0.09
Websites accurate	2.39	2	2	3	0.62	0.79	0.03	-0.44
Sufficient information	2.51	3	3	3	0.68	0.82	-0.22	-0.50
Secure	2.31	2	2	3	0.80	0.89	0.23	-0.67
Reduces monetary cost	2.62	3	3	3	0.59	0.77	-0.26	-0.24
credit card creates difficulty	2.77	3	3	3	0.70	0.83	-0.36	-0.34
Prefer cash on delivery	3.03	3	3	3	0.69	0.83	-0.73	0.17
Provision of home delivery	2.88	3	3	3	0.68	0.82	-0.84	0.49
infrastructure underdeveloped	2.90	3	3	3	0.72	0.84	-0.60	-0.06
Hesitation	2.93	3	3	3	0.64	0.80	-0.69	0.35
More convenient	2.81	3	3	3	0.57	0.76	-0.62	0.39
Out your personal information	2.73	3	3	3	0.68	0.82	-0.52	-0.12
No Boundaries between countries	2.89	3	3	3	0.88	0.93	-0.68	-0.30
Time consuming	2.81	3	3	3	0.64	0.80	-0.46	-0.06

(Safe during COVID19) on average response is 2.78. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.90 and standard deviation is 0.95. Skewness shows negative and kurtosis platykurtic for this variable.

(Saves time) on average response is 2.94. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.59 and standard deviation is 0.76. Skewness shows negative and kurtosis leptokurtic for this variable.

(Shop at any time) on average response is 3.00. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.65 and standard deviation is 0.80. Skewness shows negative and kurtosis leptokurtic for this variable.

(Prefer traditional shopping) on average response is 2.62. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.74 and standard deviation is 0.86. Skewness shows negative and kurtosis platykurtic for this variable.

(Is online shopping risky) on average response is 2.88. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.72 and standard deviation is 0.85. Skewness shows negative and kurtosis platykurtic for this variable.

(Supersede traditional shopping) on average response is 2.58. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.62 and standard deviation is 0.79. Skewness shows negative and kurtosis platykurtic for this variable.

(Delivery of products) on average response is 2.76. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.63 and standard deviation is 0.79. Skewness shows negative and kurtosis platykurtic for this variable.

(Availability) on average response is 2.82. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.65 and standard deviation is 0.80. Skewness shows negative and kurtosis platykurtic for this variable.

(Websites accurate) on average response is 2.39. Median and mode tells us that mostly people responses are disagree with this question. Maximum responses are agreed. Variance is 0.62 and standard deviation is 0.79. Skewness shows positive and kurtosis platykurtic for this variable.

(Sufficient information) on average response is 2.51. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.68 and standard deviation is 0.82. Skewness shows negative and kurtosis platykurtic for this variable.

(Secure) on average response is 2.31. Median and mode tells us that mostly people responses are disagree with this question. Maximum responses are agreed. Variance is 0.80 and standard deviation is 0.89. Skewness shows positive and kurtosis platykurtic for this variable.

(Reduces monetary cost) on average response is 2.62. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.59 and standard deviation is 0.77. Skewness shows negative and kurtosis platykurtic for this variable.

(Credit card creates difficulty) on average response is 2.77. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.70 and standard deviation is 0.83. Skewness shows negative and kurtosis platykurtic for this variable.

(Prefer cash on delivery) on average response is 3.03. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.69 and standard deviation is 0.83. Skewness shows negative and kurtosis leptokurtic for this variable.

(Provision of home delivery) on average response is 2.88. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.68 and standard deviation is 0.82. Skewness shows negative and kurtosis leptokurtic for this variable.

(Infrastructure underdeveloped) on average response is 2.90. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.72 and standard deviation is 0.84. Skewness shows negative and kurtosis leptokurtic for this variable.

(Hesitation) on average response is 2.93. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.64 and standard deviation is 0.80. Skewness shows negative and kurtosis platykurtic for this variable.

(More convenient) on average response is 2.81. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.57 and standard deviation is 0.76. Skewness shows negative and kurtosis platykurtic for this variable.

(Out your personal information) on average response is 2.73. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.68 and standard deviation is 0.82. Skewness shows negative and kurtosis leptokurtic for this variable.

(No Boundaries between countries) on average response is 2.89. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.88 and standard deviation is 0.93. Skewness shows negative and kurtosis leptokurtic for this variable.

(Time consuming) on average response is 2.81. Median and mode tells us that mostly people responses are agree with this question. Maximum responses are agreed. Variance is 0.64 and standard deviation is 0.80. Skewness shows negative and kurtosis leptokurtic for this variable.

Table 3
Descriptive Analysis by Gender and Age

Variable	Mean	Median	Mode	Range	Variance	Std. Deviation	Skewness	Kurtosis
Gender	1.65	2	2	1	0.23	0.47	-0.61	-1.63
Age	24.31	23	22	39	35.86	5.98	3.44	14.22

(Gender) on average response is 1.65. Median and mode tells us that mostly people responses are female with this question. Maximum responses are male. Variance is 0.23 and standard deviation is 0.47. Skewness shows negative and kurtosis leptokurtic for this variable.

(Age) on average response is 24.31. Median tells us that mostly people responses are of 23 and mode tells 22 for this question. Maximum responses are 39. Variance is 0.23 and standard deviation is 0.47. Skewness shows positive and kurtosis platykurtic for this variable.

Table 4
Percent Contribution by Gender

Gender	Frequency	Percent	Cumulative Percent
Male	71	35.3	35.3
Female	130	64.7	100.0

In our respondent 35% are male and 65% are female and maximum respondents fall in the age group 22-35 years.

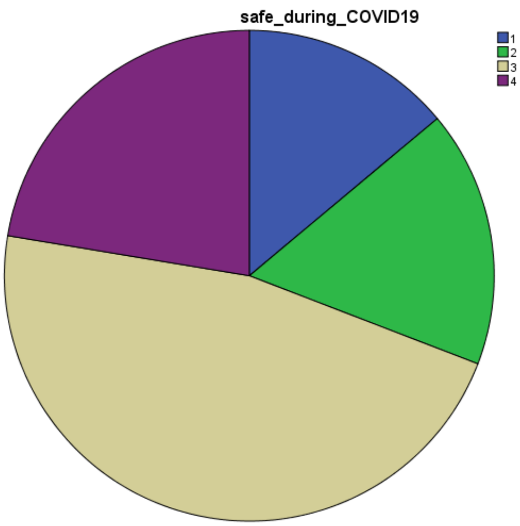


Fig 1: Pie Chart of Online Shopping Safe during COVID-19

Variable of this Fig 1 is Safe during COVID19. According to this graph 23.3% are strongly agree and 46.5% people are agree moreover 13.9 are strongly disagree and 16.3% are disagree with question. The question was is online shopping safe during COVID19?

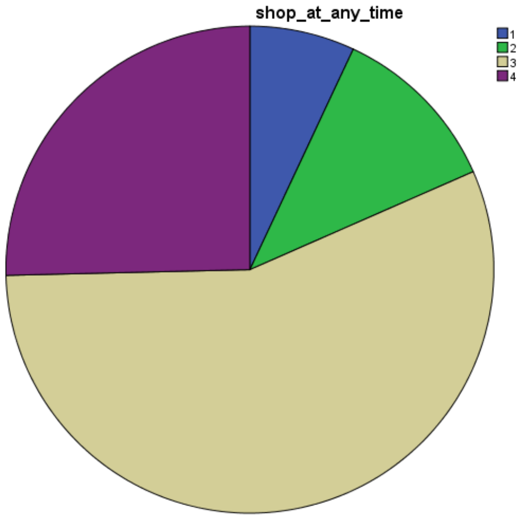


Fig 2: Pie Chart of Shop at any time

Variable of this Fig 2 is Shop at any time. According to this graph 26.2% are strongly agree and 55.9% people are agree moreover 26.2% are strongly disagree and 6.9% are disagree with question. The question was “It is a great advantage to be able to shop at any time of the day”?

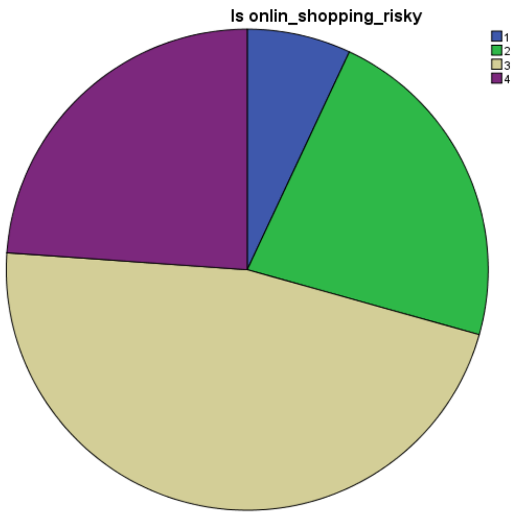


Fig 3: Pie Chart of Online Shopping Risky

This graph is related to online shopping is risky. According to this graph 24.9% are strongly agree and 43.7% people are agree moreover 7% are strongly disagree and 20.9% are disagree with question. The question was “Is online shopping is risky”?

4. COMMENTS AND CONCLUSION

According the survey/ responses which are taken from different age group peoples. The estimated results show that online shopping saves time and safe during COVID-19 over the traditional shopping. Moreover, traditional shopping is much better than online but due to COVID-19 many countries prefer online shopping

REFERENCES

1. Burke, R.R. (2002). Technology and the customer interface: what consumers want in the physical and virtual store? *Journal of the academy of Marketing Science*, 30(4), 411-432.
2. Heilbrunn, B. (2010). *La consommation et ses sociologies: Domaines et approches*. Armand Colin.
3. Edgell, S., Hetherington, K. and Warde, A. (1997). *Consumption matters: the Production and Experience of Consumption*. Oxford (United Kingdom), Blackwell.
4. Huseynov, F. and Yıldırım, S.Ö. (2014). Internet users’ attitudes toward business-to-consumer online shopping: A survey. *Information Development*, 32(3), 452-465.
5. Vegiayan, K.D., Ming, C.W. and Harun, M.L.O. (2013). Online shopping and customer satisfaction in Malaysia. *International Journal of Marketing Practices*, 1(1), 43-51.
6. Hooda, S. and Singh, N. (2018). Online apparel shopping. *Man-Made Textiles in India*, 46(3), 79-87.
7. Lim, S. and Lee, B. (2015). Loyalty programs and dynamic consumer preference in online markets. *Decision Support Systems*, 78, 104-112.
8. Puccinelli, N.M., Goodstein, R.C., Grewal, D., Price, R., Raghubir, P. and Stewart, D. (2009). Customer experience management in retailing: understanding the buying process. *Journal of Retailing*, 85(1), 15-30.
9. Hirschman, E.C. and Holbrook, M.B. (1982). Hedonic Consumption; emerging concepts, methods and proportions, *Journal of Marketing*, 48(3), 92-101.
10. Babin, B.J., Darden, W.R and Griffin, M. (1994). Work and /or fun; measuring hedonic and utilitarian shopping value. *Journal of Consumer Research*, 20, 644-656.
11. Lim, H. and Dubinsky, A.J. (2004). Consumers' perceptions of e-shopping characteristics: an expectancy-value approach. *Journal of Services Marketing*, 18(7), 500-513.
12. Zuroni, M.J. and Goh, H.L. (2012). Factors influencing consumers’ attitude towards e-commerce purchases through online shopping. *International Journal of Humanities and Social Science*, 2(4), 223-230.
13. Chin, S.L. and Goh, Y.N. (2017). Consumer Purchase Intention toward Online Grocery Shopping: View from Malaysia. *Global Business & Management Research*, 9, 221-238.

14. Hennig-Thurau, T. and Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology & Marketing*, 14(8), 737-764.
15. Vijay, S. T., Prashar, S. and Sahay, V. (2019). The influence of online shopping values and web atmospheric cues on e-loyalty: Mediating role of e-satisfaction. *Journal of Theoretical and Applied Electronic Commerce Research*, 14(1), 1-15.
16. Haque, A., Anwar, N., Yasmin, F., Sarwar, A., Ibrahim, Z. and Momen, A. (2015). Purchase intention of foreign products: A study on Bangladeshi consumer perspective. *SAGE Open*, 5(2). doi:10.1177/2158244015592680.
17. Urban, G.L., Amyxb, C. and Lorenzonc, A. (2009). Online Trust: State of the Art, New Frontiers, and Research Potential. *Journal of Interactive Marketing*, 23, 179-190.
18. Holt, D.B. (1995). How consumers consume: a typology of consumption practices. *Journal of Consumer Research*, 22, 1-16.
19. Reibstein, D.J. (2002). What attracts customers to online stores, and what keeps them coming back? *Journal of the Academy of Marketing Science*, 30(4), 465-473.
20. Bhatt, A. (2014). Consumer Attitude towards Online Shopping in Selected Regions of Gujarat. *Journal of Marketing Management*, 2(2), 29-56.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 247-251

COVID-19 EFFECTS-ANALYSIS

Ammara Nawaz Cheema and **Khurram Ashfaq**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

There are two methods of shopping. One is online shopping and other is traditional or conventional shopping. The word online is related with a framework which for this situation is the web. One of the numerous investigations accessible online has finished up some significant focuses. This examination was directed in a specific setting and time span in association with comprehend the truth of internet shopping experience. The variables that decides the purchaser's mentality towards internet shopping incorporates, sexual orientation, age, pay, and guidelines. For the purpose, the basic statistical analysis has been carried out to check the effectiveness of said dataset. The estimated results are compared for online vs traditional shopping.

1. INTRODUCTION

Human history is observing a very unfamiliar time fighting an unseen foe. COVID ailment 2019 [COVID-19] is an infectious sickness brought about by serious intense respiratory disorder COVID 2 (SARS-CoV-2)."CO" represents corona, "VI" represents infection, "D" represents malady. In December 2019, the initial case of COVID-19 was identified in Wuhan, China. On January 30,2020 the outburst was proclaim as the Public Health Emergency of International Concern by WHO. Current proof suggests that COVID-19 spreads between people via direct, oblique (through contaminated objects or surfaces), or shut contact with contaminated human beings through mouth and nostril secretions. These consist of saliva, respiratory secretions or secretion droplets. These are launched from the mouth or nostril when a contaminated individual speaks, coughs or sneezes, for example. People who are in shut contact (within 1 meter) with a contaminated person can entice COVID-19 when these infectious droplets get into their mouth, nostril or eyes. It is essential to limit your outside activities, wear a mask, clean hands time and again, maintain social distancing and cover the mouth with elbow or with a tissue paper when we feel tickle in the throat and wheezing, to avoid contact with these droplets. To protect others, wearing a medical mask and cleaning hands frequently is very demanding.

The current shear of the COVID-19, which provokes the ongoing prevalent, is known as corona virus. The Chinese foundation recognized this shear of the infection on Jan 7, 2020 [1]. It on track to divide globally from country to country [2] and the gear were intensifying day by day as of June 2020, near are no booming vaccines [3-4] and no drug medicine is explicitly recommended [5] However, recovery plasma transfusion know how to be an expected treatment, however it is in trial stages. Precautionary steps are particularly super to stop its fast unfurl then again for these to be useful, powerful strategies should be taken for reasonable well-being preparing of individuals. This outbreak has seriously influenced some overall spots and their medical care frameworks have arrived at the component of weariness. In the midst of the confusion and the rising human cost, medical care administrations are existing way decentralization and discontinuity in numerous seriously influenced worldwide territories [6]. Weak people group are lopsidedly affected in this cataclysmic circumstance. This pandemic is tireless. In an emergency stricken world grasped by method of difficulties, it has revealed the weaknesses of the worldwide entrepreneur framework, pushed by the postponed reaction [7]. A torrential slide of cases has overburdened medical services structures in created nations. In creating overall territories with long not noted and underfunded open wellness parts, the rampant is principle to pandemonium [8]. The function of the Huanan fish discount marketplace in proliferating ailment is muddled. Several starting coronavirus instances have been linked to this marketplace recommending that SARS-CoV-2 was sent from animals to individual [9]. In any case, a gene based research has given confirmation that the contamination was once conveyed from an additional, still doubtful sector, through the marketplace the spot it unfurls all rapidly, although person-to-person spread could possibly have occurred beforehand [10], allot of polluted family unit individuals and logical laborers have tried the presence of individual to-individual spread [11].

Individual to-individual passing on is thought to happen between intimate links all in all by method of respiratory beads created when a contaminated person hacks or sniffles. Things like clothes, furniture or soap can furthermore be a huge wellspring of passing on, as COVID has been resolved to persevere on things like furniture till ninety-six hours [12] and diverse COVID for as long as one week plus 2 days [13]. whether or not or now no longer there is asymptomatic passing on of sickness is disputable. 1 primer examination about posted on January 30 detailed asymptomatic transmission [14], on the other hand later it was when once situated that the analysts had no longer immediately met the patient, who did in certainty have signs preceding communicating issue [15].

However training from preceding pandemic conditions point out that there is probable to be a hazard of COVID-19 victims who get higher grow signs and symptoms of pts or depression [16] it is per chance to be common through private aspects such as flexibility accessible and terrific of collective useful resource and the sufferers personal stresses identified with sickness and recuperation [17]

While medical services laborers announced improved mental trouble for the length of the SARS rife, comprehension of danger and impression of individual wellness intercede the advancement of mental morbidity [18].

To comprise catching illnesses like COVID-19, powers that be and management officers alike advocate a sequences of deterrent way of acting, for example offer purity, and escaping way of acting, for example communal estrangement or (optional) lock up (for example, [19-24]). Before simulations and obtainable day critiques verify that a grouping of completely strategies has the high-quality hit charges in containing the illness [25-26].

2. METHODOLOGY AND ANALYSIS

To elaborate the significance of the study a real data set of infected, recovered and confirmed cases of virus has been taken from the region of Pakistan. Over the passage of time there seems a clear increase in confirmed cases of virus as shown in Figure 1-2.

Moreover, it is clear from the figure 2 that no case is reported under the age of 5 years as researches already shown that strong vaccinations of kids save them to infect with this virus, whereas; the maximum cases are reported for the age group 41 and above.

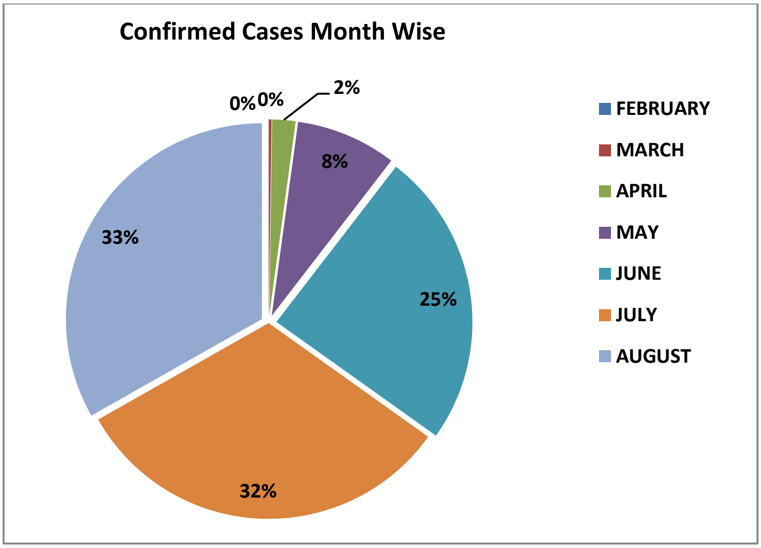


Figure 1: Confirmed Cases of Virus Month wise

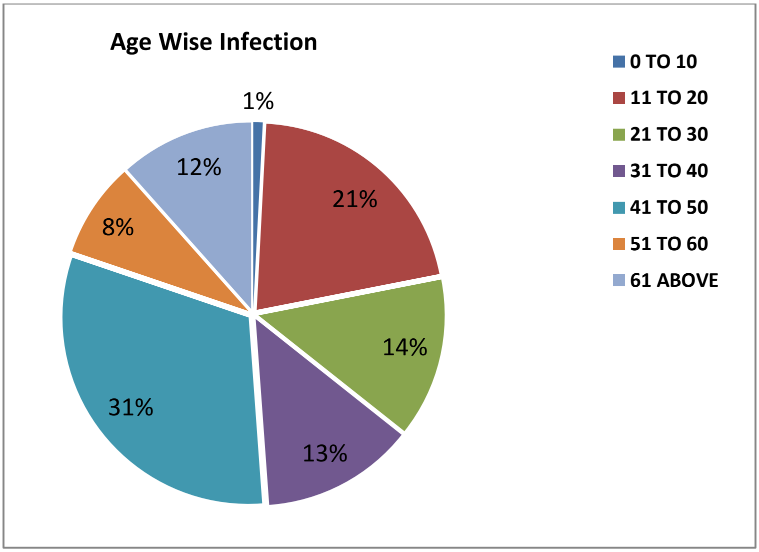


Figure 2: Age Wise Infection

Table 1
Ratio Statistics for Confirmed Cases v.s Tests, Recoveries and Deaths

Variables	Related Differential	Coefficient of Dispersion	Coefficient of Variations
Tests	0.920	0.189	21.7%
Recoveries	0.588	0.789	102.4%
Deaths	0.824	0.203	38.6%

In ratio statistics, the coefficient of dispersion (COD) is an indicator of variability. The CODs are 0.189 and 0.203, respectively, indicating the least variability in the test rate, reported cases, and deaths. The COD 0.789 suggests that the recovery and reported cases have more variability.

3. COMMENTS AND CONCLUSION

In this study, the COVID-19 pattern is explored. The statistical analysis about COVID-19 under the real data study is conducted. The findings conclude that under the age of 5 years virus do not react due to the stronger other vaccinations of kids. Whereas, the age group 60+ is more crucial group of infected in which the chances of recoveries are also decline as compare to age group 25-30. The ratio statistics indicates small variability in the deaths with the confirmation of infections.

REFERENCES

1. Burke, R.R. (2002). Technology and the customer interface: what consumers want in the physical and virtual store, *Journal of the Academy of Marketing Science*, 30(4), 411-432.
2. Heilbrunn, B. (2010). *La consommation et ses sociologies: Domaines et approches*. Armand Colin.
3. Edgell, S., Hetherington, K. and Warde, A. (1997). *Consumption matters: the Production and Experience of Consumption*. Oxford (United Kingdom), Blackwell.
4. Huseynov, F. and Yıldırım, S.Ö. (2014). Internet users' attitudes toward business-to-consumer online shopping: A survey. *Information Development*, 32(3), 452-465.
5. Vegiayan, K., Ming, C. and Harun, M.L. (2013). Online shopping and customer satisfaction in Malaysia. *International Journal of Marketing Practices*, 1(1), 43-51.
6. Hooda, S. and Singh, N. (2018). Online apparel shopping. *Man-Made Textiles in India*, 46(3), 79-87.
7. Lim, S. and Lee, B. (2015). Loyalty programs and dynamic consumer preference in online markets. *Decision Support Systems*, 78, 104-112.
8. Puccinelli, N.M., Goodstein, R.C., Grewal, D., Price, R., Raghubir, P. and Stewart, D. (2009). Customer experience management in retailing: understanding the buying process. *Journal of retailing*, 85(1), 15-30.
9. Hirschman, E.C. and Holbrook, M.B. (1982). Hedonic consumption: emerging concepts, methods and propositions. *Journal of Marketing*, 46(3), 92-101.

10. Babin, B.J., Darden, W.R. and Griffin, M. (1994). Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of Consumer Research*, 20(4), 644-656.
11. Lim, H. and Dubinsky, A.J. (2004). Consumers' perceptions of e-shopping characteristics: an expectancy-value approach. *Journal of Services Marketing*, 18(6), 500-513.
12. Jusoh, Z.M. and Ling, G.H. (2012). Factors influencing consumers' attitude towards e-commerce purchases through online shopping. *International Journal of Humanities and Social Science*, 2(4), 223-230.
13. Chin, S.L. and Goh, Y.N. (2017). Consumer Purchase Intention Toward Online Grocery Shopping: View from Malaysia. *Global Business & Management Research*, 9(4s), 221-238.
14. Hennig-Thurau, T. and Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology & marketing*, 14(8), 737-764.
15. Vijay, S.T., Prashar, S. and Sahay, V. (2019). The influence of online shopping values and web atmospheric cues on e-loyalty: Mediating role of e-satisfaction. *Journal of Theoretical and Applied Electronic Commerce Research*, 14(1), 1-15.
16. Haque, A., Anwar, N., Yasmin, F., Sarwar, A., Ibrahim, Z. and Momen, A. (2015). Purchase intention of foreign products: A study on Bangladeshi consumer perspective. *Sage Open*, 5(2), 2158244015592680.
17. Urban, G.L., Amyx, C. and Lorenzon, A. (2009). Online trust: state of the art, new frontiers, and research potential. *Journal of Interactive Marketing*, 23(2), 179-190.
18. Holt, D.B. (1995). How consumers consume: A typology of consumption practices. *Journal of consumer research*, 22(1), 1-16.
19. Reibstein, D.J. (2002). What attracts customers to online stores, and what keeps them coming back? *Journal of the academy of Marketing Science*, 30(4), 465-473.
20. Bhatt, A. (2014). Consumer attitude towards online shopping in selected regions of Gujarat. Bhatt, A. (2014). Consumer Attitude Towards Online Shopping in Selected Regions of Gujarat. *Journal of Marketing Management*, 2(2), 29-56.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 253-260

RISK FACTORS ASSOCIATED WITH CANCER PATIENTS

Metehi-ur-Rehman and **Ammara Nawaz Cheema**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Cancer is one of the most dangerous diseases all over the world. Our purpose of the study of cancer is to determine risk factor of cancer which is very harmful for human life. We determine the some important risk factor of cancer which becomes the main cause of cancer. Main risk factor of cancer is age, medical history etc. We use statistical analysis to conclude results. This study is for the welfare of people to tell that how much risk factor can be cause of cancer.

1. INTRODUCTION

Cancer, basically cells are grow out and invade tissues. After that cells become cancer due to accumulation of the mutation in their DNA. Certain genetic defects and infections can be harmful and leading cause of cancer. Basically some main risk factors which are important cause of cancer includes; environmental factors like air pollution and the alcohol, smoking, and overweight leads toward cancer. Some chemicals which are used in food are included in the risk factor of cancer. Some risk factor of cancer which we cannot control like age and medical history of family. A family history is an important factor of certain cancers can be a sign of a possible inherited cancer. Cancer has type's breast cancer, brain cancer, skin cancer, lung cancer etc. Skin cancer is the increasing day by day and most common cancer in men and women is the lung cancer and skin cancer which are leading cause of death in the entire world.

Deviate and Rosenberg (2012) study the first case of cancer which was discovered by British surgeon Percival Pot in 1775. After discover of the disease physicians started working together and draw a conclusion about it. In the 18th century then, finally discovered that the "Cancer Poison" spread from the tumor through the lymph nodes. Between 1871 to 1874 English surgeon Campbell De Morgan identify first view of disease. The first method of treatment of cancer was surgery and the result of surgery was not satisfied [1].

According to the research report of *JH Rowland and Marriott (2007) that* the number of cancer cases has grown simultaneously from the past three decades. Approximately the 12 million people of US are the survivor of cancer. This almost 4% from US population and increase in the cancer survivor is almost 4 million people, increment of 1.4% .Cervical cancer which is cancer of uterine cervix is almost this year leading cancer among women at international level[2]. According to the report of the WHO of (2017) stated that 520,000 cases was reported and the confirmed death was 274,000. HPV and the cervical statistic in India was high in last year as compare to the other types of cancer. In US the death rate decrease (per year) in men, women and children from 2006 to 2015 was respectively 1.8%, 1.4% and 1.4%."[3].

Mahmood et al. (2017) shows that “among all the provinces KPK was leading province in cancer cases which has 25.2%. Main cancer in Pakistan among the women was the breast cancer which has ratio 19.9% and then colon cancer which have almost the 7.3. Skin cancer almost have 4.5 % ratio in registered cancer cases. Cancer have different stages highest ratio in the stages the stages II with the ratio of 25.8”[4].

US Rc in (2015) stated that “Risk factor are associated with cancer patient are alcohol, tobacco, lack of activity etc. In the previous some researches showed that some risk factor are become the main cause of cancer .in US research the smoking and drinking is leading cause of error in men [5]. WHO in (2018) shows the World cancer statistics for the most common cancers (excluding non-melanoma skin cancer) in 2018, the latest year available. There were an estimated 18 million cancer cases around the world in 2018, of these 9.5 million cases were in men and 8.5 million in women”[6].

Melkonian SC and Jim MA in (2020) stated the American Indians and Alaska natives have high percentage of cancer patient involve in lung, colorectal, stomach, kidney, skin, breast and other cancer as compare to non – Hispanic white people in united states [7]. McCollum J and White MC in (2019), the cancer risk factor in six region of the American. The people may got in to trouble if they involve in commercial use of tobacco, obesity, lack of physical activity, diabetes and viral heptads infection [8]. Reuse of NCCI in (2019), the death rate of cancer is decreasing in men, women and children. The rate of new cancer cases is increasing in men, women and children [9].

NCCI report in (2018), the genetic behavior is individual like genetic mutations genetic modifier and polymorphism can change both men and women risk factor of cancer [10]. Muhammad Abbas Khokar, Muhammad Mohsen Ali and Samir Liana in (2017), there are some major problems are to reduce rate of cancer deaths and increase recovery rate in Punjab are lack of laboratories, lack of hospital, lack of research program, lack of awareness and poor drug access [11]. Virginia F. Borges in the (2019), the women in pregnancy increase the risk factor for breast cancer according to the researches the women at the age of greater than 45 years have high chances of risk factor [12]. Leslie L. Robison in (2019) the pulmonary complications due to chemotherapy and radiation exposure can be acute in onset or may develop insidiously, months or years after treatment. Our analyses show that pulmonary complication continue to manifest greater than 5 years after diagnosis and that treatment related factors are important identified of risk [13].

2. METHODOLOGY AND ANALYSIS

The descriptive analysis includes: Mean value is the most common value in the data set. Moreover, it determines the central tendency where the sum of the deviations of each value from the *mean* is always zero. Mean value is 96.05 .This value is the central or Average value from our data. The mean value is in the age factor is 9.36 of age 55-.This value describe that the maximum cases of cancer in this age group.

Median is the middle mark of the data set. If the number of data is odd we choose the central value from the data set. In case of the number of data is even then we take middle two numbers of data and the find the average of two numbers. In data set the median is 28.50 which describe the central value in the data.

Mode is the highest value in the data set. Usually the mode represent the in the bar graph. Mode is very rarely used in the continuous data. Mode not provide good central tendency about the data. In data the highest value is 2 which describe the maximum value of data.

Standard deviation tells us how data spread from mean value. Low standard deviation describe the data is close to mean value and the high standard deviation describe that data is more spread from the mean value. Standard deviation of data is 343.619 which describe that the data are spread out.

Variance is the average square deviation from the mean. Generally it give information how our data is spread from the mean value. The determine value of the data is 118073.785. This value describes how the values are spread from the mean value.

It is degree of distortion that describe the symmetric bell curve .the symmetric curve gives us the 0 skewness. There are two types of Skewness Positive and negative. Positive means the tail of the curve on the right of the normal distribution and negative skew means the tail is on the left side of distribution. Our data is positively have skewed.

Kurtoses define how the tail of distribution differs from the normal distribution. Three types of Kurtoses are Mesokurtic, leptokurtic and platykurtic. Mesokurtic mean distribution have medium value. Leptokutric means that distribution is highly peaked. Platykurtic means distribution is flat.

Table 1
Descriptive Analysis

Age	Mean	Median	Mode	Std.dev	Variance	Skewness	Kurtosis	Range
Gender	.51	1.00	1	0.502	0.2522	-.041	-2.040	1
1-	3.13	0.00	0	10.77	116.11	5.601	37.496	86
1-5	2.18	0.00	0	9.098	82.775	5.660	33.992	67
5-10	.86	0.00	0	4.168	17.374	6.902	51.217	35
10-15	1.89	0.00	0	5.942	35.311	5.484	35.720	47
15-20	4.68	1.00	0	15.06	227.04	7.042	57.840	135
20-25	5.93	1.00	0	20.70	428.61	7.822	68.955	193
25-30	8.98	1.00	0	35.31	1248.6	7.611	63.193	319
30-35	8.09	1.00	0	44.03	1938.6	9.664	95.369	439
35-40	10.93	2.00	0	47.07	2215.7	7.609	61.107	417
40-45	8.40	2.00	0	31.33	981.87	8.079	70.530	292
45-50	9.33	3.00	0	34.27	1174.4	8.123	71.041	320
50-55	9.36	2.00	0	33.55	1126.2	8.236	73.649	317
55-60	8.00	1.50	0	28.36	804.54	7.948	70.000	265
60-65	5.28	1.00	0	18.39	338.46	6.679	49.964	157
65-70	3.24	1.00	0	11.52	132.81	7.191	58.686	103
70-75	1.42	0.00	0	4.482	20.084	5.927	41.860	37
75-80	.75	0.00	0	2.728	7.4422	6.855	54.842	24
80-85	.25	0.00	0	0.809	0.6544	5.006	29.860	6

Table 2
Descriptive Analysis of Total Cases

Quantity	Total
Mean	96.05
Median	28.50
Mode	2
St. Deviation	343.619
Variance	118073.785
Skewness	8.193
Kurtosis	72.304

In view of regression analysis: Gender as dependent variable and the age factor is independent variable. Variables are defined in following Table 3 and results of regression analysis are reported in Table 4.

Table 3
Variables of Age Factor

Age Factor	Variables
1-	X_1
1-5	X_2
5-10	X_3
10-15	X_4
15-20	X_5
20-25	X_6
25-30	X_7
30-35	X_8
35-40	X_9
40-45	X_{10}
45-50	X_{11}
50-55	X_{12}
55-60	X_{13}
60-65	X_{14}
65-70	X_{15}
70-75	X_{16}
75-80	X_{17}
80-85	X_{18}

The variables of age group are defined above. Corresponding values of Beta are below:

Table 4
Regression Analysis

Variable	Coefficients	P Value
X_1	-.087	.710
X_2	.044	.840
X_3	.069	.794
X_4	-.224	.650
X_5	.309	.666
X_6	-.014	.989
X_7	-.220	.878
X_8	-.741	.168
X_9	.685	.615
X_{10}	-1.561	.231
X_{11}	-1.992	.167
X_{12}	2.445	.180
X_{13}	2.319	.089
X_{14}	-1.603	.106
X_{15}	1.023	.307
X_{16}	-1.512	.024
X_{17}	1.034	.066
X_{18}	-.078	.756

The regression equation which determined is here:

$$
\begin{aligned}
Y = -.87X_1 + 0.44X_2 + 0.69X_3 - .22X_4 + .309X_5 - .014X_6 - .220X7X_7 \\
- .741X_8 + .685X_9 - 1.561X_{10} - 1.99X_{11} + 2.45X_{12} \\
+ 2.31X_{13} - 1.603X_{14} + 1.02X_{15} - 1.512X_{16} + 1.03X_{17} \\
- .078X_{18}
\end{aligned}
$$

The values of p describe the significants. In the age factor the P has significant values if $p > 0.05$ then p has no significant values if $p < 0.05$ means significant. In data all values of p less than 0.05 are significant .The most cases lie in X13.

Table 5
Durbin Watson Test Table

R Square	A.R. Square	Error of Estimation	Durbin Watson
.166	.020	.507	.368

The R square value defines how close data are fitted on regression. In generally this is coefficient of determination. This is use for determination of data.

Durbin Watson is a simple test which is used for existence of correlation in data. Durbin Watson has value between the 0 to 4. If value of Durbin Watson is 2 then it have no correlation detected. In data the Durbin Watson value is .368 which means data have correlation.

Table 6
95% of Confidence Interval

Variables	Lower bound	Upper bound
X_1	.397	.649
X_2	-.026	.017
X_3	-.021	.026
X_4	-.055	.071
X_5	-.101	.064
X_6	-.037	.058
X_7	-.046	.046
X_8	-.044	.037
X_9	-.021	.004
X_{10}	-.022	.036
X_{11}	-.066	.016
X_{12}	-.071	.012
X_{13}	-.017	.091
X_{14}	-.006	.089
X_{15}	-.097	.009
X_{16}	-.042	.131
X_{17}	-.316	-.023
X_{18}	-.013	.394

Confidence interval shows we are 95 % confident that our true values lies in between the lower bound and upper bound. The other 5% is margin of error.

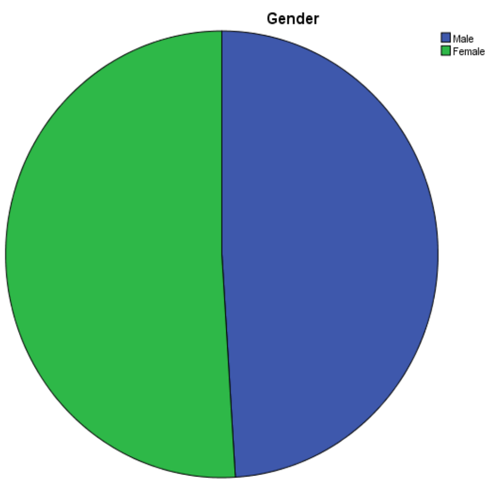


Figure 1: Gender vise Cancer

49% are male and 51% are female.

The basic regression analysis of Cancer risk factor data set was verified by considering gender as a responsible variable and other variables of risk factor is age group independent variables and results indicate basic regression model is appropriate towards practical significance for risk factor data set. This analysis was performed to observe relationship between gender and other multiple independent variables. The statistical summary of descriptive analysis gave results with certainty which was outside of measurement error. To separate the variables which have least effect on productivity of cancer influential observations analysis is performed. We note that the behavior of Risk factor in cancer is influenced by variables including gender and age factor. Rate of Risk factor can be decreased by implementing this. The main risk is in age group of X13.

3. COMMENTS AND CONCLUSION

The recommendation made in this study will helpful for doctors and suffering patients from cancer and also helpful for people to avoid from risk factor. The selection of variables that contributes more towards performance of cancer risk factor by applying best prediction model to overcome risk factor of cancer. This research work will also help to cancer patients and doctors for treatment.

REFERENCES

1. Pori, M.L. and Sen., P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons, Inc., New York.
2. Rae, C.R. (1965). *Linear Statistical Inference and Its Applications*. John Wiley and Sons, Inc., New York.
3. Sprott, D.A. and Kalbfleisch, J.D. (1969). Examples of likelihoods and comparison with point estimates and large sample approximations. *Journal of the American Statistical Association*, 64(326), 468-484.
4. Mahmood, S., Faraz, R., Yousaf, A., Quader, A., Asif, H., Atif, A., Nadee, L., Parveen, N., and Farhana Badar, F. (2017). *Annual Cancer Registry Report-2017, of the Shaukat Khanum Memorial Cancer Hospital & Research Center, Pakistan*. Online available at http://shaukatkhanum.org.pk/wp-content/uploads/2018/08/acrr-2017.pdf
5. Mahmood, S., Faraz, R., Yousaf, A., Quader, A., Asif, H., Atif, A., Nadeem, L., Parveen, N., Tanveer, R., Hussain, A., Zafar A., and Badar, F. (2019). Collective Cancer Registry Report from December 1994 to December 2019, of the Shaukat Khanum Memorial Cancer Hospital & Research Center, Pakistan. Online available at https://shaukatkhanum.org.pk/wp-content/uploads/2020/08/Collective-Cancer-Registry-Report-Dec.-1994-to-Dec.-2019.pdf
6. Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2), 185-198.
7. Melkonian, S.C., Jim, M.A., Haverkamp, D., Wiggins, C.L., McCollum, J., White, M.C., Kaur, J.S. and Espey, D.K. (2019). Disparities in cancer incidence and trends among American Indians and Alaska Natives in the United States, 2010–2015. *Cancer Epidemiology and Prevention Biomarkers*, 28(10), pp.1604-1611.
8. Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), 267-278.
9. Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481.

10. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*. 50(3), 163-170.
11. Oxvig, C. (2015). The role of PAPP-A in the IGF system: location, location, location. *Journal of cell communication and signaling*, 9(2), 177-187.
12. Ries, L.A.G. (Ed.). (1999). *Cancer incidence and survival among children and adolescents: United States SEER program, 1975-1995* (No. 99). National Cancer Institute.
13. Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6, 1-11.
14. Howard, S. (1972). Contribution to the discussion of a paper by DR Cox: Regression models and life-tables. *Journal of the Royal Society*, Series B, 34, 210-211.
15. Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 80, 89-99.
16. Breslow, N.E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 43(1), 45-57.
17. Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Rival Statistical Society*, Series B 34, 187-220.
18. Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.
19. Farewell, V.T. and Prentice, R.L. (1977). A study of distributional shape in life testing. *Technometrics*, 19(1), 69-75.
20. Godwin II, J.D. and Brown, C.C. (1975). Some prognostic factors in survival of patients with cancer of the colon and rectum. *Journal of chronic diseases*, 28(9), 441-454.
21. Hankey, B.F. and Myers, M.H. (1971). Evaluating differences in survival between two groups of patients. *Journal of chronic diseases*, 24(9), 523-531.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 261-269

TRENDS OF CYBER CRIMES

Ammara Nawaz Cheema and **Rimsha Aamir**
Department of Mathematics, Air University, Islamabad, Pakistan
Email: ammara.au@gmail.com; ammara.cheema@mail.au.edu.pk

ABSTRACT

Cyber-crime is any criminal activity done by a person using a computer or network. Cyber-crime is an illegal activity committed on the internet. The internet in Pakistan is growing rapidly. The first cyber-crime ever recorded was in France during 1820. There are laws related to cyber-crimes but barely 5% are aware of laws and concerned authorities. It has proved that Cybercrimes are mutely present; directly or indirectly. It is executed by organizations or the individuals which is the biggest threat of developed or under developing countries. Our findings conclude that Cyber-crime has turn into one of the most difficult problems and threatening for the judicial system, public at large and the law enforcement. It seems hard to put the criminals under the justice. It is difficult to get details or facts to the computer crime or internet that is desirable to prove someone a perpetrator in a courtyard of justice. As computer forensic investigates of cyber security incident or crimes but it is still new; most of the countries has not literature or expertise.

1. INTRODUCTION

Cyber-crime is any criminal activity done by a person using a computer or network. Cyber-crime is an illegal activity committed on the internet. The internet in Pakistan is growing rapidly. The first cyber-crime ever recorded was in France during 1820. The person involved was Joseph-Marie Jacquard. He was a textile manufacturer, he made looms. The first spam email took place in 1978 when it was sent over the Arpanet. The first virus was installed on an Apple computer in 1982 when a high school student, Richard Skrenta, developed the EIK Cloner. Internet also has its own disadvantages i.e. Cybercrime. We can define "Internet" as the collection of billion of computers that give a network to the electronic connections between devices or computer. There are million or trillion of devices which are connected through internet. Everyone appreciate internet but there is another dark side of internet that is cyber-crime. Cyber- crime is actually a crime related to the use of social media, computers and mobile phones etc. The cyber-crimes are getting very vast and fast now a days. While most the people use this commanding technology for illegal purposes as well. This paper is mainly explores the issues that has been faced and the describes the implemented laws and also suggestions which are helpful to enhance the laws. That's why I have choose this topic. Now a days cyber-crime has increased to the extent that no one is safe, either he is a poor person living hand to mouth or a business man, either he is a boy or girl, old or young, no one is safe. I am doing research on cyber-crime because people has very less knowledge and knowhow of types of cyber-crimes. My research will help the people to know the types of cyber-crimes, laws and concerned authorities for different kind of cyber-crimes. Some of the kinds of Cyber-criminals are mentioned as below.

- **Crackers:** These individuals break into others system and cause the loss to boost antisocial motives or usually just for fun. Many people who create the computer viruses also fall in this classification.
- **Hackers:** These individuals get into others computer systems to learn something, to compete with other hackers or just to satisfy their curiosity. They usually may be attempting to get the use of more better computer, gain the appreciation from other fellow hackers and getting the title of an expert without any degree.
- **Pranksters:** These individuals perform tricks on people. They usually do not direct any specific or long lasting hurt.
- **Cyber terrorists:** There are many types of cyber terrorism, sometimes its the group of users with same mind who get on a website together and crash it by making a rush there, sometimes its a smart hacker breaking into a government websites**.**

[1] The new smartphone and cloud computing technology, electronic commerce and social applications are mainly transforming the market outlook of the world. Business organizations are designing policies to achieve the emerging advantages of this new technology in the exchange of information and direct internet commerce.

[2] Therefore, demand for high-level authentication is enhanced to minimize the risks of contract loss and consumer unhappiness, the overall amount of company deal done online is about eighty percent of the total transactions. Cyber Defense is not restricted to entering the organization's local information technology base; rather it is growing to a much wider border network and IT base.

[3] In the current progress and implementation of essential computing and connections, the role of cyber security cannot be overlooked.

[4] The core players for the development of cyber security and the defense of a basic digital base are national security and fair outcomes. In 978-1-4799-7620-1/15/$31.00 ©2015 IEEE to restore utilities and government policy, it is all necessary to make the internet more stable and protect the internet user from cyber attacks

[5] It is all important to include warn of cybercrimes as a part of national plan for safety of computer and communication base and cyber security. A full outlook is thus needed for the development and performance of plan on national scale for cyber security

[6] It is essential to include cybercrime alerts as part of the National Information Protection and Connectivity Base and Cyber Security Strategy. A complete outlook is therefore necessary for the implementation and implementation of a national cyber security strategy.

[7] In the fight against cybercrime, these national or individual networks are required players. In order to stand up against cybercrimes, a full offer is needed; offering only technological solutions is not enough to deter any crime. Law enforcement authorities should be allowed to investigate and prosecute cybercrime quickly and efficiently.

[8] Malicious program that connects itself to other programs. Examples of malicious malware that ruins the victim machine (virus, worms, Trojan Horse, Time Bomb, Logic Bomb, Rabbit and Bacterium).

[9] A new study ranked India as the fourteenth nation to host phishing websites worldwide in 2008.

[10] The economic effect of security negligence is of concern to enterprises seeking to determine where to position their data security budget and to insurance providers who have cyber-risk policies.

[11] A ruling in favor of Ingram. Micro observed that physical harm is not limited to physical destruction or computer circuit disruption, but requires lack of use and intent.

[12] In addition, enterprises can fear that the public announcement of a security violation is of negative financial value. Previous analysis shows that media news about an occurrence that is normally viewed as detrimental can cause a decline in the stock price of the company.

[13] Using standard accounting-based metrics such as the Return on Investment (ROI) approach, risk management may be achieved.

[14] Furthermore since non-physical costs are not clearly observable, possible hypothetical damages such as lack of strategic advantage arising from neglect and loss of honor are not included.

[15] There are 90 percent garbage and 10 percent decent security schemes on the Internet,

[16] Because of management shame, fear of potential offences, businesses are unable to report these abuses.

2. METHODOLOGY AND ANALYSIS

In this section we describe the basic statistical analysis of cyber crimes by their categories, Region, Year, and types.

Table 1
Descriptive Analysis of Robbery

		Region	Sub Region	Country	Year
Mean		2.19	7.33	41.94	2010.94
Median		2.00	7.00	37.50	2011.00
Mode		1	1	6[a]	2010
Variance		1.696	20.212	2715.996	.659
Skewness		.447	.372	9.414	.118
Std. Error of Skewness		.177	.177	.177	.177
Kurtosis		-1.424	-.510	112.173	-1.473
Std. Error of Kurtosis		.353	.353	.353	.353
Range		4	18	667	2
Percentiles	25	1.00	4.00	19.00	2010.00
	50	2.00	7.00	37.50	2011.00
	75	3.00	10.75	55.75	2012.00

The average of the region is 2.19 and the median is 2.00. Mode (maximum value) is 1. The variance of the region is 1.696 and skewness is 0.447(positively skewed). The kurtosis of the region is -1.424 and the range of region is 4. Whereas the mean(average) of the sub-region is 7.33 and the median is 7.00. Mode (maximum value) is 1. The variance of the sub-region is 20.212 and skewness is 0.372 (positively skewed). The kurtosis of the sub-region is -.510 and the range of the sub-region is 18 respectively. The mean(average) of the country is 41.94 and the median is 37.50. Mode (maximum) 6[a]. The variance of the country is 2715.996 and skewness is 9.414 (positively skewed). The kurtosis of the country is 112.173 and the range of the country is 667. The mean(average) of the year is 2010.94 and the median is 2011.00 and the Mode is 2010. The variance of the year is .659 and skewness is .118 (positively skewed). The kurtosis of the year is -1.473 and the range of the year is 2. The mean(average) of the count is 38172.12 and the median is 2784.50. Mode (maximum) is 1. The variance of the count is 25671937316.842 and skewness is 5.493 (positively skewed). The kurtosis of the count is 29.885 and the range of the count is 1087058. The mean(average) of the rate is 99.866 and the median is 45.941. Mode (maximum). 5419. The variance of the rate is 24016.494 and skewness is 2.983 (positively skewed). The kurtosis of the rate is 9.886 and the range of the rate is 942.9035.

Table 2
Descriptive Analysis of Theft

		Rate	Region	Sub region
Mean		99.866589	2.78	6.41
Median		45.941465	3.00	5.00
Mode		.5419	1[a]	3
Variance		24016.494	1.961	21.184
Skewness		2.983	-.126	.933
Std. Error of Skewness		.177	.177	.177
Kurtosis		9.886	-1.422	-.148
Std. Error of Kurtosis		.353	.353	.353
Range		942.9035	4	17
Percentiles	25	15.406815	1.00	3.00
	50	45.941465	3.00	5.00
	75	101.543126	4.00	11.00

The mean(average) of the region is 2.78 and the median is 3.00. Mode (maximum value) is 1[a]. The variance of the region is 1.961 and skewness is -1.26(negatively skewed). The kurtosis of the region is -1.424 and the range of region is 4. The mean(average) of the sub-region is 6.41 and the median is 5.00. Mode (maximum value) is 3. The variance of the sub-region is 21.184 and the skewness is .933 (positively skewed). The kurtosis of the sub-region is -.148 and the range of the sub-region is 17. The mean(average) of the country is 51.52 and the median is 13.50. Mode (maximum) 6. The variance of the country is 2702.443 and skewness is .361 (positively skewed). The kurtosis of the country is -1.728 and the range of the country is 145. The mean(average) of the year is 2010.49 and the median is 2010.00 and the Mode is 2006. The variance of the year is 18.369 and skewness is .123 (positively skewed). The kurtosis of the year is -1.122 and the range of the year is 15. The mean(average) of the count is 100135.86 and the median is 5659.00. Mode (maximum) is 494[a]. The variance of the count is 25118077216.312 and skewness is 1.808 (positively skewed). The kurtosis of the count is 2.409 and the range of the count is 623688. The mean(average) of the rate is 783.844 and the median is 451.335. Mode (maximum).9.0193. The variance of the rate is 69000.438 and skewness is .961 (positively skewed). The kurtosis of the rate is -.505 and the range of the rate is 3170.918.

Table 3
Descriptive Analysis of Money Laundering

Descriptive		Sub-region	Country	Year	Count
Mean		7.63	61.55	2015.67	724.60
Median		7.00	62.50	2016.00	116.00
Mode		1	5	2016[a]	4
Variance		22.353	1347.896	2.682	3240956.348
Skewness		.300	.212	-.144	4.094
Std. Error of Skewness		.177	.177	.177	.177
Kurtosis		-.496	-.921	-1.094	17.550
Std. Error of Kurtosis		.353	.353	.353	.353
Range		17	132	5	11541
Percentiles	25	3.25	31.00	2014.00	25.25
	50	7.00	62.50	2016.00	116.00
	75	11.00	86.00	2017.00	463.75

The mean(average) of the year is 2015.67 and the median is 2016.00 and the Mode is 2016[a]. The variance of the year is 2.682 and skewness is -.144 (negatively skewed). The kurtosis of the year is -1.094 and the range of the year is 5. The mean(average) of the count is 724.60 and the median is 116.00. Mode (maximum) is 4. The variance of the count is 3240956.348 and skewness is 4.094 (positively skewed). The kurtosis of the count is 17.550 and the range of the count is 11541. The mean(average) of the rate is 8.578 and the median is 1.800 Mode (maximum) .100. The variance of the rate is 783.169 and skewness is 4.788 (positively skewed). The kurtosis of the rate is 22.644 and the range of the rate is 180.600.

Table 4
Descriptive Analysis of Kidnapping

		Region	Subregion	Country
Mean		2.61	6.34	46.30
Median		3.00	4.00	12.00
Mode		1	3[a]	15
Variance		1.790	22.812	2492.501
Skewness		-.018	1.048	.539
Std. Error of Skewness		.177	.177	.177
Kurtosis		-1.443	.117	-1.566
Std. Error of Kurtosis		.353	.353	.353
Range		4	17	150
Percentiles	25	1.00	3.00	6.00
	50	3.00	4.00	12.00
	75	4.00	11.00	100.00

The mean (average) of the region is 2.61 and the median is 3.00. Mode (maximum value) is 1. The variance of the region is 1.790 and skewness is -.018(negatively skewed). The kurtosis of the region is -1.443 and the range of region is 4. The mean(average) of the sub-region is 6.34 and the median is 4.00. Mode (maximum value) is 3[a] the variance of the sub-region is 22.812 and the skewness is 1.048 (positively skewed). The kurtosis of the sub-region is .117 and the range of the sub-region is 17. The mean(average) of the country is 46.30 and the median is 12.00. Mode (maximum) 15. The variance of the country is 2492.501 and skewness is .539 (positively skewed). The kurtosis of the country is -1.566 and the range of the country is 150. The mean(average) of the year is 2010.84 and the median is 2011.00 and the Mode is 2006. The variance of the year is 19.032 and skewness is -.029(negatively skewed). The kurtosis of the year is -1.172 and the range of the year is 15. The mean(average) of the count is 164.84 and the median is 19.00. Mode (maximum) is 5. The variance of the count is 107131.978 and skewness is 2.277 (positively skewed). The kurtosis of the count is 3.913 and the range of the count is 1303. The mean(average) of the rate is 2.263 and the median is .883 Mode (maximum) .000. The variance of the rate is 10.101 and skewness is 1.724 (positively skewed). The kurtosis of the rate is 1.991 and the range of the rate is 14.671.

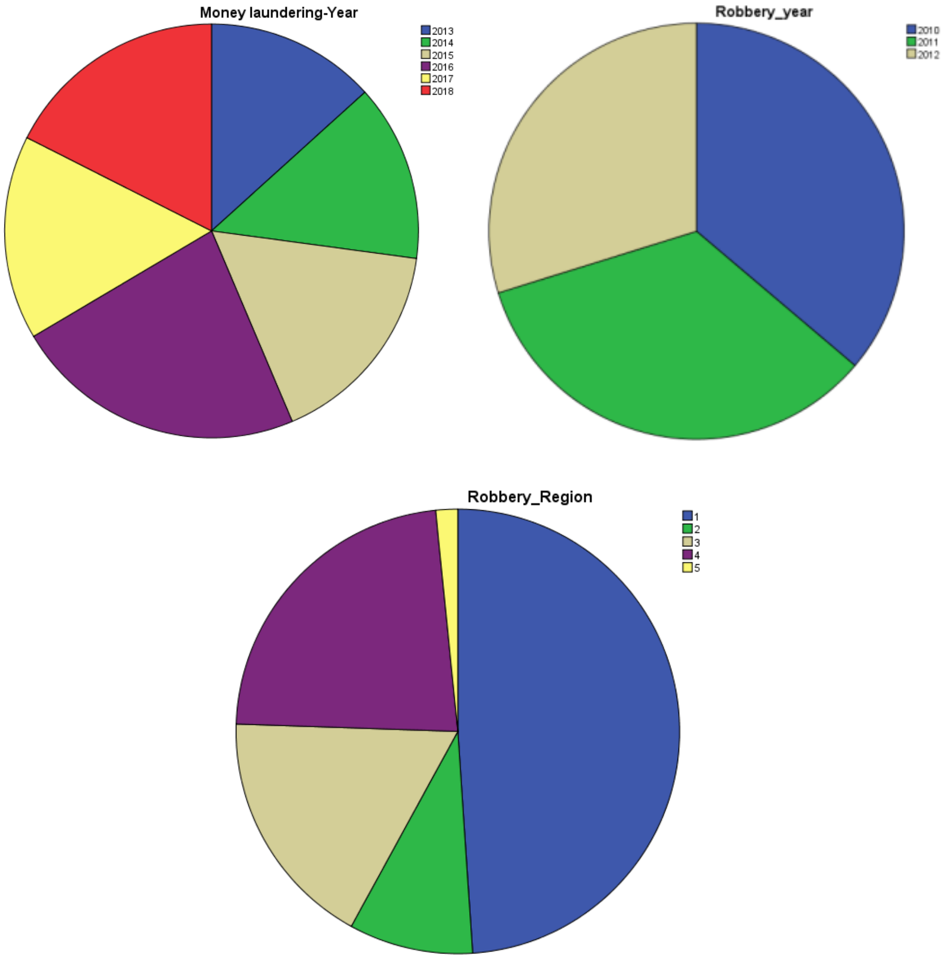


Fig 1: Pie Chart of Cyber Crimes by Year and Region

From above figures it is clear that maximum crimes are done in Europe region then America then Asia, Africa and very less ratio of crimes in Oceania region.

3. COMMENTS AND CONCLUSION

In this study, the statistical data on the Cyber Security cases in Pakistan is studied. The statistical data is collected and compiled. The estimated results are used to infer the main methods of Cyber Crime attempts in different regions of Pakistan. This data and topic has been chosen as this is one of the most emerging crime now a days and needs to be addressed properly, have to find out different methods that are being used and have to find out how to stop them. Main goal is to spread awareness on how to stay safe from any of the cybercrime.

REFERENCES

1. Rezgui, Y. and Marks, A. (2008). Information security awareness in higher education: An exploratory study. *Computers & Security*, 27(7-8), 241-253.
2. Broadhurst, R. and Chang, L. Y. (2013). Cybercrime in Asia: trends and challenges. In *Handbook of Asian criminology* (pp. 49-63). Springer, New York, NY.
3. Gartenstein-Ross, D. and Dabruzzi, K. (2007). The convergence of crime and terror. *Policing terrorism report*, (1). DOI: https://www.rieas.gr/researchareas/editorial/707-the-convergence-of-crime-and-terror
4. Hecht, J. (2007). When web browsers turn bad. *New Scientist*, 194(2602), 28-29.
5. Dashora, K. (2011). Cybercrime in the society: Problems and preventions. *Journal of Alternative Perspectives in the social sciences*, 3(1), 240-259.
6. Kshetri, N. (2006). The simple economics of cybercrimes. *IEEE Security & Privacy*, 4(1), 33-39.
7. Gercke, M. (2011). *Understanding Cybercrime: A Guide for Developing Countries*.
8. International Telecommunications Union, Cybercrime Legislation Resources. https://www.itu.int/ITU-D/cyb/cybersecurity/docs/ ITU_Guide_A5_12072011.pdf
9. Chung, W., Chen, H., Chang, W. and Chou, S. (2006). Fighting cybercrime: a review and the Taiwan experience. *Decision Support Systems*, 41(3), 669-682.
10. Wolanin, N., Smith, R.G. and Worthington, G. (2003). E-crime solutions and crime displacement. *Trends and Issues in Crime and Criminal Justice*, (243), 1-6.
11. Glossary, V. (2006). *Virus Dissemination*, Available at: http://www.virtualpune.com/citizencentre/html/cyber_crime_glossary.shtml
12. *India emerging as major cybercrime centre* (2009), Available at: http://wegathernews.com/ 203/indiaemerging-as-major-cyber-crime-centre/, Visited: 10/31/09.
13. Gordon, L.A., Loeb, M.P. and Sohail, T. (2003). A framework for using insurance for cyber-risk management. *Communications of the ACM*, 46(3), 81-85.
14. Ariz, D. (2000). American guarantee & liability insurance co. v. Ingram Micro. *Inc*, 104, 99-185.
15. Sprecher, C.R. and Perti, M.A. (1988). Intra-industry effects of the MGM Grand fire. *Quarterly journal of Business and Economics*, 27(2), 96-116.

16. Baskerville, R. (1991). Risk analysis: an interpretive feasibility tool in justifying information systems security. *European Journal of Information Systems*, 1(2), 121-130.
17. D'Amico, A. (2000). *What Does a Computer Security Breach Really Cost*? Global Information Assurance Certification Paper, The Sans Institute
18. Aseef, N., Davis, P., Mittal, M., Sedky, K. and Tolba, A. (2005). Cyber-criminal activity and analysis. *White Paper*. DOI: https://courses.cs.washington.edu/courses/csep590/05au/whitepaper_turnin/team2-whitepaper.pdf
19. Hoffer, J.A. and Straub Jr, D.W. (1989). The 9 to 5 underground: are you policing computer crimes?. *MIT Sloan Management Review*, 30(4), 35-43.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 271-282

GENERALIZED ESTIMATOR OF THE FINITE POPULATION MEAN USING AUXILIARY VARIABLE IN PRESENCE OF MEASUREMENT ERROR

Sarah Kanwal, Mahnaz Makhdum and **Jiya Ali**
Lahore College for Women University, Lahore, Pakistan
Email: Sarahkanwal39@gmail.com
mahnaz.makhdum@lcwu.edu.pk
jiya.amir16@gmail.com

ABSTRACT

In this paper, generalized difference cum exponential type estimator has been developed for estimating the population mean of the study variable using auxiliary information under simple random sampling scheme in the presence of measurement error non-response with two special cases. The expressions for the bias and mean square errors of proposed estimator have been derived. A comparative study is made among the proposed estimators, the Hansen and Hurwitz (1946) estimator, the Cochran's (1977) estimator, Rao (1986) Estimator, Bahl and Tuteja (1991) Estimators, Kumar and Bhougal (2011) Estimator.

KEYWORDS

Auxiliary variable, bias, dual estimator, exponential estimator, measurement error, mean square error, non-sampling errors, non-response.

1. INTRODUCTION

In sample survey, we estimate of population parameter on the basis of collected data. It may originate from various kinds of sampling methods such as simple random sampling, stratified sampling, systematic sampling etc. Various methods of estimation are used under the assumption that observations collected are true (error free). In real life this kind of situations are not tenable. The real data contains observational error due to many reasons like memory failure, over-reporting, under-reporting, prestige bias etc. These are also called measurement error [Cochran (1997) and Cheng, and Van Ness (1999)].

Shalabh (1997) presented a contribution on the performance of ratio method of estimation in presence of measurement error. Manisha and Singh (2001) used a combination of ratio estimator with sample mean estimator to express the useful performance in presence of measurement error. Shrivastava and Shalabh (1997) discussed a new property of stain procedure assuming measurement error model. Kumar (2016) examined measurement error and suggested an exponential ratio method of estimation in the presence of measurement error. Sharma et al. (2018) developed improved class of estimators of population mean in presence of measurement error using auxiliary variable based on arithmetic mean, geometric mean and harmonic mean of the usual unbiased estimator, Shalabh estimator (1997) and estimators due to Bahal and Tuteja (1991).

Some other useful contributions over applications of measurement error models are due to Fuller (1987), Cochran (1997), Kumar et al. (2011), Murthy (1977), Sabir & Sanaullah (2019), Shukla et al. (2012) etc. This paper presents an estimation strategy under measurement error model in the light of Azeem & Hanif (2017) methodology. Azeem & Hanif (2017) a scheme to jointly integrate the non-response and measurement error in the estimators of population mean with auxiliary information in SRS.

2. NOTATIONS

A simple random sample of size n is selected from the population of size N by a simple random sampling without replacement (SRSWOR) method. Let $(\boldsymbol{x_i}, \boldsymbol{y_i})$ be the observed values and $(\boldsymbol{X_i}, \boldsymbol{Y_i})$ be the true values on two characteristics $(\boldsymbol{x}, \boldsymbol{y})$ respectively associated with the $\boldsymbol{i^{th}}(\boldsymbol{i = 1, 2, \ldots, n})$ sample unit. Let the measurement error associated with the study variable be

$$\boldsymbol{U_i = y_i - Y_i} \tag{2.1}$$

$$\boldsymbol{V_i = x_i - X_i} \tag{2.2}$$

The measurement error is assumed to be random in nature and they are uncorrelated with mean zero and variances $\boldsymbol{S_v^2} = \frac{\boldsymbol{\sum_{i=1}^{N}(v_i - \bar{V})^2}}{\boldsymbol{N_1 - 1}}$ and $\boldsymbol{S_u^2} = \frac{\boldsymbol{\sum_{i=1}^{N}(u_i - \bar{U})^2}}{\boldsymbol{N_1 - 1}}$ respectively. It is also assumed that the true values of the variables Y and X are independent of measurement error. Also assume the following notation

Let introduced some notation

$$\omega_y = \sum_{i=1}^{n}(Y_i - \bar{Y}) \tag{2.3}$$

$$\omega_x = \sum_{i=1}^{n}(X_i - \bar{X}) \tag{2.4}$$

$$\omega_u = \sum_{i=1}^{n} U_i \tag{2.5}$$

$$\omega_v = \sum_{i=1}^{n} V_i \tag{2.6}$$

In (2.3) $Y_i - \bar{Y}$ is the ith unit deviation of true values from the mean of variable and then deviation are summed over all sample, here we may recall the measurement error linked with the study Y and auxiliary X variables

$$U_i = y_i - Y_i;\ V_i = x_i - X_i$$

$$\omega_y + \omega_u = \sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n} U_i \tag{2.7}$$

by dividing both sides by n, we get

$$\frac{1}{n}(\omega_y + \omega_u) = \frac{1}{n}\left[\sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n} U_i\right]$$

$$\frac{1}{n}(\omega_y + \omega_u) = \frac{1}{n}\left[\sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n}(y_i - Y_i)\right]$$

$$\frac{1}{n}(\omega_y + \omega_u) = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \bar{Y})\right]$$

or

$$\bar{y} = \bar{Y} + \frac{1}{n}(\omega_y + \omega_u) \tag{2.8}$$

Similarly

$$\bar{x} = \bar{X} + \frac{1}{n}(\omega_x + \omega_v) \tag{2.9}$$

Squaring the both sides of (4.6), we have:

$$(\omega_y + \omega_u)^2 = \left[\sum_{i=1}^{n}(Y_i - \bar{Y}) + \sum_{i=1}^{n} U_i\right]^2$$

$$(\omega_y + \omega_u)^2 = \left\{\sum_{i=1}^{n}(Y_i - \bar{Y})\right\}^2 + \left\{\sum_{i=1}^{n} U_i\right\}^2 + 2\left\{\sum_{i=1}^{n}(Y_i - \bar{Y})\right\}\left\{\sum_{i=1}^{n} U_i\right\}$$

Applying expectation on both sides of above equation we get:

$$E\,(\omega_y + \omega_u)^2 = E\left\{\sum_{i=1}^{n}(Y_i - \bar{Y})\right\}^2 + E\left\{\sum_{i=1}^{n}(U_i - \bar{U})\right\}^2 + 2E\left\{\sum_{i=1}^{n}(Y_i - \bar{Y})\right\}\left\{\sum_{i=1}^{n} U_i\right\}$$

Since the true values of variables Y are independent of measurement error, thus on applying the expectation the cross product term become zero;

$$E\,(\omega_y + \omega_u)^2 = \sum_{i=1}^{n} E\,(Y_i - \bar{Y})^2 + \sum_{i=1}^{n} E\,(U_i - \bar{U})^2$$

$$E\,(\omega_y + \omega_u)^2 = \sum_{i=1}^{n} S_Y^2 + \sum_{i=1}^{n} S_U^2 = n(S_Y^2 + S_U^2)$$

Substituting f.p.c in above equation

$$E\left(\omega_y+\omega_u\right)^2 = n(1-f)(S_Y^2+S_U^2)$$

$$E\left(\omega_y+\omega_u\right)^2 = \frac{n^2(1-f)}{n}(S_Y^2+S_U^2)$$

$$E\left(\omega_y+\omega_u\right)^2 = n^2\lambda_2(S_Y^2+S_U^2)$$

Further dividing both sides by n

$$E\left(\frac{\omega_y+\omega_u}{n}\right)^2 = \lambda_2(S_Y^2+S_U^2)$$

and also

$$\left.\begin{aligned} E\left(\frac{1}{n}\left(\omega_y+\omega_u\right)\right)^2 &= \lambda_2(S_Y^2+S_U^2)\\ E\left(\frac{1}{n}\left(\omega_x+\omega_v\right)\right)^2 &= \lambda_2(S_X^2+S_V^2)\\ E\left[\frac{1}{n}\left(\omega_y+\omega_u\right).\frac{1}{n}\left(\omega_x+\omega_v\right)\right] &= \lambda_2\rho_{YX}S_YS_X \end{aligned}\right\} \tag{2.10}$$

and also let

$$\frac{1}{\bar{\bar{X}}}-\frac{\delta}{\bar{\bar{X}}}-\frac{\alpha}{\bar{\bar{X}}}=\beta \text{ and } \tau=\frac{1}{2\bar{\bar{X}}}-\frac{\delta}{\bar{\bar{X}}}$$

$$A_0=\beta^2\lambda_2(S_X^2+S_V^2)+\frac{1}{\bar{\bar{Y}}^2}\lambda_2(S_Y^2+S_U^2)+2\beta\frac{1}{\bar{\bar{Y}}}\lambda_2\rho_{YX}S_YS_X$$

$$A_1=\left(\frac{1}{\bar{\bar{Y}}^2}\lambda_2(S_X^2+S_V^2)\right)$$

$$A_2=1+\frac{\delta^2}{\bar{\bar{X}}^2}\lambda_2(S_X^2+S_V^2)+\frac{1}{\bar{\bar{Y}}^2}\lambda_2(S_Y^2+S_U^2)+4\tau\frac{1}{\bar{\bar{Y}}}\lambda_2\rho_{YX}S_YS_X$$

$$A_3=2\tau\frac{1}{\bar{\bar{Y}}}\lambda_2(S_X^2+S_V^2)+\frac{1}{\bar{\bar{Y}}^2}\lambda_2\rho_{YX}S_YS_X$$

$$A_4=\beta\frac{1}{\bar{\bar{Y}}}\lambda_2(S_X^2+S_V^2)+\frac{1}{\bar{\bar{Y}}^2}\lambda_2\rho_{YX}S_YS_X$$

$$A_5=\left(\frac{\alpha\delta}{\bar{\bar{X}}^2}+\tau\beta\right)\lambda_2(S_X^2+S_V^2)+\frac{1}{\bar{\bar{Y}}}(2\beta+\tau)\lambda_2\rho_{YX}S_YS_X+\frac{1}{\bar{\bar{Y}}^2}\lambda_2(S_Y^2+S_U^2).$$

3. EXISTING ESTIMATORS

A traditional mean per unit (or Mean) estimator is a well-known in the literature and in setup of measurement error as shown in (2.1). It is unbiased also

$$MSE(\bar{y})=\lambda_2(S_Y^2+S_U^2) \tag{3.1}$$

For estimating, the sample statistic provides an unbiased estimator. In this, we are not using any additional information. To use auxiliary X, one can adopt several ways laid down under $\boldsymbol{Y}$.

Cochran (1997) developed the following ratio and product type estimator of population mean. The estimator in the presence of measurement error and non-response is given as:

$$\boldsymbol{t_r = \frac{\bar{y}}{\bar{x}}\bar{X}, t_p = \frac{\bar{y}}{\bar{X}}\bar{x}.}$$

Mean square error of above mention estimators in the presence of measurement error and non-response is given as:

$$MSE(t_r) \approx \lambda_2(S_Y^2 + R^2 S_X^2 - 2R\rho_{YX} S_Y S_X) + \lambda_2(S_U^2 + R^2 S_V^2) \tag{3.2}$$

and

$$MSE(t_p) \approx \lambda_2(S_Y^2 + R^2 S_X^2 + 2R\rho_{YX} S_Y S_X) + \lambda_2(S_U^2 + R^2 S_V^2) \tag{3.3}$$

Cochran (1997) regression estimator under measurement error is given as

$$\boldsymbol{t_{reg} = \bar{y} + b(\bar{X} - \bar{x})}$$

The minimum MSE of regression estimator is given as

$$\boldsymbol{min. MSE(t_{reg}) = \lambda_2(S_Y^2 + S_U^2) + b_0^2\left(\lambda_2(S_x^2 + S_v^2)\right) - 2b_0(\lambda_2 \rho_{YX} S_Y S_X)} \tag{3.4}$$

where

$$\boldsymbol{b_o = \frac{\lambda_2 \rho_{YX} S_Y S_X}{\lambda_2(S_x^2 + S_v^2)}}.$$

Rao (1991) suggested a difference-type estimator for population mean in the presence of measurement error and non-response is defined as:

$$\boldsymbol{T_{rd} = \omega_1(\bar{X} - \bar{x}) + \omega_2 \bar{y}}$$

where $\boldsymbol{\omega_1}$ and $\boldsymbol{\omega_2}$ are suitable chosen constants.

$$MSE(t_{rd}) \approx \bar{Y}^2 - 2\omega_2\bar{Y}^2 + \omega_2^2\bar{Y}^2 + \omega_1^2\lambda_2(S_x^2 + S_v^2) + \omega_2^2\lambda_2(S_Y^2 + S_U^2) - 2\omega_1\omega_2\lambda_2\rho_{YX}S_YS_X \tag{3.5}$$

where as

$$\boldsymbol{\omega_1 = \frac{\bar{Y}^2\, \lambda_2 \rho_{YX} S_Y S_X}{\lambda_2(S_x^2 + S_v^2)\left(\bar{Y}^2 + \lambda_2(S_Y^2 + S_U^2)\right) - (\lambda_2 \rho_{YX} S_Y S_X)^2}}$$

$$\boldsymbol{\omega_2 = \frac{\bar{Y}^2\, \lambda_2 \rho_{YX} S_Y S_X}{\lambda_2(S_x^2 + S_v^2)\left(\bar{Y}^2 + \lambda_2(S_Y^2 + S_U^2)\right) - (\lambda_2 \rho_{YX} S_Y S_X)^2}}.$$

Kumar and Bhougal (2011) suggested the following exponential-type estimator of population mean in the presence of measurement error and non-response.

$$\boldsymbol{t_{kb}} = \overline{\boldsymbol{y}}\left\{\propto \boldsymbol{exp}\left(\frac{\overline{\boldsymbol{X}} - \overline{\boldsymbol{x}}}{\overline{\boldsymbol{X}} + \overline{\boldsymbol{x}}}\right) + (\boldsymbol{1}-\propto)\boldsymbol{exp}\left(\frac{\overline{\boldsymbol{x}} - \overline{\boldsymbol{X}}}{\overline{\boldsymbol{X}} + \overline{\boldsymbol{x}}}\right)\right\} \tag{3.6}$$

where α is a real constant such that the mean square error of $\boldsymbol{t_{kb}}$ is minimum.

$$\boldsymbol{min.MSE}(\boldsymbol{t_{kb}}) \approx \boldsymbol{\lambda}_2\left[\boldsymbol{S_Y^2} + \left(\boldsymbol{\alpha_o} - \frac{\boldsymbol{1}}{\boldsymbol{2}}\right)^{\boldsymbol{2}}\boldsymbol{R^2S_X^2} - \boldsymbol{2}\left(\boldsymbol{\alpha_o} - \frac{\boldsymbol{1}}{\boldsymbol{2}}\right)\boldsymbol{R\rho_{YX}S_YS_X}\right] + \boldsymbol{\lambda}_2\left[\boldsymbol{S_U^2} + \left(\boldsymbol{\alpha_o} - \frac{\boldsymbol{1}}{\boldsymbol{2}}\right)^{\boldsymbol{2}}\boldsymbol{R^2S_V^2}\right]$$

where

$$\boldsymbol{\alpha_0} \approx \frac{\boldsymbol{1}}{\boldsymbol{2}} + \frac{\boldsymbol{\lambda_2R\rho_{YX}S_YS_X}}{\boldsymbol{\lambda_2R^2S_X^2} + \boldsymbol{\lambda_2R^2S_V^2}}.$$

4. PROPOSED ESTIMATOR

Taking motivation from Pak et al. (2018) if there is a *measurement error* on both study and auxiliary variables, we suggest a modified generalized difference cum exponential type estimator as:

$$t_{sm} = \left[\bar{y}\left\{\alpha\ exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) + (1-\alpha)\ exp\left(\frac{\bar{x} - \bar{X}}{\bar{X} + \bar{x}}\right)\right\} + \eta_1(\bar{X} - \bar{x}) + \eta_2\bar{y}\right] \times \left[\delta exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) + (1-\delta)\ exp\left(\frac{\bar{x} - \bar{X}}{\bar{X} + \bar{x}}\right)\right] \tag{4.1}$$

where (α, δ) are constants such that $0 \le (\alpha, \delta) \le 1$ and (η_1, η_2) are constants such that MSE of the proposed class of estimators $t_{sp(2)}$ is minimum.

Using the (2.8) and (2.9) we get

$$t_{sm} = \left[\bar{Y} - \frac{\omega_y + \omega_u}{n}\left\{\alpha \exp\frac{\bar{X} - \bar{X} - \frac{\omega_x + \omega_v}{n}}{\bar{X} + \bar{X} + \frac{\omega_x + \omega_v}{n}} + (1-\alpha)\exp\frac{\bar{X} - \bar{X} + \frac{\omega_x + \omega_v}{n}}{\bar{X} + \bar{X} + \frac{\omega_x + \omega_v}{n}}\right\} + \eta_1\left(\bar{X} - \bar{X} - \frac{\omega_x + \omega_v}{n}\right) + \eta_2\bar{Y} + \frac{\eta_2}{n}(\omega_y + \omega_u)\right]\left[\delta\exp\frac{\bar{X} - \bar{X} - \frac{\omega_x + \omega_v}{n}}{\bar{X} + \bar{X} + \frac{\omega_x + \omega_v}{n}} + (1-\delta)\exp\frac{\bar{X} - \bar{X} + \frac{\omega_x + \omega_v}{n}}{\bar{X} + \bar{X} + \frac{\omega_x + \omega_v}{n}}\right].$$

By expanding the Taylor's series in above equation and by ignore the higher order of approximation we get:

$$t_{sm} = \left[\left(\bar{Y} + \frac{1}{n}(\omega_y + \omega_u)\right)\left\{\alpha \exp\left\{\frac{-1}{2\bar{X}}\frac{(\omega_x + \omega_v)}{n} + \frac{1}{4\bar{X}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2\right\}\right.\right.$$
$$\left. + (1-\alpha)\exp\left\{\frac{1}{2\bar{X}}\frac{(\omega_x + \omega_v)}{n} - \frac{1}{4\bar{X}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2\right\}\right\}$$
$$- \eta_1\frac{(\omega_x + \omega_v)}{n} + \eta_2\bar{Y}$$
$$\left. + \frac{\eta_2}{n}(\omega_y + \omega_u)\right]\left[\delta\exp\left\{\frac{-1}{2\bar{X}}\frac{(\omega_x + \omega_v)}{n} + \frac{1}{4\bar{X}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2\right\}\right.$$
$$\left. + (1-\delta)\exp\left\{\frac{1}{2\bar{X}}\frac{(\omega_x + \omega_v)}{n} - \frac{1}{4\bar{X}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2\right\}\right]$$

So the Bias for the proposed estimator is given as:

$$Bias(t_{sm}) = \bar{Y}\left[\frac{\alpha\delta}{\bar{X}^2}\lambda_2(S_X^2 + S_V^2) + \left(\frac{1}{\bar{X}\bar{Y}} - \frac{\delta}{\bar{X}\bar{Y}} - \frac{\alpha}{\bar{X}\bar{Y}}\right)\lambda_2\rho_{YX}S_YS_X\right.$$
$$- \eta_1\left(\frac{1}{2\bar{Y}\bar{X}} - \frac{\delta}{\bar{Y}\bar{X}}\right)\lambda_2(S_X^2 + S_V^2)$$
$$+ \eta_2\left\{1 - \left(\frac{1}{8\bar{X}^2} - \frac{\delta}{2\bar{X}^2}\right)\lambda_2(S_X^2 + S_V^2)\right.$$
$$\left.\left. + \left(\frac{1}{2\bar{X}\bar{Y}} - \frac{\delta}{\bar{X}\bar{Y}}\right)\lambda_2\rho_{YX}S_YS_X\right\}\right] \quad (4.2)$$

So the mean square error of proposed estimator is:

$$(t_{sm} - \bar{Y})^2 = \bar{Y}^2\left[\beta^2\left(\frac{\omega_x + \omega_v}{n}\right)^2 + \frac{1}{\bar{Y}^2}\left(\frac{\omega_y + \omega_u}{n}\right)^2\right.$$
$$+ 2\beta\frac{1}{\bar{Y}}\frac{(\omega_x + \omega_v)}{n}\frac{(\omega_y + \omega_u)}{n} + \eta_1^2\frac{1}{\bar{Y}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2$$
$$+ \eta_2^2\left\{1 + \frac{\delta^2}{\bar{X}^2}\left(\frac{\omega_x + \omega_v}{n}\right)^2 + \frac{1}{\bar{Y}^2}\left(\frac{\omega_y + \omega_u}{n}\right)^2\right.$$
$$\left. + 4\frac{1}{\bar{Y}}\tau\frac{(\omega_x + \omega_v)}{n}\frac{(\omega_y + \omega_u)}{n}\right\}$$
$$- 2\eta_1\left\{\beta\frac{1}{\bar{Y}}\left(\frac{\omega_x + \omega_v}{n}\right)^2 + \frac{1}{\bar{Y}^2}\frac{(\omega_x + \omega_v)}{n}\frac{(\omega_y + \omega_u)}{n}\right\}$$
$$+ 2\eta_2\left\{\left(\frac{\alpha\delta}{\bar{X}^2} + \tau\beta\right)\left(\frac{\omega_x + \omega_v}{n}\right)^2\right.$$
$$\left. + \frac{1}{\bar{Y}}(2\beta + \tau)\frac{(\omega_y + \omega_u)}{n}\frac{(\omega_x + \omega_v)}{n} + \frac{1}{\bar{Y}^2}\left(\frac{\omega_y + \omega_u}{n}\right)^2\right\}$$
$$\left. - 2\eta_1\eta_2\left\{2\frac{1}{\bar{Y}}\tau\left(\frac{\omega_x + \omega_v}{n}\right)^2 + \frac{1}{\bar{Y}^2}\frac{(\omega_x + \omega_v)}{n}\frac{(\omega_y + \omega_u)}{n}\right\}\right]$$

By substituting the notation, the equation becomes:

$$Mse(t_{sm)}) \ = \bar{Y}^2[A_0 + \eta_1^2 A_1 + \eta_2^2 A_2 - \ 2\eta_1\eta_2 A_3 - 2\eta_1 A_4 + 2\eta_2 \ A_5\ ] \tag{4.3}$$

Differentiating Equation *(4.3)* w.r.t η_1 and η_2 and equating them to zero, we have

$$\begin{bmatrix} A_1 & -A_3 \\ -A_3 & A_2 \end{bmatrix}\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} A_4 \\ -A_5 \end{bmatrix} \tag{4.4}$$

It can be written as

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} A_1 & -A_3 \\ -A_3 & A_2 \end{bmatrix}^{-1}\begin{bmatrix} A_4 \\ -A_5 \end{bmatrix}$$

and

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \frac{1}{A_1A_2 - A_3^2}\begin{bmatrix} A_2 & A_3 \\ A_3 & A_1 \end{bmatrix}\begin{bmatrix} A_4 \\ -A_5 \end{bmatrix}$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \frac{1}{A_1A_2 - A_3^2}\begin{bmatrix} A_2A_4 - A_3A_5 \\ A_3A_4 - A_1A_5 \end{bmatrix}$$

After simplification of equation optimum values of to η_1 and η_2 are

$$\left.\begin{aligned} \eta_{11} &= \frac{A_2A_4 - A_3A_5}{A_1A_2 - A_3^2} \\ n_{21} &= \frac{A_3A_4 - A_1A_5}{A_1A_2 - A_3^2} \end{aligned}\right] \tag{4.5}$$

Substituting the value of in η_1 and η_2 in equation (4.11)

$$\begin{aligned} min.MSE(t_{sm}) \ &= \bar{Y}^2[A_0 + \eta_{11}^2 A_1 + \eta_{21}^2 A_2 - \ 2\eta_{11}\eta_{21} A_3 - 2\eta_{11} A_4 \\ &\quad + \ 2\eta_{21}\ A_5\ ] \end{aligned} \tag{4.6}$$

or we can write it as

$$min.Mse(t_{sm}) \ = \bar{Y}^2 \left[A_0 - \frac{A_2A_4^2 \ - \ 2A_3A_4A_5 \ + \ A_1A_5^2}{(A_1A_2 \ - \ A_3^2)}\right].$$

5. SIMULATION STUDY

In this section, for empirical study, by using R language program, we have generated four different populations from normal distribution with different parameters. In simulation study the population for Y and X is considered using as multivariate normal, drawn from the following populations.

By using R program a Monte-Carlo simulation study is been carried out to validate the results. The data matrix has been generated for the different population sizes dividing into two non-overlapping groups N_1 and N_2.The data matrix on X, Y, V and U have been generated using multivariate normal distribution for variables with mean vectors $(\bar{Y}, \bar{X}, 0, 0)$ and covariance matrix.

$$cov = \begin{bmatrix} S_Y^2 & \rho_{XY}S_YS_X & 0 & 0 \\ \rho_{XY}S_YS_X & S_X^2 & 0 & 0 \\ 0 & 0 & S_U^2 & 0 \\ 0 & 0 & 0 & S_V^2 \end{bmatrix}$$

The populations are given below.

Population 1:

Let the peach production of June 1946 in North Carolina and y be bushels in an orchard and x be the number of peach trees in the orchard. The statistics for this data set are:

$$N = 256;\ n = 100;\ \rho = 0.887$$

$$\mu = \begin{bmatrix} 56.47 \\ 44.45 \end{bmatrix};\ cov = \begin{bmatrix} 6430.019 & 4426.185 \\ 4426.185 & 3872.573 \end{bmatrix}.$$

Population 2:

$$N = 5000;\ n = 500;\ \rho = -0.18275$$

$$\mu = \begin{bmatrix} 1.007391 \\ 9.953722 \end{bmatrix};\ cov = \begin{bmatrix} 1.0310780 & -0.3740459 \\ -0.3740459 & 4.0629690 \end{bmatrix}.$$

Population 3:

$$N = 5000;\ n = 500\ \rho = 0.394221$$

$$\mu = \begin{bmatrix} 4.961780 \\ 4.961081 \end{bmatrix};\ cov = \begin{bmatrix} 102.24080 & 40.03396 \\ 40.03396 & 100.86800 \end{bmatrix}.$$

6. CONCLUSION

The Table 5.1 (see Appendix) shows the MSEs and PREs of existing similar type estimator in the presence of measurement error on both study and auxiliary variables. Table 5.2 shows the MSEs and PREs of proposed modified generalized difference-cum-exponential-type estimator t_{sm} estimator in the presence of measurement error on both study and auxiliary variables for various populations. It is observed from tables that the theoretical results and the results obtained through simulation study are almost equivalent which shows the validity of the simulation study.

The proposed modified generalized difference-cum-exponential-type estimator t_{sm} in the presence of measurement error on both study and auxiliary variables have highest PREs as compare to modified Hansen and Hurwitz (1946), Cochran's (1977), Rao(1986), Bahl and Tuteja (1991), Kumar and Bhougal (2011) generalized exponential estimators for all given populations.

It is also observed that the product type estimator has least MSE than some existing similar type estimators for the population 1 and 3 because of positive correlation coefficient and ratio type estimators has least PREs w.r.t modified Hansen and Hurwitz for the population 2 due to negative correlation coefficient between study and auxiliary variables.

PREs of modified generalized estimator t_{sm} increases with the increase in value of α while keeping δ constant for all populations. The highest values of PREs are obtain at the following combination of (α, δ) = (0.25, 0.75), (0.50, 0.75), (0.75, 0.75) and at (1, 0.75).

REFERENCES

1. Azeem, M. and Hanif, M. (2017). Joint influence of measurement error and non-response on estimation of population mean. *Communications in Statistics-Theory and Methods*, 46(4), 1679-1693.
2. Bahl, S. and Tuteja, R. (1991). Ratio and product type exponential estimators. *Journal of Information and Optimization Sciences*, 12(1), 159-164.
3. Cheng, L. and Van Ness, J.W. (1999). *Statistical regression with measurement error*. Arnold Publishers, London.
4. Cochran, W. (1997). *Sampling techniques*. 3rd ed, John Wiley & Sons, Inc. New York.
5. Fuller, W.A. (1987). *Measurement Error Models.* Wiley, New York.
6. Kreuter, F., Müller, G. and Trappmann, M. (2010). Nonresponse and measurement error in employment research: making use of administrative data. *Public Opinion Quarterly*, 74(5), 880-906.
7. Kumar, M., Singh, R., Sawan, N. and Chauhan, P. (2011). Exponential ratio method of estimation in the presence of measurement errors. *Int. J. Agri. Statist. Sci*., 7(2), 457-461.
8. Kumar, S. (2016). Improved estimation of population mean in presence of nonresponse and measurement error. *Journal of Statistical Theory and Practice*, 10(4), 707-720.
9. Kumar, S. and Bhougal, S. (2011). Estimation of the population mean in presence of non-response. *Communications for Statistical Applications and Methods*, 18(4), 537-548.
10. Manisha and Singh, R. (2001). An estimation of population mean in the presence of measurement error. *Journal of Indian Society Agricultural Statistics*, 54(1), 13-18.
11. Murthy, M.N. (1977). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.
12. Pal, S.K., Singh, H.P., Kumar, S. and Chatterjee, K. (2018). A family of efficient estimators of the finite population mean in simple random sampling. *Journal of Statistical Computation and Simulation*, 88(5), 920-934.
13. Rao, T. (1991). On certail methods of improving ration and regression estimators. *Communications in Statistics- Theory and Methods*, 20(10), 3325-3340.
14. Sabir, S. and Sanaullah, A. (2019). A Note on Kumar (2016). Improved Estimation of Population Mean in the Presence of Measurement Error and Non-response. *Journal of Statistical Theory and Practice*, 13(1), 25.
15. Shalabh (1997). Ratio method of estimation in the presence of measurement errors. *Journal of Indian Society of Agricultural Statistics*, 50(2), 150-155.
16. Sharma, A., Adichwal, N. K., Singh, R. and Kumar, N. (2018). Improved estimators for population mean in presence of measurement error. *International Journal of Statistics and Applied Mathematics*; 3(1), 201-210.
17. Shrivastava, A.K. and Shalabh (1997). A new property of Stein procedure in measurement error model. *Statistics & Probability Letters*, 32(3), 231-234.
18. Shukla, D., Pathak, S. and Thakur, N.S. (2012). An estimator for mean estimation in presence of measurement error. *Res. Rev. A. J. Stat.*, 1(1), 1-8.

APPENDIX

Table 5.1
MSEs and PREs of Some Modified Existing Similar Type Estimators for Various Populations

EST	Pop 1		Pop 2		Pop 3	
	MSE Theoretical (Simulated)	**PRE**	**MSE Theoretical (Simulated)**	**PRE**	**MSE Theoretical (Simulated)**	**PRE**
$\bar{y}^*$	**10.0409** (9.9950)	100	**0.0036** (0.0035)	100	**0.2318** (0.2300)	100
t_r	**2.0327** (2.0633)	493.9658	**0.0038** (0.0037)	94.0461	**0.3112** (0.3353)	74.5148
t_p	**36.706** (36.547)	27.35448	**0.0035** (0.0035)	101.259	**0.5917** (0.5739)	39.185
t_{re}	**3.7046** (3.7090)	271.0374	**0.0037** (0.0036)	97.5330	**0.2166** (0.2228)	107.036
t_{pe}	**21.042** (21.916)	47.71937	**0.0035** (0.0035)	101.273	**0.3569** (0.3479)	64.9674
t_{reg}	**1.9859** (1.9949)	505.5899	**0.0035** (0.0035)	101.429	**0.2094** (0.2102)	110.695
t_{rd}	**1.9848** (1.9925)	505.8885	**0.0035** (0.0034)	101.570	**0.2078** (0.2067)	111.611
t_{kb}	**9.3851** (9.3845)	106.9889	**0.0036** (0.0035)	100.002	**0.2095** (0.2126)	110.690

Table 5.2
MSEs and PREs of the Proposed Modified Generalized Difference Cum Exponential Type Estimator t_{sm} in the Presence of Measurement Error on Both Auxiliary and Study Variables for Various Populations

α	δ	**Pop 1**		**Pop 2**		**Pop 3**	
		MSE Theoretical (Simulated)	**PRE**	**MSE Theoretical (Simulated)**	**PRE**	**MSE Theoretical (Simulated)**	**PRE**
0.25	0.25	**1.9822** (1.9937)	506.5567	**0.0035** (0.0034)	101.786	**0.2077** (0.2058)	111.646
	0.50	**1.9776** (1.9976)	507.7385	**0.0035** (0.0034)	101.789	**0.2069** (0.2061)	112.049
	0.75	**1.9756** (2.0007)	**508.2375**	**0.0035** (0.0034)	101.790	**0.2067** (0.2069)	**112.200**
	1	**1.9777** (2.0003)	507.7400	**0.0035** (0.0034)	101.789	**0.2069** (0.2081)	112.052
0.50	0.25	**1.9843** (1.9934)	506.0901	**0.0035** (0.0034)	101.792	**0.2075** (0.2063)	111.737
	0.50	**1.9805** (1.9967)	507.0039	**0.0035** (0.0034)	101.795	**0.2067** (0.2071)	112.158
	0.75	**1.9789** (1.8103)	**507.4126**	**0.0035** (0.0034)	**101.796**	**0.2064** (0.2085)	**112.315**
	1	**1.9804** (1.0419)	507.0048	**0.0035** (0.0034)	101.795	**0.2067** (0.2010)	112.161
0.75	0.25	**1.9854** (1.9951)	505.7761	**0.0035** (0.0034)	101.798	**0.2073** (0.2077)	111.833
	0.50	**1.9827** (1.9976)	506.4233	**0.0035** (0.0034)	101.801	**0.2065** (0.2085)	112.271
	0.75	**1.9815** (1.9919)	**506.7426**	**0.0035** (0.0034)	**101.802**	**0.2062** (0.2103)	**112.434**
	1	**1.9827** (1.9356)	506.4239	**0.0035** (0.0034)	101.801	**0.2065** (0.2125)	112.274
1	0.25	**1.9859** (1.9985)	505.6142	**0.0035** (0.0034)	101.803	**0.2071** (0.2081)	111.931
	0.50	**1.9844** (1.8903)	505.9959	**0.0035** (0.0034)	101.807	**0.2063** (0.2101)	112.381
	0.75	**1.98349** (1.7653)	**506.2262**	**0.0035** (0.0034)	**101.808**	**0.2060** (0.2125)	**112.555**
	1	**1.9844** (1.80911)	505.9962	**0.0035** (0.0034)	101.807	**0.2063** (0.2151)	112.392

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 283-290

IMPROVED NON-PARAMETRIC SURVIVAL METHOD FOR CENSORED DATA

Umair Khalil, Dost Muhammad Khan and **Naseem Asghar§**
Department of Statistics, Abdul Wali Khan University Mardan, Pakistan
Email: umairkhalil@awkum.edu.pk
dostmuhammad@awkum.edu.pk
nsmsghr@gmail.com
§ Corresponding author

ABSTRACT

This study aimed to compare estimates of Kaplan-Meier (K-M) and Improved Weighted Kaplan-Meier (I-W-K-M) as an alternative method to deal with the problem of censored data for COVID-19 positive admitted patients, who recovered during their stay at the hospital from March 2020 to August 2020. The six-month recovery estimates for these patients were evaluated based on K-M and I-W-K-M. A total of 213 (57.26%) patients recovered after hospitalization for some time (in days), whereas, 159 (42.74%) patients censored at the end of the study. The median survival time for these patients was 13 days.

1. INTRODUCTION

The most useful and standard estimator in medical research is the Kaplan-Meier (K-M) also called the product-limit estimator (PLE) [1]. However, another estimator, which is an alternative of K-M, recommended by Nelson [2] and considered by Aalen [3], used in applied statistics is the Nelson-Aalen estimator. The Nelson–Aalen estimator (NAE) is also called the cumulative hazard estimator (CHE). The product-limit estimator and cumulative hazard estimator is based on counting process theory [4]. Specifically, it has just been demonstrated that both the PLE and the CHE are asymptotically comparable. Fleming and Harrington [5] demonstrated that the Kaplan Meier estimator (KME) meets to zero with increasing sample size. Estimates of survival function at some predefined time can be developed by utilizing both NAE and KME. It is realized that estimates obtained by KME are smaller than estimates obtained from CHE [6]. Klein and Moesch Berger express that the CH estimator has a superior small size performance over the PL estimator [7]. A few authors have suggested linear interpolation, rather than step-function directly, to acquire better estimates by utilizing the two estimators [8].

In statistical research, various techniques have been proposed for survival function estimates. Peto [9] recommended a non-parametric methodology and Turnbull [10] determined a self-reliable calculation for its estimation. In a strategy considered by Finkelstein [11], where she suggested the utilization of the non-parametric maximum likelihood technique to fit the Cox model. In a spline strategy, Kooperberg and stone [12] acquired estimates of the survival function. A calculation method proposed by W. Pan iteratively ascribes the unobserved event times from a prespecified model [13], [14], while a few different researchers considered Bayesian techniques [15], [16]. Lindsey and Ryan's [17] instructional exercise gives a brilliant audit of the writing. Taylor et al. [18] utilized various attribution techniques in the analysis of right-censored survival data. Their work has built up a hypothetical establishment for the utilization of the attribution strategies with regards to survival analysis.

This study aims to propose a weighted product limit estimator called the improved weighted Kaplan-Meier (I-W-K-M) estimator. The motivation behind this proposed estimator is to involve the contribution of censored observations with probabilities in weights assignment, so that to assign, weights in consistent order without rapid decay or gaps and to bring uniformity in weights. This proposed estimator will give consistent and smaller probabilities in comparison to the K-M estimator.

2. SURVIVAL FUNCTIONS

Let t_j denote the actual times of failure of n individuals in a cohort, d_j denote the number of failures at t_j and let n_j denote the number of items at risk of failure at t_j. The Kaplan-Meier estimator at the time t_j given by [1] and is defined as

$$\hat{S}_{KM}(t_j) = \prod_{j:t_j \le t} \frac{n_j - d_j}{n_j} \tag{1}$$

Another way of expressing the survival function is given by [2] and is defined as

$$S(t) = \exp\big(-H(t)\big) \tag{2}$$

where, $H(t)$ is the cumulative hazard function. Equation (2) suggests that the estimation of $S(t)$ could also be based on $H(t)$. The Nelson-Aalen estimator of $H(t)$ is given by [3] and is defined as

$$\widehat{H}_{NA}(t) = \sum \frac{d_j}{n_j} \tag{3}$$

To deal with the censored time $t_{(j)}$a weight with time $t_{(j)}$ denotd by w_j, is defined as $W_j = \left\{1 - \frac{c_j}{n_j}\right\}$ known as non-censored rate. Where $W_j = 1$ if there is no censoring and $W_j < 1$ in case of censoring at time t_j. The weighted Kaplan-Meier estimator given by [19] and defined as

$$\hat{S}_{WKM}(t) = \prod_{j:t_j \le t} w_j\{(n_j - d_j)/n_j\} \tag{4}$$

To bring consistency in weight assignment, the contribution of probabilities with censored observations are incorporated to form new weights as w_j, where $w_j = \left[n_j - \frac{c_j \times p_j}{n_j}\right]/n_j$ is a weight function and, $w_j \le 1$.

The proposed weighted product limit estimator then become

$$\hat{S}_{IWKM}(t_j) = \prod_{j:t_j \le t} w_j \left[\frac{n_j - d_j}{n_j}\right]. \tag{5}$$

3. METHODS

To assess the relative performance of the proposed method by simulation under different situations, data sets were generated with event time and censored time from an exponential distribution, with shape parameter [19]. The same was applied for censored time, as every observation was a pair of (time, event) or (time, censor). The simulation study was based on sample size generated randomly from exponential distribution and replicated a fixed number of times [20]. The estimates, standard errors, and confidence limits were calculated through which the performance of the weighted product limit estimator was assessed.

Similarly, to investigate the performance of the proposed weighted product limit estimator, a real data set was used in this study to estimate the recovery time of 372 COVID-19 admitted patients during the first wave of COVID-19 pandemic at Saidu Group of Teaching Hospitals Saidu Sharif Swat from March 2020 to August 2020. The recovery time of patients was determined from the date of admission to the date of discharge from the hospital. The expired patients were considered as censored. The Kaplan Meier estimator and improved weighted Kaplan Meier estimator were used in the analysis.

4. SIMULATION

The simulation study is based on a sample of size 100 generated randomly and replicated 5,000 times. The Greenwood and Peto's standard errors are calculated. The averages of each sample of size 100 are taken and plotted as figures 1-4. Estimates with its lower limit and upper limit are plotted in blue and green respectively. The proposed method gives the shortest confidence interval than the existing method.

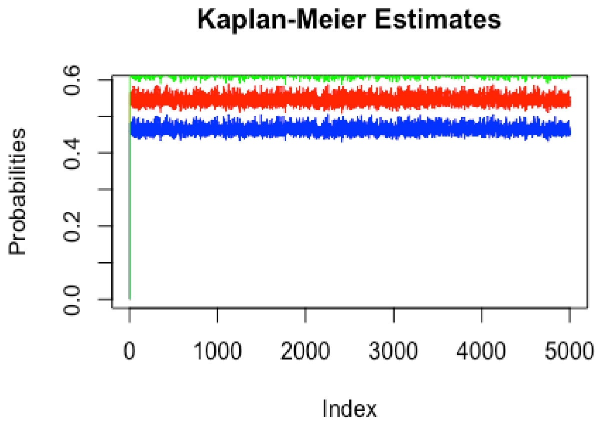


Figure 1: Mean of 95% confidence intervals for $S_{KM}(t)$ using Greenwood's, method. Survival times are Exp (10) distributed, with censoring from an Exp (50) distribution. Sample size n is 100, with 5,000 replications for each case.

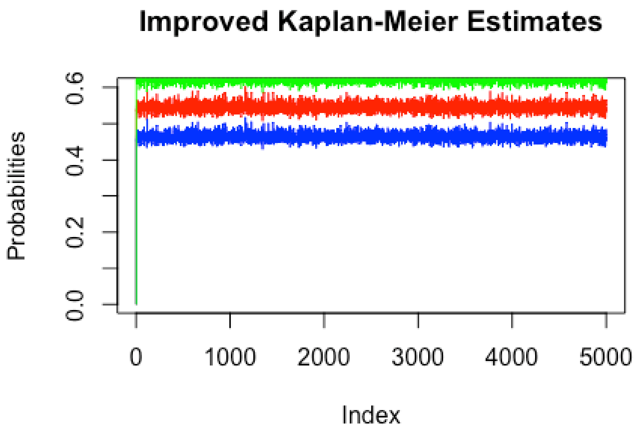


Figure 2: Mean of 95% confidence intervals for $S_{IWKM}(t)$ using Greenwood's, method. Survival times are Exp (10) distributed, with censoring from an Exp (50) distribution. Sample size n is 100, with 5,000 replications for each case.

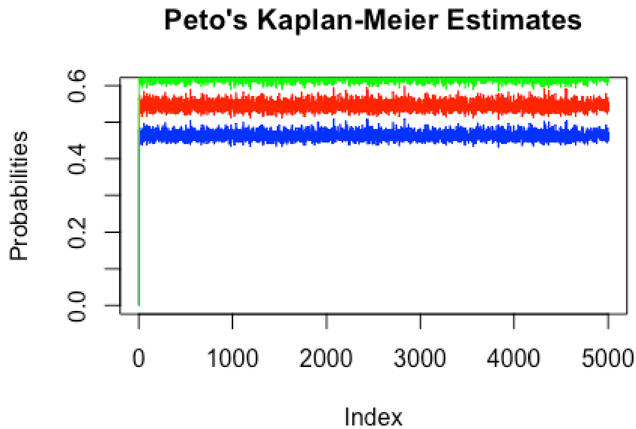


Figure 3: Mean of 95% confidence intervals for $S_{KM}(t)$ using Peto's, method. Survival times are Exp (10) distributed, with censoring from an Exp (50) distribution. Sample size n is 100, with 5,000 replications for each case.

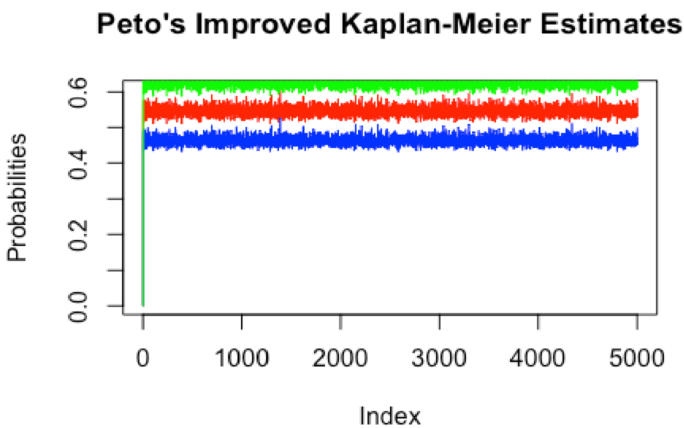


Figure 4: Mean of 95% confidence intervals for $S_{IWKM}(t)$ using Peto's, method. Survival times are Exp (10) distributed, with censoring from an Exp (50) distribution. Sample size n is 100, with 5,000 replications for each case.

Example:

Table 1 shows 372 COVID-19 patients from the initial time to its end time that is date of admission to date of discharge from the hospital. The total number of days spent by the patients is 29. Recovery estimates are calculated using $\widehat{S}_{KM}(t)$ and $\hat{S}_{IWKM}(t)$. Confidence limits are included in the table for both estimators.

Table 1
K-M and I-W-K-M Recovery Estimates of 372 COVID-19 Admitted Patients

t	d	c	n	$\hat{S}_{KM}(t)$	SE	LL	UL	w	$\hat{S}_{IWKM}(t)$	SE	LL	UL
0	0	0	372	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000
1	8	13	372	0.9785	0.0075	0.9575	0.9892	0.9999	0.9784	0.0075	0.9576	0.9892
2	27	26	351	0.9032	0.0156	0.8678	0.9295	0.9998	0.9030	0.0155	0.8679	0.9295
3	21	20	298	0.8396	0.0197	0.7965	0.8742	0.9998	0.8394	0.0197	0.7966	0.8742
4	22	12	257	0.7677	0.0232	0.7183	0.8096	0.9998	0.7676	0.0232	0.7184	0.8096
5	19	12	223	0.7023	0.0256	0.6487	0.7493	0.9998	0.7021	0.0256	0.6488	0.7493
6	27	12	192	0.6035	0.0282	0.5459	0.6563	0.9997	0.6034	0.0282	0.5459	0.6563
7	15	6	153	0.5444	0.0293	0.4851	0.5997	0.9998	0.5442	0.0293	0.4852	0.5997
8	21	11	132	0.4578	0.0301	0.3979	0.5155	0.9995	0.4575	0.0301	0.3979	0.5155
9	12	6	100	0.4028	0.0304	0.3431	0.4617	0.9995	0.4026	0.0304	0.3431	0.4617
10	7	7	82	0.3684	0.0305	0.3090	0.4279	0.9990	0.3681	0.0304	0.3091	0.4278
11	7	5	68	0.3305	0.0305	0.2715	0.3905	0.9990	0.3302	0.0305	0.2716	0.3905
12	4	4	56	0.3069	0.0305	0.2483	0.3673	0.9988	0.3065	0.0305	0.2483	0.3673
13	5	6	48	0.2749	0.0305	0.2169	0.3359	0.9977	0.2743	0.0304	0.2170	0.3358
14	3	0	37	0.2526	0.0306	0.1949	0.3143	1.0000	0.2526	0.0306	0.1949	0.3143
15	0	3	34	0.2526	0.0306	0.1949	0.3143	0.9974	0.2520	0.0306	0.1950	0.3141
16	1	1	31	0.2445	0.0307	0.1868	0.3064	0.9990	0.2442	0.0307	0.1869	0.3064
17	3	1	29	0.2192	0.0308	0.1621	0.2820	0.9989	0.2190	0.0308	0.1621	0.2820
18	2	4	25	0.2017	0.0307	0.1452	0.2648	0.9941	0.2005	0.0306	0.1454	0.2646
19	2	3	19	0.1804	0.0310	0.1245	0.2448	0.9926	0.1791	0.0307	0.1247	0.2446
20	2	2	14	0.1547	0.0314	0.0992	0.2214	0.9913	0.1533	0.0312	0.0994	0.2211
21	0	1	10	0.1547	0.0314	0.0992	0.2214	0.9900	0.1531	0.0311	0.0995	0.2210
22	2	0	9	0.1203	0.0325	0.0659	0.1922	1.0000	0.1203	0.0325	0.0659	0.1922
23	1	0	7	0.1031	0.0321	0.0512	0.1761	1.0000	0.1031	0.0321	0.0512	0.1761
24	0	1	6	0.1031	0.0321	0.0512	0.1761	0.9722	0.1002	0.0312	0.0517	0.1751
25	0	2	5	0.1031	0.0321	0.0512	0.1761	0.9200	0.0949	0.0295	0.0527	0.1732
28	1	1	3	0.0687	0.0353	0.0203	0.1590	0.9259	0.0636	0.0327	0.0211	0.1560
29	1	0	1	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000

5. COMMENTS AND CONCLUSION

Monte Carlo simulations were performed using R to compare the behavior of the standard estimator with the proposed estimator. The simulation study was based on exponential distribution with a mean of 10 for survival time and a mean of 50 for event/censor time. The confidence interval was calculated using Greenwood and Peto's standard errors. The simulation study shows almost the same but better results for proposed improved weighted K-M method than K-M.

Columns 5 and 10 of table-1 shows estimates of K-M and improved weighted K-M respectively. A close look of column 5 and 10 shows that

$$\hat{S}_{KM}(t_j) \geq \hat{S}_{IWKM}(t_j) \Longrightarrow \hat{S}_{KM}(t_j) - \hat{S}_{IWKM}(t_j) \geq 0 \text{ for all } j \tag{6}$$

If there is no observation censored then

$$\hat{S}_{KM}(t_j) = \hat{S}_{IWKM}(t_j) \Longrightarrow \hat{S}_{KM}(t_j) - \hat{S}_{IWKM}(t_j) = 0 \text{ for all } j \tag{7}$$

If there is no event occur but the observation censored then

$$\hat{S}_{KM}(t_j) > \hat{S}_{IWKM}(t_j) \Longrightarrow \hat{S}_{KM}(t_j) - \hat{S}_{IWKM}(t_j) > 0 \text{ for all } j \tag{8}$$

If there are events occur with observation censored then

$$\hat{S}_{KM}(t_j) > \hat{S}_{IWKM}(t_j) \Longrightarrow \hat{S}_{KM}(t_j) - \hat{S}_{IWKM}(t_j) > 0 \text{ for all } j \tag{9}$$

If there are no events occur consecutively but observations are censoring then $\hat{S}_{KM}(t_j)$ is a constant repeatedly, whereas, $\widehat{S}_{IWKM}(t_j)$ is an estimate changing with respect to censoring time.

From practical results for K-M and improved weighted K-M: $\hat{S}(t_j) = 1$ when there is no failure, $\hat{S}(t_j) = 0$ when all members of the population failed. Also, for improved weighted K-M, $\hat{S}_{IWKM}(t_j) = 0$, when the last single observation is censored but will give estimates when more than one observation is censored.

The K-M method causes survival probability to be constant at points when no event occurs consecutively while the number of subjects at risk decreases markedly. Improved weighted K-M estimator gives zero weight to the last censored observation but gives proper weight when a group of observations is censored at the end of the study, which is a good indication. Hence the improved weighted K-M gives zero weight to the last censored observation, appropriate weight to the group of censored observations at the end of the study, and close probabilities to the K-M.

REFERENCES

1. Meier, P. (1975). Estimation of a distribution function from incomplete observations. *Journal of Applied Probability*, 12(S1), 67-87.
2. Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4), 945-966.
3. Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4), 701-726.

4. Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.
5. Fleming, T.R. and Harrington, D.P. (1991). *Counting processes and survival analysis*. John Wiley & Sons. Inc. New York.
6. Bohoris, G.A. (1994). Comparison of the cumulative-hazard and Kaplan-Meier estimators of the survivor function. *IEEE Transactions on Reliability*, 43(2), 230-232.
7. Klein, J.P. and Moeschberger, M.L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.
8. Lee, E.T. and Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.
9. Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society: series C (Applied Statistics)*, 22(1), 86-91.
10. Waller, L.A. and Turnbull, B.W. (1992). Probability plotting with censored data. *The American Statistician*, 46(1), 5-12.
11. Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 845-854.
12. Kooperberg, C. and Stone, C.J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4), 301-328.
13. Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56(1), 199-203.
14. Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics*, 58(1), 64-70.
15. Sinha, D., Chen, M.H. and Ghosh, S.K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, 55(2), 585-590.
16. Calle, M.L. and Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference*, 98(1-2), 73-87.
17. Lindsey, J.C. and Ryan, L.M. (1998). Methods for interval-censored data. *Statistics in medicine*, 17(2), 219-238.
18. Taylor, J.M., Murray, S. and Hsu, C.H. (2002). Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, 58(3), 221-232.
19. Jan, B. (2004). *Improved Inferences in the context of Survival/Failure Time* (Doctoral dissertation, Ph.D. Thesis University of Peshawar, Pakistan).
20. Harezlak, J. and Tu, W. (2006). Estimation of survival functions in interval and right censored data using STD behavioural diaries. *Statistics in medicine*, 25(23), 4053-4064.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 291-304

SOME CHARACTERIZATIONS OF UNIMODAL, RESERVE J-SHAPED DISTRIBUTION

Aqsa Rafique and **Naz Saud**
Lahore College for Women University, Lahore, Pakistan.
Email: nazahmad16@yahoo.com
aqsa.aim@gmail.com

ABSTRACT

Some characterizations of the unimodal, reverse J-shaped distribution named as Exponentiated Generalized Frechet Geometric (EGFG) distribution are presented in paper. These characterizations are based on three different methods:

i. Left and right truncated moments.
ii. Certain function of random variable.
iii. Based on truncated moment of the first order statistic.

1. INTRODUCTION

The characterization of probability distribution is important before its application to any real world phenomena to confirm whether the given probability distribution is suitable for the specific data set or not. It plays a vital role in different fields of mathematical and statistical sciences. Many authors have discussed the characterizations of probability distributions. For example, Ahsanullah et al. (2017) studied the characterization of Lindely distribution by truncated moments. Recently Ahsanullah and Shakil (2020) discussed some characterizations of Chaudhry and Zubair Extended Generalized Inverse Gaussian Distribution.

2. UNIMODAL, RESERVE J-SHAPED DISTRIBUTION

The unimodal, reverse J-shaped distribution is presented along with structural properties that are used in this article. Some structural properties had already been studied in Rafique and Saud (2019). The distribution function and density function with parameters $\eta, \kappa, \varphi, \delta \geq 0$ and $\tau \epsilon (0,1)$ are defined respectively,

$$f(y;\underline{\theta}) = \frac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}} \quad (1)$$

$$F(y;\underline{\theta}) = 1 - \frac{(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}$$

where $\underline{\theta} = \{\kappa, \varphi, \delta, \tau, \eta\}$ (2)

with κ, φ, δ, τ are shape parameters and scale parameter is η. We can write, expansion of distribution function and density function as,

$$F(y,\underline{\theta}) = 1 - \sum_{z=1}^{\infty}\sum_{b=0}^{\infty}\sum_{m=0}^{\infty}\sum_{y=0}^{\infty}(1-\tau)\tau^{z-1}\binom{z}{b}\binom{\delta b}{m}\binom{\kappa m}{x}(-1)^{b+m+x}e^{-x\left(\frac{\eta}{y}\right)^{\phi}}$$

$$F(y,\underline{\theta}) = 1 - M_{zbmx}e^{-x\left(\frac{\eta}{y}\right)^{\phi}} \tag{3}$$

where

$$M_{zbmx} = \sum_{z=1}^{\infty}\sum_{b=0}^{\infty}\sum_{m=0}^{\infty}\sum_{x=0}^{\infty}(1-\tau)\tau^{z-1}\binom{z}{b}\binom{\delta b}{m}\binom{\kappa m}{x}(-1)^{b+m+x} \tag{4}$$

and

$$f(y) = \sum_{b=0}^{\infty}\sum_{m=0}^{\infty}\sum_{a=1}^{\infty}\sum_{z=1}^{\infty}(1-\tau)\tau^{z-1}\binom{z}{a}\binom{\delta(a)-1}{m}\binom{\kappa(m+1)-1}{b}(-1)^{a-1+m+b}$$

$$\kappa(a\delta)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(b+1)\left(\frac{\eta}{y}\right)^{\phi}} \tag{5}$$

The hazard rate function and reverse hazard rate function of the EGFG distribution is defined as

2.1 Characterization of Distribution by Truncated Moments

We will give two characterization theorems for the EGFG distribution, with cdf (2) and pdf (1), by truncated moments. The first characterization theorem is based on the relation between hazard rate and left truncated moment. The second characterization theorem is based on the relation between reverse hazard rate and right truncated moment. To state these theorems, we will present two supportive lemmas. Many researchers such as Shakil et al. (2018), Nofal and Ahsanullah (2019) and Ahsanullah et al. (2016) have studied the characterization by truncated moments.

$$h(y) = \frac{\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)} \tag{6}$$

$$r(y) = \frac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{-1}}{\left[1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)-(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right]} \tag{7}$$

2.1.1 Characterization of EGFG under Left Truncated Moments

Here, we give supportive lemma for the characterization of the EGFG by left truncated moment.

Lemma 1

Suppose the absolutely continuous random variable Y has cdf F(y) with $F(0) = 0, F(y) > 0 \ \forall \ y > 0$, pdf $f(y) = F'(y)$, and $h(y) = \frac{f(y)}{1-F(y)}$ and $0 < E(Y\ ) < \infty$. If

$$E[Y\ |Y \geq y] = u(y)h(y), y > 0 \tag{8}$$

where $u(y)$ is a differentiable function in $y\epsilon(0, \infty)$, then

$$f(y) = Dexp\left[-\int_0^y \frac{t+u'(t)}{u(t)} \quad dt\right],\ y > 0$$

where $D > 0$ is a normalizing constant.

Proof:

Since

$$E[Y|Y \geq y] = \frac{1}{1-F(y)} \int_y^\infty tf(t)dt \tag{9}$$

Comparing (8) and (9), we have

$$u(y)h(y) = \frac{1}{1-F(y)} \int_y^\infty tf(t)dt$$

Substituting $h(y)$ in above equation, we obtain

$$f(y)u(y) = \int_y^\infty tf(t)dt \ or \quad u(y) = \frac{\int_y^\infty tf(t)dt}{f(y)} \tag{10}$$

Differentiating (10) w.r.t 'y', we obtain

$$u(y)f\ '(y)\ +\ u'(y)f(y)\ =\ -yf(y)$$

and hence

$$\frac{f'(y)}{f(y)} = -\frac{y+u'(y)}{u(y)}$$

Taking integral on both sides of the equation, we have

$$f(y) = Dexp\left[-\int_0^y \frac{t+u'(t)}{u(t)}dt\right]$$

where D is determined by

$$\int_0^\infty f(y)dy\ =\ 1$$

Theorem 1

Suppose the absolutely continuous random variable Y has cdf F(y) with $F(0)=0$, $F(y)>0\ \forall\ y>0$, p.d.f $f(y)=F'(y)$, then

$$E[Y\,|Y\ \geq\ y]=u(y)h(y), y>0$$

where and

$$h(y)=\frac{f(y)}{1-F(y)}$$

$$u(y)=\frac{\dfrac{y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}+\displaystyle\int_{y}^{\infty}M_{zbmx}e^{-x\left(\frac{\eta}{y}\right)^{\phi}}dt}{\dfrac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}}}$$

where, M_{zbmx} are defined in (4) and if and only if Y has pdf (1).

Proof:

$$E[Y|Y\geq y]=\frac{1}{1-F(y)}\int_{y}^{\infty}tf(t)dt$$

or

$$u(y)f(y)=\int_{y}^{\infty}tf(t)dt$$

$$u(y)=\frac{\int_{y}^{\infty}tf(t)dt}{f(y)}=\frac{-t(1-F(t))|_{y}^{\infty}+\int_{y}^{\infty}(1-F(t))dt}{f(y)}$$

Substituting (2) and (1). Then it is easily seen that

$$u(y)=\frac{\dfrac{y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}+\displaystyle\int_{y}^{\infty}\frac{(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{t})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{t})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}dt}{\dfrac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}}} \tag{11}$$

using (3)

$$\frac{(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{t})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{t})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}dt = M_{zbmx}e^{-x\left(\frac{\eta}{y}\right)^{\phi}} \tag{12}$$

where, M_{zbmx} are defined in (4)

Substituting (12) in (11), then we have

$$u(y) = \frac{\dfrac{y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)} + \displaystyle\int_y^{\infty} M_{zbmx}e^{-x\left(\frac{\eta}{y}\right)^{\phi}}dt}{\dfrac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}}}$$

Simple differentiation and simplification gives $u'(y) = -y - u(y)A(y)$, where

$$A(y) = \frac{f'(y)}{f(y)} = \tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]$$

$$\left[\frac{\tau^{-1}(\delta-1)\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{-\delta}}{\delta}\right.$$

$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{-1}}{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]}$$

$$\left.-2\left(1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\right]\Bigg\}$$

Then we have

$$u'(y) = -y - u(y)\Bigg\{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]$$
$$\Bigg[\frac{\tau^{-1}(\delta-1)\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{-\delta}}{\delta}$$
$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{-1}}{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]}$$
$$-2\left(1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\Bigg]\Bigg\}$$

from which we obtain

$$-\frac{y+u'(y)}{u(y)} = \tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]$$
$$\Bigg[\frac{\tau^{-1}(\delta-1)\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{-\delta}}{\delta}$$
$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{-1}}{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]}$$
$$-2\left(1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\Bigg]$$

By lemma (1), we have

$$\frac{f'(y)}{f(y)} = -\frac{y+u'(y)}{u(y)}$$

It follows that

$$\frac{f'(y)}{f(y)} = \left\{ \tau \left[\kappa\delta\phi\eta^{\phi} y^{-(\phi+1)} e^{-(\frac{\eta}{y})^{\phi}} \left(1 - e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1} \left[1 - \left(1 - e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1} \right] \right.$$

$$\left[\frac{\tau^{-1}(\delta-1)\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{-\delta}}{\delta} \right.$$

$$+ \frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{-1}}{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]}$$

$$\left. -2\left(1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\right]\right\}$$

On integrating the above expression with respect to 'y' and simplifying, we obtain

$$\ln f(y) = \ln\left[\frac{D(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}}\right]$$

Since D is determined by $\int_0^{\infty} f(y) = 1$, we have the pdf (1).

2.1.2 Characterization of Distribution by Right Truncated Moments

Here, we provide an essential lemma as a tool for the characterization of the EGFG distribution by right truncated moments.

Lemma 2

Suppose the absolutely continuous random variable Y has cdf $F(y)$ with

$F(0) = 0$, $F(y) > 0 \forall y > 0$, pdf $f(y) = F'(y)$, and $r(y) = \frac{f(y)}{F(y)}$ and $0 < E(Y) < \infty$.

If

$$E[Y \mid Y \leq y] = n(y)r(y), y > 0 \tag{13}$$

where $n(y)$ is a differentiable function in $y \varepsilon (0, \infty)$, then

$$f(y) = Dexp\left[\int_0^{y} \frac{t - n'(t)}{n(t)} \quad dt\right], \; y > 0$$

where $D > 0$ is a normalizing constant.

Proof:

Since

$$E[Y|Y \leq y] = \frac{1}{F(y)} \int_0^y tf(t)dt \tag{14}$$

Comparing (13) and (14), we have

$$n(y)r(y) = \frac{1}{F(y)} \int_0^y tf(t)dt$$

Substituting r(y) in above equation, we obtain

$$f(y)n(y) = \int_0^y tf(t)dt \ or \quad n(y) = \frac{\int_0^y tf(t)dt}{f(y)} \tag{15}$$

Differentiate (15) on both sides w.r.t 'y', we obtain

$$yf(y) = n(y)f'(y) + n'(y)f(y)$$

$$\frac{f'(y)}{f(y)} = \frac{y - n'(y)}{n(y)}$$

Taking integral on both sides of the equation, we have

$$f(y) = Dexp\left[\int_0^y \frac{t - n'(t)}{n(t)} dt\right]$$

and D is such that

$$\int_0^\infty f(y)dy = 1$$

Theorem 2

Suppose the absolutely continuous random variable Y has c.d.f. $F(y)$ with F(0) = 0, F(y) > 0 ∀ y > 0, p.d.f $f(y) = F'(y)$, then

$$E[Y|Y \leq y] = n(y)r(y), y > 0$$

where

$$r(y) = \frac{f(y)}{F(y)}$$

$$n(y) = \frac{\dfrac{-y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^\phi}\right)^\kappa\right]^\delta\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^\phi}\right)^\kappa\right]^\delta\right)} + \int_0^y M_{zbmx} e^{-x\left(\frac{\eta}{y}\right)^\phi} dt}{\dfrac{(1-\tau)\kappa\delta\phi\eta^\phi y^{-\phi-1} e^{-(\frac{\eta}{y})^\phi}\left(1-e^{-(\frac{\eta}{y})^\phi}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^\phi}\right)^\kappa\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^\phi}\right)^\kappa\right]^\delta\right)\right)^2}}$$

where, M_{zbmx} are defined in (4) and if and only if Y has pdf (1).

Proof:

$$E[Y|Y\leq y]=\frac{1}{F(y)}\int_0^y tf(t)dt$$

or

$$n(y)f(y)=\int_0^y tf(t)\ dt$$

$$n(y)f(y)=-t(1-F(t))|_0^y\int_0^y 1-F(t)\ dt$$

Substituting (2), then we have

$$n(y)=\frac{\frac{-y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}+\int_0^y\frac{-y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}}{\frac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^2}}$$

Substituting (12), we have

$$n(y)=\frac{\frac{-y(1-\tau)\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}{1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)}+\int_0^y M_{zbmx}e^{-x\left(\frac{\eta}{y}\right)^{\phi}}dt}{\frac{(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^2}}$$

where M_{zbmx} are defined in (4).

Simple differentiation and simplification gives $n'(y)=\ y-n(y)A(y)$, where

$$A(y) = \frac{f'(y)}{f(y)} = \tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]$$

$$\Bigg[\frac{\tau^{-1}(\delta-1)\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{-\delta}}{\delta}$$

$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{-1}}{\tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]}$$

$$-2\Big(1-\tau\Big(1-\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta}\Big)\Big)\Bigg]$$

then we have

$$n'(y) = y - n(y)\Bigg\{\tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]$$

$$\Bigg[\frac{\tau^{-1}(\delta-1)\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{-\delta}}{\delta}$$

$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{-1}}{\tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]}$$

$$-2\Big(1-\tau\Big(1-\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta}\Big)\Big)\Bigg]\Bigg\}$$

Thus

$$\frac{y-n'(y)}{n(y)} = \tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]$$

$$\Bigg[\frac{\tau^{-1}(\delta-1)\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{-\delta}}{\delta}$$

$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{-1}}{\tau\Big[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa-1}\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta-1}\Big]}$$

$$-2\Big(1-\tau\Big(1-\Big[1-\Big(1-e^{-(\frac{\eta}{y})^{\phi}}\Big)^{\kappa}\Big]^{\delta}\Big)\Big)\Bigg]$$

By lemma (14), we have

$$\frac{f'(y)}{f(y)}=\frac{y-n'(y)}{n(y)}$$

it follows that

$$\frac{f'(y)}{f(y)}=\left\{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]\right.$$

$$\left[\frac{\tau^{-1}(\delta-1)\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{-\delta}}{\delta}\right.$$

$$+\frac{-(\phi+1)y^{-1}+\phi\eta^{\phi}y^{-(\phi+1)}-(\kappa-1)\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{-1}}{\tau\left[\kappa\delta\phi\eta^{\phi}y^{-(\phi+1)}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}\right]}$$

$$\left.\left.-2\left(1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)\right]\right\} \tag{16}$$

On integrating (16) w.r.t to 'y', we have

$$\ln f(y)=\ln\left[\frac{D(1-\tau)\kappa\delta\phi\eta^{\phi}y^{-\phi-1}e^{-(\frac{\eta}{y})^{\phi}}\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa-1}\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta-1}}{\left(1-\tau\left(1-\left[1-\left(1-e^{-(\frac{\eta}{y})^{\phi}}\right)^{\kappa}\right]^{\delta}\right)\right)^{2}}\right]$$

Since D is determined by $\int_0^\infty f(y)dy = 1$, we have the pdf (1).

2.2 Characterizations via Certain Function of the Random Variable

Bhatti et al. (2019b), Bhatti et al. (2019a) and Bhatti et al. (2018) characterized the probability distributions through certain function of the random variable.

Proposition 1

Let $Y : \zeta \rightarrow (0,\infty)$ be a continuous random variable with its probability density function (1). Let $\zeta(y)$ and $Q(y)$ be two differentiable functions on $(0,\infty)$ such that

$$\int_0^\infty \frac{\zeta'(y)}{[\zeta(y)-Q(y)]}dy=\infty.$$

Then $E[Q(y)|Y > y)] = \zeta(y)$, $0 < y < \infty$, implies

$$F(y)=1-exp\left\{\int_0^y \frac{\zeta'(t)}{[\zeta(t)-Q(t)]}dt\right\},\ 0\leq y$$

Proof:

We have

$$\int_{y}^{\infty} Q(u)f(u)du = \left(1 - F(y)\right)\zeta(y)$$

Differentiating above expression w.r.t 'y', and then rearrange the terms, we have

$$\frac{f(y)}{1-F(y)} = \frac{\zeta'(y)}{[\zeta(y) - Q(y)]},\ 0 < y < \infty\,.$$

Integrating the last equation from 0 to y, we obtain

$$F(y) = 1 - exp\left\{\int_0^y \frac{\zeta'(t)}{[\zeta(t) - Q(t)]}dt\right\},\ 0 < y < \infty\,.$$

Remark 1

Taking

$$Q(y) = \left[\frac{(1-\tau)\left(1-\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)}\right)^{\kappa}\right]^{\delta}\right)\right)}{1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)}\right)^{\kappa}\right]^{\delta}\right)}\right]^{-\frac{1}{2}} \text{ and } \zeta(y) = 2Q(y),$$

Proposition (1) gives a characterization of the EGFG distribution. Clearly there are other choices of these functions.

2.3 Characterizations via Truncated Moment of the 1st Order Statistic

Let $Y_{1:n} \le Y_{2:n} \le, \ldots \le Y_{n:n}$ be the n order statistics from a continuous distribution function cdf $F(y)$ and its pdf $f(y) = F'(y)$. The characterization result have already been defined in Hamedani (2015) and Hamedani et al. (2017). We just state a short proof of it here for the sake of completeness.

Proposition 2

Let $Y\colon \zeta \to R$ be a continuous random variable with cdf F. Let $\zeta(y)$ and $Q(y)$ be two differentiable functions on $(0, \infty)$ such that

$$\lim_{n\to\infty} \zeta(y)[1 - F(y)]^{n} = 0, \qquad \int_0^{\infty} \frac{Q'(t)}{[Q(t) - \zeta(t)]}dt = \infty$$

Then

$$E[\zeta(Y_{1:n})|Y_{1:n} > t] = Q(t), t > 0 \tag{17}$$

implies

$$F(y) = 1 - exp\left\{-\int_0^y \frac{Q'(t)}{n\,[Q(t) - \zeta(t)]}dt\right\}$$

Proof:

If (17) holds, then taking integration by parts of (17) and the assumption

$$\lim_{y\to\infty} \zeta(y)[1 - F(y)]^n = 0,$$

We have,

$$\frac{\int_t^\infty \zeta(y)(f(y))}{(1-F(y))^n} dy = Q(t)$$

$$\zeta(t)(1-F(t))^n + \int_t^\infty \zeta'(y)(1-F(y))^n = Q(t)$$

$$\int_t^\infty \zeta'(y)(1-F(y))^n = [Q(t)-\zeta(t)]\,(1-F(t))^n$$

Differentiating above expression w.r.t 'y', and then rearrange the terms, we have

$$\frac{f(t)}{1-F(t)} = \frac{Q'(t)}{n[Q(t)-\zeta(t)]},\ 0 < y < \infty.$$

Integrating the last equation from 0 to y, we obtain, $\int_0^\infty \frac{Q\prime(t)}{[Q(t)-\zeta(t)]} dt = \infty$ a cdf F defined by (2).

Remark 2

Taking, for instance, $\zeta(y) = \left[\frac{(1-\tau)\left(1-\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)}\right)^{\kappa}\right]^{\delta}\right)\right)}{1-\tau\left(1-\left[1-\left(1-e^{-\left(\frac{\eta}{y}\right)}\right)^{\kappa}\right]^{\delta}\right)}\right]^n$ and $Q(y) = \frac{1}{2}\zeta(y)$, in Proposition (2), we attain (1).

3. CONCLUSION

In this article, we have considered the five-parameter Unimodal, Reverse-J shaped distribution presented by Rafique and Saud (2019), and provided its different characterization methods. We hope the results of the article will be convenient for the practitioners in different fields of statistics.

REFERENCES

1. Ahsanullah, M., Ghitany, M. and Al-Mutairi, D. (2017). Characterization of lindley distribution by truncated moments. *Communications in Statistics-Theory and Methods*, 46(12), 6222-6227.
2. Ahsanullah, M. and Shakil, M. (2020). Some characterizations of chaudhry and zubairs extended generalized inverse gaussian distribution. *International Journal of Computational and Theoretical Statistics*, 7(1), 15-24.

3. Ahsanullah, M., Shakil, M. and Kibria, B. (2016). Characterizations of continuous distributions by truncated moment. *Journal of Modern Applied Statistical Methods*, 15(1), 17-20.
4. Bhatti, F.A., Hamedani, G., Najibi, S.M. and Ahmad, M. (2019a). On the extended chen distribution: development, properties, characterizations and applications. *Annals of Data Science*, 8, 159-180.
5. Bhatti, F.A., Hamedani, G.G., Korkmaz, M.Ç. and Ahmad, M. (2019). On the Modified Burr XII-Power Distribution: Development, Properties, Characterizations and Applications. *Pakistan Journal of Statistics and Operation Research*, 15(1), 61-85.
6. Bhatti, F.A., Hamedani, G.G., Sheng, W. and Ahmad, M. (2018). On extended quadratic hazard rate distribution: development, properties, characterizations and applications. *Stochastics and Quality Control*, 33(1), 45-60.
7. Hamedani, G. (2015). Characterizations of NWP, ETGR and TWL Distributions. *Pakistan Journal of Statistics and Operation Research*, 11(4), 473-480.
8. Hamedani, G.G., Cordeiro, G.M., Lima, M. and Nascimento, A. (2017). Some extended classes of distributions: Characterizations and properties. *Pakistan Journal of Statistics and Operation Research*. 13(4), 893-908.
9. Nofal, Z.M. and Ahsanullah, M. (2019). A new extension of the fréchet distribution: Properties and its characterization. *Communications in Statistics-Theory and Methods*, 48(9), 2267–2285.
10. Rafique, A. and Saud, N. (2019). Exponentiated generalized fréchet geometric distribution. In *15th Islamic Countries Conference on Statistical Sciences (ICCS-15)*, 34, 189-198.
11. Shakil, M., Ahsanullah, M. and Kibria, B.G. (2018). On the characterizations of chens two-parameter exponential power life-testing distribution. *Journal of Statistical Theory and Applications*, 17(3), 393-407.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 305-315

TRENDS IN URBAN/RURAL AND REGIONAL FERTILITY DIFFERENTIALS IN PAKISTAN USING RELATIONAL GOMPERTZ MODEL

Adan Kamran and **Asifa Kamal**
Lahore College for Women University, Lahore, Pakistan
Email: adankamran85@gmail.com
asifa.k53@gmail.com

ABSTRACT

For developing countries like Pakistan, the direct methods' assumptions are not realistic and are directed to bias estimates of fertility. In current study, indirect method of fertility estimation i.e. Relational Gompertz model was used to overcome this problem. Urban/rural and regional differentials in fertility were estimated. It is concluded that direct method had under estimated TFRs for both urban/rural areas and regions of Pakistan. TFRs estimated from Relational Gompertz model were found higher as compared to the TFR from direct method. Decline in urban and rural fertility in all four surveys was observed. Fertility distribution was narrowed for both urban and rural areas and process of narrowing was more obvious for urban areas. It was also found that fertility distribution got narrowed with respect to time. The observed fertility distribution for Pakistan in all PDHS had shown slightly older and narrowed pattern as compared to the standard irrespective of urban/rural settings. The observed fertility distributions for all regions of Pakistan, in all PDHS was slightly older than the standard except for Baluchistan (PDHS 1990-91). Spread of the fertility distribution was slightly narrowed then standard for all waves of PDHS and for all regions except for Baluchistan (PDHS 2006-07).

1. INTRODUCTION

Fertility is an essential factor as life continuation depends on it. For a development project or any project related to family planning, policymakers need to analyze fertility. For under developing countries, analysis is essential to get control of the problem of rapid population growth. As Pakistan is a developing country, rapid population growth is a problem, leading the nation to economic and social disturbance (Pritchett, 1994).

In Pakistan, fertility rate estimates for demographic research are taken from annual censuses. Still, information on births and deaths, as given by the official registration system, has been incomplete in terms of availability and coverage (Afzal and Ahmed, 1974). Due to these problems, the demographic surveys and censuses that collect current fertility and lifetime births data are not reliable, and errors and omissions are expected. These errors are more frequent in the data obtained from rural areas as compared to urban areas. For developing countries like Pakistan, the direct methods' assumptions are not realistic and are directed to bias estimates of fertility rates. To reduce the risk of unreliable values and errors, indirect ways to measure fertility rates and trends are recommended and used. The reliability of these indirect methods depends on the data's assumptions and the methods used to calculate them. The Relational Gompertz model is an indirect method that is useful as it automatically adjusts the impact of errors and omissions. It removes the errors and omissions from the parities described by the low age group women (Famule, 1992).

It is observed that urbanization plays a vital role in fertility decline. The decline in fertility of urban women is more pronounced than her rural counterpart due to access to education, health facilities, and her participation in the labor market. Women's autonomy and exposure to mass media are more prevalent in urban societies (White, et al., 2008). Almost one-third of Pakistani women interviewed in PDHSs reside in the urban area. Pakistan is an agricultural country. The majority of the population lives in rural areas are engaged in the agriculture sector. Fertility intentions of rural residents are higher than urban as children mean an increase in the household labor force. Male dominance and son preference is also more common in rural areas of Pakistan. All these may result in higher fertility rates as compared to urban areas (Mughees, 2014).

Pakistan is a country with regional diversity in the shape of race, language, and culture. On an administrative basis, Pakistan is divided into four provinces, Federally Administrative Tribal Areas (FATA), the Gilgit Baltistan, and Azad Jammu Kashmir (AJK) regions. The four provinces show regional variations in the characteristics of the population. Alam and Shah (1986) documented that the fertility decline trend was found in Pakistan from 1960-1975. Hakim (1999) reported variability in fertility in the four provinces. Mean children ever born were found highest for NWFP, followed by Baluchistan, Punjab, and Sindh; when controlled for other factors, sequence slightly altered, i.e., fertility was found higher in Baluchistan, followed by NWFP, Sindh, and Punjab. United Nations (1987, page 28) also confirmed that fertility was observed lowest in Punjab, followed by Sindh and NWFP in the period 1960-1970.

In the current study, trends and urban/rural and regional differentials in fertility are investigated using four waves of the Pakistan Demographic and Health Survey.

2. METHODOLOGY

Source of Data

The current study used data from PDHS (1990-91, 2006-07, 20112-13, 2017-18). These surveys were conducted by DHS (Demographic and Health surveys); they collect the data on the birth histories of a woman. The unit of analysis for this was de facto women interviewed. This file is recoded as IR Individual (Women's) Recode file. IR Individual file contains all the data collected in the Women's questionnaire for de facto women with some variables from the household questionnaire. The total number of women interviewed in PDHS 1990-91, PDHS 2006-07, PDHS 2012-13 and PDHS 2017-18 were 6,611, 10,023, 13,558 and 15,068 respectively. MS Excel worksheet (Moultrie, 2013) was used for analysis.

Relational Gompertz Model

The Relational Gompertz model provides a framework to estimate the fertility rates. It defines the shape of fertility, which is usually sigmoid. It omits the chances of error, which occurs in the fertility data due to under-reporting or over-reporting of births.

Misreporting of lifetime births and misreporting of age by older women because of memory lapses also generates errors in fertility data that can be handled by indirect fertility methods (Avery, Clair, Levin & Hill, 2013).

The Relational Gompertz model is the modified form of the Brass P/F ratio. As the Brass P/F ratio method assumes constant fertility which is not fulfilled in the data obtained from the census or survey and the Relational Gompertz model does not require such an assumption. That is why the relational Gompertz model is used on survey or census data. Brass (1964) used the relational Gompertz model to estimate fertility rates using a fixed experimental change with age scale to represent ASFR's. Considering the age and fertility of women, constant Zaba (1981) used this model to adjust corrected fertility distribution.

This method is used to obtain the ASFR's and TFR's by regulating the shapes of the fertility obtained from direct methods by determining the level of these fertilities from the average parity (Moultrie, 2013).

The model overcomes the errors mostly found in fertility data. Misreporting the age of the older women or births as too few or too many in the recommended period and under-reporting of lifetime fertility are the fertility data errors (Moultrie, 2013).

The shape of the cumulated Gompertz distribution that is sigmoid is the property on which the Relational Gompertz model depends. The cumulated Gompertz distribution is:

$$G(x) = e^{ae^{b.x}}$$

The fertility and average parities of women are well demonstrated by the Gompertz distribution. The shape of cumulated Gompertz is sigmoid not a straight line so a double negative log transformation is used to linearize the distribution. This transformation approximates a straight line for all age groups. This transformation is known a Gompit (Moultrie, 2013). The gompit is:

$$Y(x) = -\ln\left(-\ln\left(G\left(x\right)\right)\right)$$

The linear function for Relational Gompertz model takes the observed data of fertility as a linear function of the gompit of standard fertility. The function is:

$$Y\left(x\right) = \alpha + \beta Y^{s}\left(x\right)$$

where, $Y^{s}\left(x\right)$ is the gompit of standard fertility schedules derived by Booth (1984). The parameter α is an index of location where the childbearing age differs from the standards. The parameter β is an inverse measure of spread and width of ASF distribution. It shows the spread of childbearing age differs from the standards (Moultrie, 2013).

The standards used for fertility were computed from a set of high fertility schedules obtained by the Coale and Trussell's model and are subjected to the constraints imposed by the model. These can be used in medium to high fertility population.

The standards used here are the modified form of Booth standards by Zaba (1981). Booth standards are not suitable for the younger age groups so Zaba modified Booth standards are used to obtain better fit the childbearing pattern of younger age group (Moultrie, 2013).

The standards by Zaba give identical results for un-shifted age and shifted age. Zaba (1981) obtained the values of $G\left(x+\frac{1}{2}\right)$, $G\left(x+\frac{3}{2}\right)$ by inserting the values of $G(x)$, $G(x+1), G(x+2)$. Still, the gompit transformation linearize $G(x)$ which is understandable to insert the gompits of $G(x)$ for half year age shift and demonstrate the value for $G\left(x+\frac{1}{2}\right)$, $G\left(x+\frac{3}{2}\right)$ by taking anti-gompit (Moultrie, 2013).

3. RESULTS

Current analysis showed difference in fertility due to place of residence and region, which gives insight into role of urbanization and regional differentials in declining fertility in Pakistan.

Urban/Rural Differentials in Fertility

In younger ages, current urban fertility was higher than completed fertility. Same was observed for rural fertility. It indicated age misreporting, omissions, and errors in fertility data for both urban and women.

Decline in ASFRs was consistent in all surveys for both urban and rural Pakistan (Figure 1). This decrease in ASFRs was almost observed in all age groups. However, pace of decline was not similar. It can be observed that decline in ASFRs for rural areas was more in younger and older ages between PDHS 1990-91 and PDHS 2006-07. For PDHS 2006-07 and PDHS 2012-13, ASFRs were almost same for younger ages, decline started after age 20. In urban areas decline was more noticeable between PDHS 1990-91 and PDHS 2006-07 for all ages but it was more pronounced after 24. Decline in ASFRs for urban areas was obvious from 20 to 34 between PDHS 2006-07 and PDHS 2012-13. For both of these PDHSASFRs for younger than 20 and older to 34 were almost same (both curves are almost overlapping). It can be seen from both average parity and ASFRs that fertility decline was faster in urban areas of Pakistan than rural areas.

It can also be observed from figure 1 that fertility distribution is narrowing down for both urban and rural areas and process of narrowing is more obvious for urban areas in all surveys except for PDHS 2012-13. For the PDHS 2012-13 the peak of the fertility distribution was for age group 20-24 and after that fertility started to decline, but this decline was not steep as compared to other survey decline pattern. It was slow over the age groups. Values of beta from Relational Gompertz model also depict same scenario. Most of child bearing is in younger age that is 20-29. Negative value of alpha from Relational Gompertz model also confirms it.

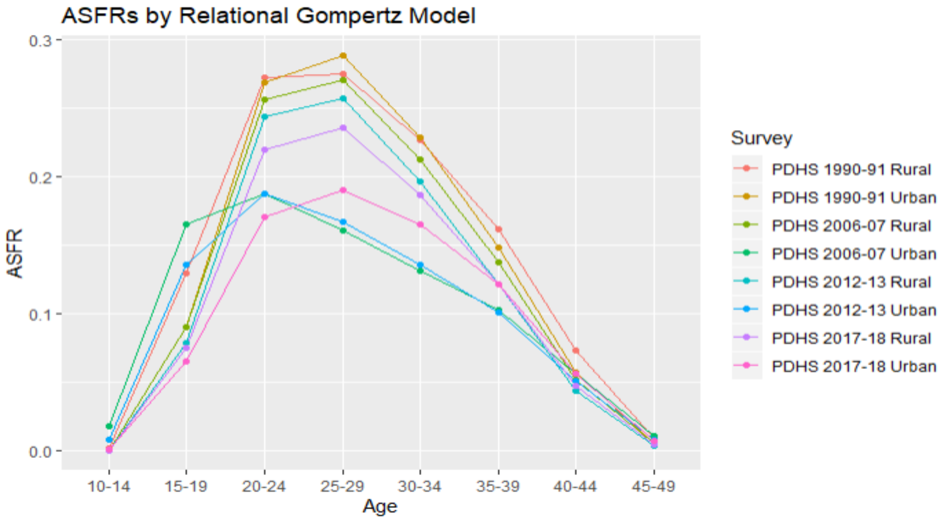


Figure 1: ASFRs for Urban and Rural using Relational Gompertz Model, PDHS 1990-91 to PDHS 2017-18

Average parity computed from Relational Gompertz model had shown that there was consistent decline in average parity for both rural and urban areas with respect to time (Figure 2). More significant decline can be observed in average parity for urban areas between PDHS 1990-91 and PDHS 2006-07. The decline was not more pronounced in average parity for rural areas for these surveys.

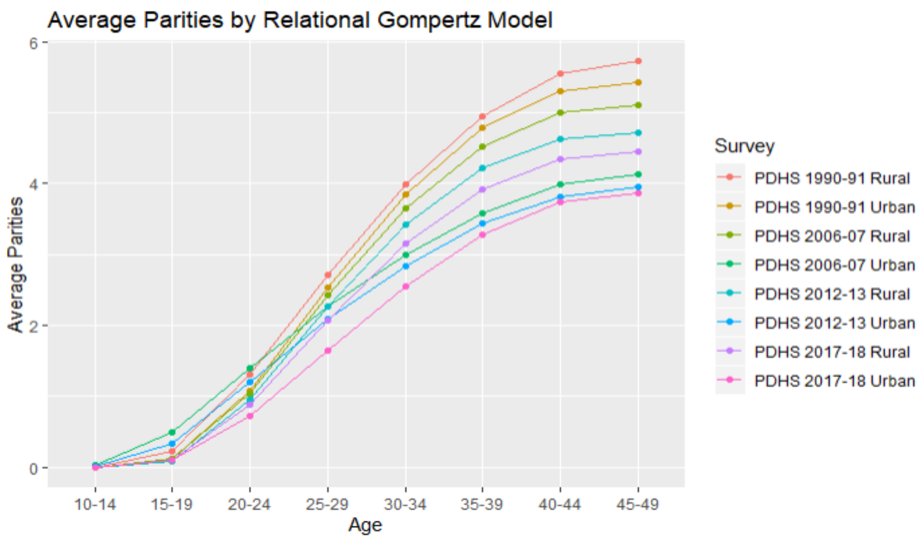


Figure 2: Average Parities for Urban and Rural using Relational Gompertz Model, PDHS 1990-91 to PDHS 2017-18

Estimates of TFRs using Relational Gompertz model for all surveys were higher than reported (Table 1) TFRs for both urban and rural Pakistan. All surveys showed that urban fertility was higher as compared to rural fertility. This difference became wider in later two surveys. Difference became more obvious for PDHS 2006-07, remained consistent for PDHS 2012-13, and had shown downward trend for both urban and rural fertility.

Table 1
TFR by Direct and Indirect Methods (Relational Gompertz Model) for Urban and Rural, PDHS 1990-91 to PDHS 2017-18

Survey	Urban		Rural	
	TFR Direct Method	TFR Relational Gompertz Model	TFR Direct Method	TFR Relational Gompertz Model
PDHS 1990-91	4.9	5.4	5.6	5.7
PDHS 2006-07	3.3	4.2	4.5	5.1
PDHS 2012-13	3.2	3.8	4.2	4.7
PDHS 2017-18	2.9	3.8	3.9	4.4

Parameter estimates for all surveys were within the range (Table 2). Value of alpha was negative for all surveys for both urban and rural areas, which had depicted older distribution of ages at child bearing than in standard. Value of alpha for urban areas was more than rural areas, which indicated increase in age at first birth in urban areas. Beta was the measure of spread of the fertility distribution. Since value was greater than "1" imply narrower distribution. Distribution of fertility was narrower for urban areas than rural areas in all surveys. Pattern of beta for both urban and rural areas was consistent except that difference was slightly narrowed down between urban and rural fertility distribution in PDHS 2012-13 as compared to other surveys. No obvious trend exist in the values of alpha and beta with respect to time.

Table 2
Values of Alpha and Betas for Different Surveys using Relational Gompertz Model (Urban and Rural)

Surveys	Urban		Rural	
	Alpha	Beta	Alpha	Beta
PDHS 1990-91	-0.1305	1.1273	-0.0708	1.0187
PDHS 2006-07	-0.1016	1.2494	-0.1087	1.1265
PDHS 2012-13	-0.1306	1.2383	-0.1029	1.1701
PDHS 2017-18	-0.2002	1.0164	-0.1279	1.1225

Regional Differentials in Fertility

Average parity computed from Relational Gompertz model had shown that the pattern was almost same for Sindh, KPK and Punjab for PDHS 2017-18 (Figure 3). Curves showing average parities almost overlapped before age 40 for the three regions. Little difference in average parity can be observed after age 40. Average parity for KPK was slightly higher as compared to Sindh and average parity for Sindh was slightly higher to Punjab after age 40. Average parity for Baluchi women was highest to all other three regions above age 24 and difference became widened as age increased.

Pattern of average parity computed from indirect method (Relational Gompertz Model) for PDHS 2006-07 was quite different from other PDHS. Before age 39, average parity of Baluchi women was lowest as compared to other regions. After 39, it was above Punjab and overlapped with Sindh. Average parity for Sindh and KPK was almost same for PDHS2006-07. After age 39, average parity of Punjabi women was lowest among all other regions. For PDHS 1990-91, average parity of Baluchistan was higher to other regions before age 34. After age 34, it had tended towards lower as compared to other regions. Average parity for Punjab, Sindh KPK was almost overlapping. In all surveys, parity became stable after age 39. Average parity of Blochi women was lowest in 1990-91, after age 39 it became highest in 2012-13 for all ages. Some trends in average parities were observed for the four regions in PDHS 2017-18 as it was observed in PDHS 2012-13.

Pattern of ASFRs across region was observed same in earlier two surveys. Later two surveys showed almost same pattern of ASFRs but quite different from earlier two surveys (Figure 4). In earlier two surveys, KPK had higher ASFRs while in later two surveys Baluchistan had shown higher ASFRs as compared to other regions. Decline in ASFRs was almost same for Punjab, Sindh and KPK before age 29 for PDHS 2012-13. After age 29, ASFR declined faster for Punjab followed by Sindh, Baluchistan and KPK. Before age 20, ASFRs for all regions were showing same overlapping pattern but after age 20, ASFRs for all regions indicated regional differences. ASFRs were higher for Baluchistan and KPK as compared to Punjab and KPK after age 20. Pace of decline was slow for KPK. Baluchistan had highest ASFRs for women below 34, after 34, it steadily start declining and even became lower to KPK for PDHS 2012-13.

ASFRs for Baluchistan was found lower to ASFRs of other regions before age 29, after age 29 it tend above all regions for PDHS 2006-07. Current fertility was slightly highest for Sindh, followed by Punjab, KPK and Baluchistan in the younger age groups. In older age groups, fertility declined more sharply for Punjab, followed by Sindh, KPK and Baluchistan. Pace of decline was slowest for Baluchistan. For PDHS 1990-91, current fertility pattern was same for Punjab, Sindh and KPK. Before age 24, Baluchi women had sharp increase in fertility, after age 24, there was sharp decline in fertility for Baluchi as compared to other regions.

It can be concluded from graphs of average parities and ASFRs for all regions and in all surveys followed more or less same pattern except Baluchistan. Pattern of fertility had changed drastically for Baluchistan, from one survey to another survey.

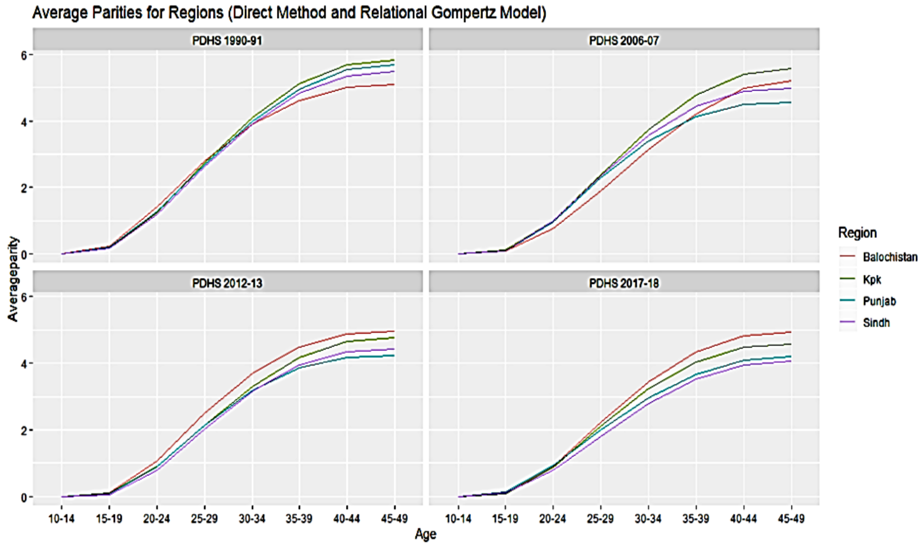


Figure 3: Average Parities for Regions using Relational Gompertz Model, PDHS 1990-91 to PDHS 2017-18

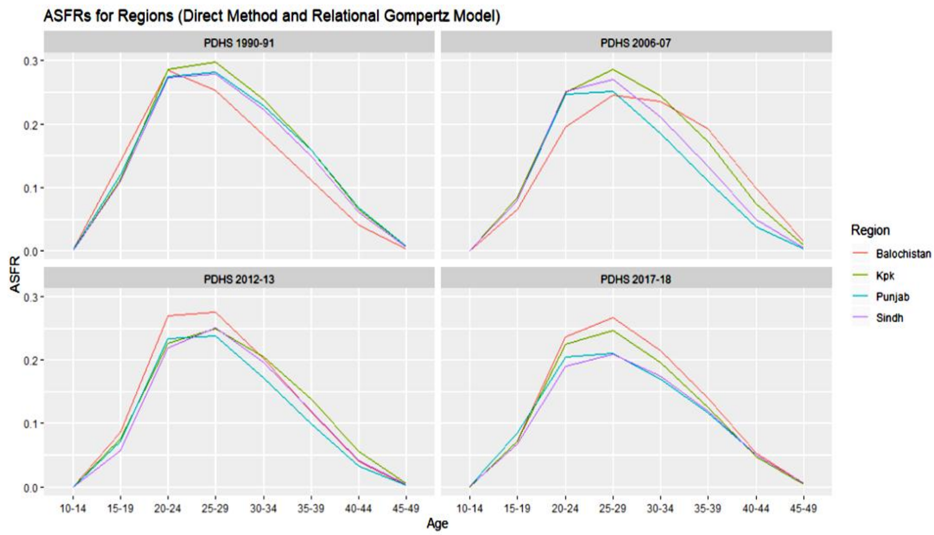


Figure 4: ASFR's for Regions using Relational Gompertz Model, PDHS 1990-91 to PDHS 2017-18

Difference can be observed in total fertility rates computed from direct method of ASFRs (reported in survey reports) and indirect method i.e. from Relational Gompertz model (Table 3) for different regions. Estimates of TFRs using Relational Gompertz model were higher than reported TFRs for different region of Pakistan for all surveys. Difference was more for KPK and Baluchistan for recent two surveys. Wide gap can be observed for observed and expected TFRs for Sindh on the basis of PDHS 1990-91.

Table 3
TFRs by Direct and Indirect Method Across Regions of Pakistan.

Survey	Punjab		Sindh	
	TFR Direct Method	TFR Indirect Method	TFR Direct Method	TFR Indirect Method
PDHS 1990-91	5.05	5.66	4.51	5.68
PDHS 2006-07	3.93	4.58	4.32	5.1
PDHS 2012-13	3.78	4.26	3.91	4.46
PDHS 2017-18	3.47	4.24	3.57	4.08
Survey	**KPK**		**Baluchistan**	
	TFR Direct Method	**TFR Indirect Method**	**TFR Direct Method**	**TFR Indirect Method**
PDHS 1990-91	5.13	5.53	4.63	5.14
PDHS 2006-07	4.34	5.36	4.09	5.23
PDHS 2012-13	3.92	4.63	4.23	5.01
PDHS 2017-18	3.76	4.21	3.85	4.93

Table 4
Values of Alpha and Betas (Relational Gompertz Model)
for Different Surveys (Regional Differentials)

Surveys	Punjab		Sindh		KPK		Baluchistan	
	ALPHA	BETA	ALPHA	BETA	ALPHA	BETA	ALPHA	BETA
PDHS 1990-91	-0.1095	1.0620	-0.0610	1.0538	-0.0281	1.1218	0.2368	1.0937
PDHS 2006-07	-0.1006	1.2052	-0.0574	1.1409	-0.1867	1.0602	-0.3800	0.9697
PDHS 2012-13	-0.0714	1.2452	-0.1629	1.1706	-0.1330	1.0607	-0.0770	1.2281
PDHS 2017-18	-0.0909	1.0607	-0.1703	1.0731	-0.1468	1.1370	-0.1804	1.1360

Value of alpha was negative for all regions and in all surveys except Baluchistan for PDHS 1990-91. Alpha was more negative for PDHS 2012-13 for Punjab and Baluchistan and for PDHS 2006-07 for Punjab and Sindh, it means child bearing in these regions, in older age was found higher as compared to other regions. For PDHS 1990-91, KPK had more child bearing at older age than other regions. Alpha was positive for PDHS 1990-91 for Baluchistan, which indicated that most of child bearing in this region was in younger age as compared to other regions during that period. Distribution of fertility was less narrowed for KPK as compared to other regions, which had shown that the childbearing was more in middle (25-29) age group and less in younger (15-19) and older (45-49) age groups for PDHS 2012-13. Almost consistent pattern as of PDHS 2012-13 can be observed for PDHS 2006-07 fertility distribution.

4. DISCUSSION

There is no doubt that fertility in Pakistan has started declining after 1990. However, this study is based on direct estimates of fertility, which are generally prone to error. It necessitates computing reliable estimates of the fertility level and trends using indirect techniques after adjusting data errors. This is helpful in assessing actual trend in fertility decline in Pakistan

Fertility in urban areas is found low as compared to women living in rural areas of Pakistan in all surveys. Education, good health facilities, exposure to media and women participation in labour market played vital role in the decline of fertility in urban areas (White, et al., 2008). Iqbal and Shah (1986) on the basis of 1972 survey also reported low fertility rates in urban areas as compared to rural areas. Hakim (1999) also found that fertility rates were lower of women living in urban areas as compared to women living in rural areas.

Regional disparity in fertility was also reported by Hakim (1999). It was found that fertility rates were higher for Baluchistan followed by KPK, Sindh and Punjab. Similar results were observed in current study. Baluchistan had highest fertility rate among all other regions of Pakistan. Iqbal and Shah (1986) on the basis of 1972 survey found that fertility rates were highest for Punjab followed by KPK, then Sindh and lowest fertility rate in Baluchistan.

5. CONCLUSION

It is concluded that direct method had under estimated TFRs for both urban/rural areas and regions of Pakistan. Decline in urban and rural fertility in all four surveys was observed. Fertility distribution was narrowed for both urban and rural areas and process of narrowing was more obvious for urban areas. It was also found that fertility distribution got narrowed with respect to time. The observed fertility distribution for Pakistan in all PDHS had shown slightly older and narrowed pattern as compared to the standard irrespective of urban/rural settings. The observed fertility distributions for all regions of Pakistan, in all PDHS was slightly older than the standard except for Baluchistan (PDHS 1990-91). Spread of the fertility distribution was slightly narrowed then standard for all waves of PDHS and for all regions except for Baluchistan (PDHS 2006-07).

REFERENCES

1. Afzal, M. and Ahmed, T. (1974). Limitations of Vital Registration System in Pakistan against Sample Population Estimation Project: A Case Study of Rawalpindi. *Pakistan Institute of Development Economics, Islamabad.* 13(3), 325-334.
2. Avery, C., St. Clair, T., Levin, M. and Hill, K. (2013). The ‘Own Children’ fertility estimation procedure: A reappraisal. *Population Studies*, 67(2), 171-183.
3. Brass, W. (1964). *Uses of census or survey data for the estimation of vital rates*. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa 14-19 December 1964. Document No.E/CN.14/CAS.4/V57. New York: United Nations.

4. Famule, F.D. (2005). Application of the Gompertz Relational Model to Re-Estimation of Fertility Rates in Osun State, Nigeria. *Multidisciplinary Journal of Research Development*, 6(3), 90.
5. Hakim, A. (1999). Demographic, socio-economic, and regional fertility differentials in Pakistan. *The Pakistan Development Review*, 643-660.
6. Iqbal, A. and Shah, N.M. (1986). Population Composition, Mortality, and Fertility. In Shah, N.M. (ed.) *Pakistani Women: A Socioeconomic and Demographic Profile*. Islamabad: Pakistan Institute of Development Economics, 53-86.
7. Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I M. and Zaba, B. (2013). *Tools for demographic estimation*. International Union for the Scientific Study of Population.
8. Mughees, S. (2014). Cubism in Pakistan: An evolutionary glance. *Sunway Academic Journal*, 10, 1-21.
9. NIPS and ICF (2019). *Pakistan Demographic and Health Survey 2017-18*. National Institute of Population Studies Islamabad, Pakistan and Rockville, Maryland, USA: NIPS and ICF.
10. NIPS and ICF (2013). *Pakistan Demographic and Health Survey 2012-13*. National Institute of Population Studies Islamabad, Pakistan and Calverton, Maryland, USA: NIPS and ICF International.
11. NIPS and Macro International Inc. (2008). *Pakistan Demographic and Health Survey 2006-07*. National Institute of Population Studies Islamabad, Pakistan and Macro International Inc.
12. NIPS and IRD/Macro International Inc. (1992). *Pakistan Demographic and Health Survey 1991-92*. National Institute of Population Studies Islamabad, Pakistan and Macro International Inc.
13. Pritchett, L.H. (1994). Desired fertility and the impact of population policies. *Population and Development Review*, 1-55.
14. United Nations (1987). *Levels and Trends of Fertility for Small Geographical Areas in Pakistan Using census Data*, Bangkok: ESCAP. (Asian Studies No. 62-1).
15. White, M.J., Muhidin, S., Andrzejewski, C., Tagoe, E., Knight, R. and Reed, H. (2008). Urbanization and fertility: An event-history analysis of coastal Ghana. *Demography*, 45(4), 803-816.
16. Zaba, B. (1981). Use of the Relational Gompertz Model in Analysing Fertility Data Collected in Retrospective Surveys. *Centre for Population Studies Research Paper*. 81-82. London: Centre for Population Studies, London School of Hygiene & Tropical Medicine.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan - February 18-20, 2021, Vol. 35, pp. 317-324

STATISTICAL ANALYSIS OF WIND SPEED DATA: A CASE STUDY FROM LAHORE, PAKISTAN

Naz Saud, Aqsa Rfique, Faiza Mushtaq and **Aqsa Riaz Khan**
Lahore College for Women University, Lahore, Pakistan
Email: naz.saud17@gmail.com
raqsa760@gmail.com
faizamushtaq912@gmail.com

ABSTRACT

The wind energy is the most appropriate and cheap source of energy for human beings which can be accessed from measurements of wind speed. It has long been a topic of interest in the field of meteorology in establishing a probability distribution that provides a good fit to wind speed data. Several studies has been conducted to analyze wind speed data. The present study comprises of obtaining the best fitted probability distribution to model wind speed data. For this purpose, average monthly wind speed data for 31 years (1988-2018) of Lahore, Pakistan is used. Weibull, Gamma, Rayleigh, Lognormal, Frechet and Burr Distributions are applied to wind speed data and parameters of these distributions are evaluated by using maximum likelihood method. Three Goodness of Fit tests i.e. Kolmogorove-Smirnove (KS), Anderson Darling (AD) and Cramer Von Mises (CVM) test are used to evaluate the performance of the distributions. Fitted PDFs, fitted CDFs, and PP plots are also presented to confirm the goodness of fit of these probability distributions. It can be observed from the results obtained that weibull distribution is the best fitted probability distribution for wind speed data.

KEYWORDS

Wind speed measurements, Probability density function, maximum likelihood method, goodness of fit tests.

1. INTRODUCTION

In present era, the whole world is striving to meet the energy crisis especially the third world countries by such natural resources which are not only easily available but also economical as well. Among these resources wind energy is considered to be the most economical one. As wind speed has many applications in the real world ranging from naval to aerospace and construction industry to growth of plants. In light of its importance many work had been done in past to model wind speed data which simply provides a statistical mechanism for forecasting as well.

Statistical parameters to express wind speeds distribution are very useful and considerable work has been carried out in past years, e.g., Justus (1976), Wentink (1976), Widger (1977) and Hennessey (1977). Widger (1976) developed a method of calculating wind power values from average wind speeds only, which was simple and appeared to have considerable value if generally valid. Akpinar and Akpinar (2004)used five years wind speed data based on hourly recorded observations to narrate the wind energy potential. Rayleigh and Weibull distributions were used for the sake of analysis and results revealed that Weibull distribution gives better fit than Rayleigh model. Xiao et al. (2006) concluded that three-parameter Weibull and Type-1 Weibull are appropriate probability models as compared with two-parameter Weibull model for extreme wind speed data recorded in Hong Kong. Zaharim et al. (2009) applied Burr, Frechet and Lognormal distribution to analyse the data of wind speed for a particular region in Pahang, Malaysia. After analysis, on the basis of goodness of fit tests and graphical results, Burr distribution was best fitted distribution for this data set. Furthermore, Wadagel et al. (2011) used five probability models namely, Lognormal, Uniform, Gamma, Weibull and Beta, to the given data and concluded that according to the results obtained from the comparison of Anderson Darling and Modified Anderson Darling Test, Weibull model is most appropriate. In addition, Kamran Abbas et al. (2012) used Gamma, Burr, Rayleigh, Frechet, Lognormal and Weibull for wind speed data and afterwards applied goodness of fit tests and narrated that Lognormal, Gamma and Burr were better distributions to model the data as compared to the rest of three distributions. Balpetek, N. et al. (2018) applied Weibull and Rayleigh distributions following chi-square, RMSE and R-square as measures of goodness of fit to identify the model for hourly wind speed data and concluded that Weibull distribution was more appropriate than Rayleigh. In accordance to the above context, this paper aims to identify the suitable probabilistic model for average monthly wind speed data.

2. MATERIALS AND METHODS

Data Source: The secondary data for the study was taken from meteorological observatory at Weather Radar Jail Road Lahore, Pakistan and is comprised of average monthly wind speed observations for 31 years (1988-2018) of Lahore.

Methodology: The data was analyzed to identify the distribution of average wind speed received in a month. At initial, the suitable distributions were selected by analyzing the frequency curve of data, which shows that positively skewed distributions are appropriate to be fitted in. These distributions were namely Weibull, Burr, Rayleigh, Gamma, Lognormal and Frechet. Afterwards, to obtain the best fitted model, goodness of fit tests and their respective pp-plots are used, whereas these goodness of fit tests are Kolmogorov Smirnov (KS) test, Cramer-von-Mises (CVM) test and Anderson Darling (AD) test. Finally, the best fitted model was obtained by evaluating the results obtained from above respective tests and plots. The probability density function (pdf) and cumulative distribution function (cdf) for the above mentioned distributions are presented below.

3. DISTRIBUTIONS AND GOODNESS OF FIT TESTS

Gamma Distribution:

The pdf and cdf of Gamma distribution are given below

$$f(t;\delta,\tau)\,\frac{t^{\delta-1}}{\tau^{\delta}\,\Gamma(\delta)}\,e^{-\left(\frac{t}{\tau}\right)}, t>0, \delta,\tau>0$$

and

$$F(t) = \frac{\Gamma(t/\tau)\delta}{\Gamma(\delta)}$$

where the shape parameter is γ and scale parameter is τ.

Weibull Distribution:

The given function is the cdf of Weibull distribution.

$$F(t;\ \delta, \tau) = 1 - e^{-\left(\frac{t}{\tau}\right)^{\delta}}$$

where shape parameter is δ and scale parameter is τ.

The corresponding pdf is defined as

$$f(t;\ \delta, \tau) = \frac{\delta}{\tau}\left(\frac{t}{\tau}\right)^{\delta-1} e^{-\left(\frac{t}{\tau}\right)^{\delta}}, t > 0\ , \delta, \tau > 0$$

Lognormal Distribution:

The pdf of Lognormal distribution is given by:

$$\mathrm{f}(\mathrm{t}; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{ln\, t - \mu}{\sigma}\right)^2\right]\ t > 0, \sigma > 0, -\infty < y < \infty.$$

The corresponding cdf of is

$$\mathrm{F(t)} = \emptyset\left(\frac{\ln t - \mu}{\sigma}\right)$$

where ∅ represent Laplace integral.

Rayleigh Distribution:

The pdf of Rayleigh distribution with parameter σ is:

$$\mathrm{f}\ (\mathrm{t}; \sigma) == \frac{t}{\sigma^2} \mathrm{e}^{\left[-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right]}\ t > 0, \gamma, \sigma > 0,$$

and cdf is

$$F(t) = 1 - \mathrm{e}^{\left[-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right]}$$

where σ represent scale parameter.

Burr Distribution:

The Burr distribution has following pdf and cdf:

$$f(t; \sigma, k, \tau) = \frac{\sigma k \left(\frac{t}{\tau}\right)^{\delta-1}}{\tau\left(1 + \left(\frac{t}{\tau}\right)^{\sigma}\right)^{k+1}}\ t > 0, \sigma, \kappa, \tau > 0$$

$$F(t;\sigma,k,\tau) = 1-\left(1+\left(\frac{t}{\tau}\right)^{\sigma}\right)^{-k}$$

Frechet Distribution:

The pdf and cdf of Frechet distribution is obtained by

$$f(t;\tau,\delta) = \frac{\tau}{\delta}\left(\frac{\delta}{t}\right)^{\tau+1} e^{-\left(\frac{\delta}{t}\right)^{\tau}} \; t>0,\delta,\tau>0$$

$$F(t;\tau,\delta) = e^{-\left(\frac{\delta}{t}\right)^{\tau}}$$

Goodness of Fit (GOF):

It measures that how closely a random sample follows a theoretical probability distribution.

These tests are applied for testing the following null hypothesis:

H_0 : Average monthly wind speed data observes the given distribution.

H_1 : Average monthly wind speed data do not observes the given distribution.

In this study, we applied three GOF test at five percent level of significance to check that the GOF for the fitted distribution in which 'T' denote the wind speed random variable and 'n' represent the sample size. The GOF tests are defined below.

(i) Kolmogorov Smirnov Test:

The most popular nonparametric test named as Kolmogorov-Smirnov (KS)test was suggested by Massey Jr (1951). This test is based on the difference between the hypothesized distribution function and empirical cumulative distribution function (ECDF). It is represented by K.

K = max (K^{+}, K^{-})

where

$$K^{+} = \max\left(\frac{j}{n} - F(t_j)\right)$$

and

$$K^{-} = \max\left(F(t_j) - \frac{j-1}{n}\right)$$

(ii) Anderson Darling Test:

Anderson and Darling (1954) introduced an empirical distribution function based goodness of fit test named as Anderson Darling (AD) statistic. It compares the better fit between observed and expected cumulative distribution functions.

$$A^2 = n - \frac{1}{n}\sum_{k=1}^{n}(2k-1)\left[lnF(t_k) + ln\left(1 - F(T_{n-k+1})\right)\right]$$

and it also give more weight to the tails than the Kolmogorov Smirnov test.

(iii)Cramer-von Miser Test

It uses sum squared differences between observed and expected cumulative proportion as test statistic.

$$T = \frac{1}{12n} + \sum \left[\frac{2j-1}{2n} - F(t_j)\right]^2$$

4. ANALYSIS AND RESULTS

Table 1
Descriptive Statistics of Wind Speed

Year	N	Mean	SD	Skewness	Kurtosis	Min	Max
1988-2018	358	1.328	0.9500	0.9333	2.4982	0.1	4

The descriptive statistic of wind speed data is discussed in table 1. Since the measure of skewness is 0.9333 which indicates that the data is positively skewed and, in this context, right skewed distributions can be applicable to this respective dataset.

Now proceeding further, Lognormal, Gamma, Rayleigh, Weibull, Frechet and Burr distributions are fitted to wind speed data along with the estimation of parameters by Maximum Likelihood Method.

Table 2
Parameter Estimates

S#	Distribution	Parameters	Estimates
1	Weibull	δ T	1.1145 1.3704
2	Burr	σ k τ	1.3893 4010.7 570.16
3	Frechet	δ τ	0.97797 0.57564
4	Gamma	γ τ	1.8112 0.70587
5	Lognormal	μ σ	0.93372 -0.05643
6	Rayleigh	σ	1.1475

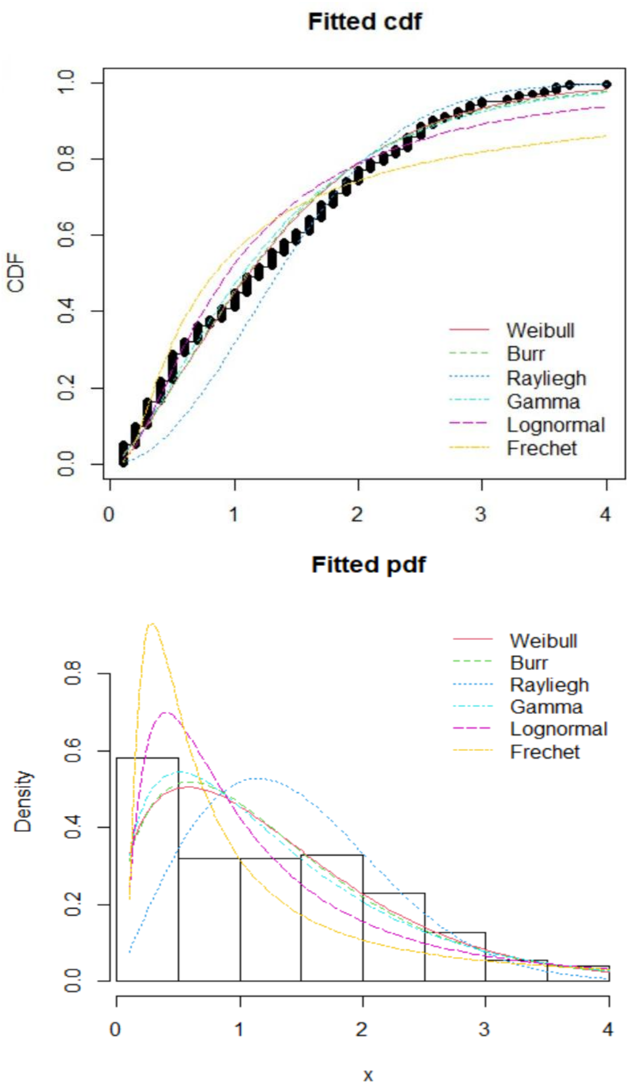


Figure 1: Fitted CDF and Fitted PDF

The above figure shows the fitted CDF. It can be seen that Burr and Weibull distribution are much closer to empirical distribution and also indicate that the following two distributions are appropriate for the given data, whereas the rest of the four distributions show deviation from the empirical distribution.

Table 3
Goodness of Fit Tests

Distribution	KS (p-value)	AD (p-value)	CVM (p-value)	AIC	BIC
Weibull	0.087346 (0.008)	2.8034 (0.0345)	0.44551 (0.0548)	872.3516	880.1127
Lognormal	0.1148 (0)	7.2846 (0.0002)	71.1801 (0)	930.566	938.2177
Rayleigh	0.19994 (0)	27.451 (0)	3.093 (0)	955.4722	959.3527
Gamma	0.0842 (0.0124)	3.2803 (0.0197)	0.54192 (0.0311)	880.1716	887.9327
Burr	0.088862 (0.0078)	3.2497 (0.0205)	0.5079 (0.0379)	881.2648	892.9064
Frechet	0.1478 (0)	14.822 (0)	2.3356 (0)	1042.748	1050.509

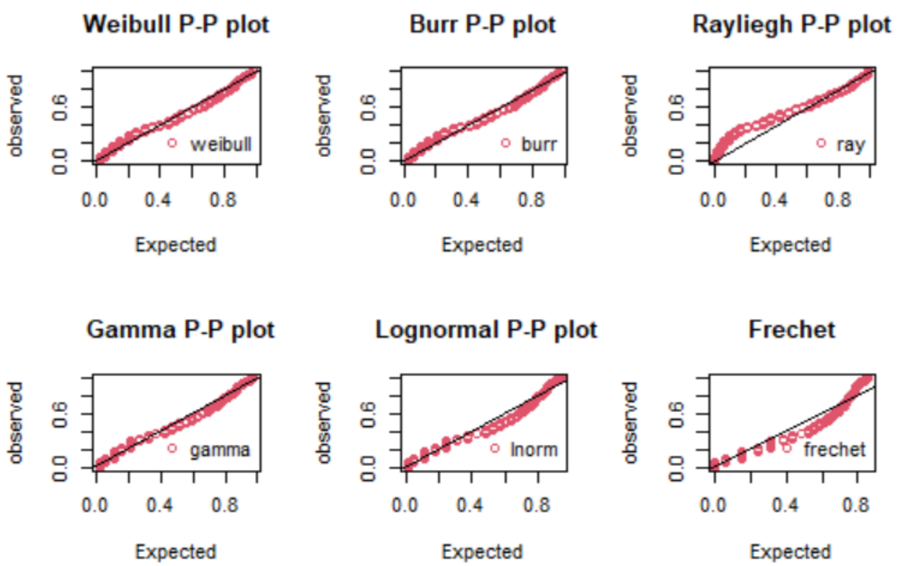


Figure 2: PP-Plots

From the above results since all the p-values are less than 0.05 indicates that all the distributions are good fit for data while the fitted pdf and PP-plots indicate that Burr and Weibull distributions are good fit however, consideration of AIC and BIC criteria Weibull distribution provides the best fitted model for average monthly wind speed data.

5. CONCLUSION

Keeping in view, the importance of wind energy and its enormous applications in the real world, multiple probability distributions can be fitted to this sort of data in order to estimate and forecast the patterns and average wind speed in the future. In context of the analysis performed in this study, among six statistical probability distributions, Weibull distribution provides the good fit to model average monthly wind speed data.

REFERENCES

1. Abbas, K., Alamgir, K.S., Ali, A., Khan, D.M. and Khalil, U. (2012). Statistical analysis of wind speed data in Pakistan. *World Applied Sciences Journal*, 18(11), 1533-1539.
2. Akpinar, E.K. and Akpinar, S. (2004). Statistical analysis of wind energy potential on the basis of the Weibull and Rayleigh distributions for Agin-Elazig, Turkey. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 218(8), 557-565.
3. Anderson, T.W. and Darling, D.A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268), 765-769.
4. Balpetek, N. and Akpınar, E.K. (2018). Weibull ve Rayleigh dağılımlarına göre Elazığ ilinin rüzgâr enerjisi potansiyelinin istatiksel analizi. *Journal of the Faculty of Engineering & Architecture of Gazi University*, 2018, 1-18.
5. Hennessey Jr, J.P. (1977). Some aspects of wind power statistics. *Journal of Applied Meteorology and Climatology*, 16(2), 119-128.
6. Justus, C.G. and Mikhail, A. (1976). Height variation of wind speed and wind distributions statistics. *Geophysical Research Letters*, 3(5), 261-264.
7. Massey Jr, F.J. (1951). The kolmogoro/v-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253), 68-78.
8. Wadagale, A.V., Thatkar, P.V., Dase, R.K. and Tandale, D.V. (2011). Modified Anderson Darling test for wind speed data. *International Journal of Computer Science & Emerging Technologies*, 2(2), 249-251.
9. Wentink, T. (1976). Study of Alaskan wind power potential and its possible applications. *Final Report (Rep. No. NSF/RANN/SE/AER 74-00239/FR-76/1),* Geophysical Institute, University of Alaska.
10. Widger Jr, W.K. (1976). Estimating wind power feasibility. *Power Engineering*, 80, 58-61.
11. Widger Jr, W.K. (1977). Estimations of wind speed frequency distributions using only the monthly average and fastest mile data. *Journal of Applied Meteorology*, 16(3), 244-247.
12. Xiao, Y.Q., Li, Q.S., Li, Z.N., Chow, Y.W. and Li, G.Q. (2006). Probability distributions of extreme wind speed and its occurrence interval. *Engineering Structures*, 28(8), 1173-1181.
13. Zaharim, A., Razali, A.M., Abidin, R.Z. and Sopian, K. (2009). Fitting of statistical distributions to wind speed data in Malaysia. *European Journal of Scientific Research*, 26(1), 6-12.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 325-330

COMPARATIVE STUDY OF PREDICTIVE STRENGTH OF CLASSICAL RISK FACTORS OF CORONARY ARTERY DISEASE IN LAHORE, PAKISTAN

Anmol Tahir[1], **Yasmin Zia**[2] and **Asma Zeb**[1]
[1] Lahore College for Women University Lahore, Pakistan
[2] Govt. Post Graduate College (W) Samanabad, Lahore, Pakistan
Email: anmoltahir97@gmail.com
yasminzia457@gmail.com
asma_zeb13@yahoo.com

ABSTRACT

The main aim of this research was to take a perspective over the factors which are significantly responsible for Coronary Artery Disease (CAD) and to model them to estimate the chance of Coronary Artery Disease in the presence or absence of these risk factors. For the fulfillment of the objectives, a hospital-based cross-sectional study was planned to include both descriptive and analytical components at two randomly selected public sector hospitals of Lahore. A sample of 200 patients, who come for their first angiography, was selected by a face-to-face survey method. Data included 100 cases and 100 control aged 20 years and above. In this study 100(50%) males and 100(50%), females were collected by a well-designed questionnaire. Data collected through 20 risk factors age, income, gender, marital status, obesity, blood pressure, diabetes, high cholesterol, family history, eating habits, sleeping habits, anxiety, blood clotting, smoking, lifestyle, drug addict, depression, area, daily exercise, physical inactivity. To test the main effects and the interaction effects of risk factors, multiple logistic regression was used for overall risk factors and odds ratios for each risk factor were calculated. It was found that area, exercise, income, depression, blood clotting, high cholesterol, sleeping habits were the significant risk factors of coronary artery disease. These factors play a major role in the development of the disease.

1. INTRODUCTION

The heart is a muscular organ in humans which pumps blood through the blood vessels of the circulatory system. Blood provides the body with oxygen and nutrients, as well as assists in the removal of metabolic wastes. In humans, the heart is located between the lungs, in the middle compartment of the chest.

Heart disease also known as "CARDIOVASCULAR DISEASE" (CVD), refers to the blockage of blood vessels and cardiovascular system and this condition generally leads the heart attacks and heart failures which can be harmful if not treated immediately. Coronary artery disease is the narrowing or blockage of the coronary arteries, usually caused by atherosclerosis. Atherosclerosis (sometimes called "hardening" or "clogging" of the arteries) is the buildup of cholesterol and fatty deposits (called plaques) on the inner walls of the arteries.

Pakistan is situated in the South Asia region and has the highest rate of coronary artery disease prevalence. Khan et al. (2007) revealed that approximately 10 million people suffered from acute myocardial infarction (heart attack) in Pakistan. It was also concluded that there is more chance of developing coronary artery disease in men than women at the early age of the Pakistan population. Iqbal et al. (2006) indicated that 16% of heart attack patients were the age of 45 or less and 93% of them were men. Cardiovascular disease is responsible for 50% of adults death all over the world and most of the deaths were due to coronary artery disease. The people of subcontinent Asia who lived in the UK had 50% more mortality from coronary artery disease than from UK citizens. Ardeshna et al. (2018) revealed that mortality rate due to CAD in Pakistan reached 196258 or 15.36% of total deaths. The main objective of the study is to determine the risk factors which are significant to the problem of CAD in Lahore.

2. MATERIALS AND METHODS

The sample was selected from Punjab Institute of Cardiology (PIC) Lahore and Sheikh Zayed hospital Lahore. A sample of 200 patients both males and females coming to the Angiography ward was taken. Data was collected by frequent visits to the hospitals. By using the case to control the ratio of 1:1, 100 patients with the age of 20 years or above were selected as cases and 100 patients as control from two hospitals.

A questionnaire designed for this study was used to collect the required information from patients. The questions used in this study comprises the possible risk factors relating to CAD by with the helped of given literature. Direct personal interviewing was adopted for data collection. The most important phase of research is data analysis after the collection of data. All the variables were pre-coded for computer analysis and all the records were rechecked to ensure correct data entry. Statistical analysis has been performed using SPSS 22.0. Descriptive analysis of various risk factors with CAD is constructed and interpreted. In the analytical section, bivariate analysis was performed. The binary logistic regression model was fitted using all risk factors to determine the effect of each risk factor on CAD.

3. RESULTS

Through descriptive analysis, it was obtained frequencies and percentages of each risk factor of Coronary Artery Disease. A total number of 200 patients were included in this study, out of which 100(50%) were cases and 100(50%) were controlled. There are 100 (50%) males and 100(50%) females. The result of bivariate analysisis shown in Table 1.

Table 1
Bivariate Analysis

Predictor	Chi-Square	P-Value	Predictor	Chi-Square	P-Value
Age	3.923	0.270	drug addict	0.177	0.674
Gender	0.8	0.777	sleeping habit	10.332	0.006
Area	1.447	0.229	daily exercise	10.632	0.001
Education	2.067	0.151	Depression	5.12	0.024
Obesity	0.083	0.774	eating habit	7.771	0.049
Marital Status	2.571	0.463	high cholesterol	9.288	0.030
Lifestyle	8.68	0.013	Diabetes	0.368	0.544
Income	7.959	0.035	high blood pressure	9.582	0.010
family history of heart disease	0.082	0.775	Anxiety	0.781	0.377
physical inactive	2.02	0.155	blood clotting	53.056	0.000
Smoking	5.650	0.002			

The risk factors lifestyle, income, smoking, sleeping habit, daily exercise, depression, eating habits, high cholesterol, high blood pressure, blood clotting were found to be statistically significantly associated with Coronary Artery Disease.

Table 5
Model Coefficients, Odds Ratios & Confidence Intervals for Odds Ratio

Factors	B	S.E.	Wald	Sig.	Exp (B)	95% C.I. for EXP(B)	
						Lower	Upper
Area	1.106	.443	6.226	.013	3.021	1.268	7.202
Income	-.647	.265	5.947	.015	.524	.311	.881
Sleeping Habit	1.002	.282	12.665	.000	2.724	1.569	4.731
Daily Exercise	- 1.272	.411	9.606	.002	0.280	0.125	0.626
Depression(1)	.894	.409	4.776	.029	2.446	1.097	5.456
High Cholesterol(1)	1.188	.536	4.919	.027	3.280	1.148	9.370
Blood Clotting	3.243	.517	39.341	.000	25.623	9.299	70.599
Constant	-4.337	.755	33.016	.000	.013		

The predictive logistic regression model is:

$$\begin{aligned} CAD = & -4.337 + 1.106\ \text{Area} - 0.647\ \text{İncome} + 1.002\ \text{Sleeping Habit} \\ & + 1.272\ \text{Daily Exercise (1)} + 0.894\ \text{Depression (1)} \\ & + 1.188\ \text{High Cholesterol (1)} + 3.243\ \text{Blood Clotting} \end{aligned}$$

As shown in the above table, the risk factors area, income, sleeping habit, daily exercise, depression, high cholesterol, and blood clotting were found to be significant against CAD as their p-value is < 0.05

Since the coefficients of the area were found to be 1.106 indicating a direct relationship between area and CAD. The odds ratio was 3.021, indicating that the patients of urban areas were a 3.021 times more chance of having CAD than rural areas.

The odds ratio for income was found to be 0.524, indicating that there was less chance of not having CAD in patients with high income as compared to those patients who had a low income. The patients who have high income (1-0.524) = 0.476 = 47.6% protection against CAD. The 95% confidence did not include 1 (0.311, 0.881) indicating that income had a statistically significant negative association with CAD.

The odds ratio for sleeping habits was found to be 2.724 indicating that with an increased in sleeping hours there was an increased risk of 2.724 times of CAD in patients. Daily exercise is found to be protective against the disease. Its coefficients were negative. Odds ratio is 0.280 and 95% confidence did not include 1 (0.125, 0.626). Risk of CAD could decreased by (1- 0.280 = 0.72) 72% if patients do exercise. Exercise is an essential factor that affects the presence and absence of the disease.

The odds ratio for depression was found to be 2.446 and 95% confidence interval (1.097, 5.456) which means that patients with depression have 2.446 times more chance of having CAD than patients without depression keeping all other factors constants.

The odds ratio for high cholesterol was found to be 3.280 and 95% confidence interval (1.148, 9.370) indicating that the patients with high cholesterol have 3.280 times more chance of having Coronary Artery Disease than patients without high cholesterol keeping all other factors constants.

The odds ratio for high cholesterol was found to be 25.623 and 95% confidence interval (9.299, 70.599) which means that patients who have blood clotting 25.623 more chance of having Coronary Artery Disease than patients who do not have blood clotting keeping all other factors constants.

4. COMMENTS AND CONCLUSION

A hospital-based comprehensive study was designed that includes 200 patients, who underwent treatment for their heart at two randomly selected public sector hospitals of Lahore named Punjab Institute of Cardiology(PIC), Sheikh Zaid hospital. The scope of the study is to take a perspective over the factors which are significantly responsible for CAD. Out of 200 patients, 100 are diagnosed as having CAD and therefore treated as cases whereas the remaining 100 are founded to have not CAD treated as controls.

To determine the predictive strength of various risk factors and to predict the chance of suffering from Coronary Artery disease based on the presence of risk factors, a multiple logistic regression model is fitted. The purpose was to estimate the extent to which a factor is contributing towards CAD, odds ratio, confidence intervals for the odds ratio were also found. Age, area, marital status, gender, diabetes, blood pressure, cholesterol, obesity, lifestyle, income, depression, anxiety, sleeping habit, eating habit, drug addict,

family history, smoking, daily exercise, education, physically inactive, and blood clotting are included in the overall model as significant risk factors.

The model showed that seven factors are playing a significant role in the presence or absence of the disease. These were area, sleeping habit, blood clotting, income, daily exercise, depression, and high cholesterol. Area, sleeping habit, exercise, depression, cholesterol, blood clotting were positively associated while income is negatively associated. The result produced in this study is according to medical science and also literature review.

The results and conclusion of this study lead to several recommendations, noted below. The results suggest that major as well as minor risk factors contribute to the incidence of coronary artery disease. In Pakistan, risk factors are different as compared to the western countries due to differences in diet and social life. Such as risk factor drinking alcohol are not significant in Pakistan as well as smoking in females are not prevalence in Pakistan.

As cholesterol is one of the most significant risk factors so patients must be checked the cholesterol on a regular and precautionary should be adopted to control the blood cholesterol level.

High blood pressure is also a major risk factor so the patient must be checked and control the blood pressure level. Diabetes is also one of the most significant risk factors so the patient must be checked and control the blood sugar level.

Extra care should be done if the person has a family history of Coronary Artery disease both in male and female. Avoidance of smoking, drug addiction, control hypertension, careful in eating and sleeping habits, and an active lifestyle should be aimed at the prevention of coronary artery disease events thus reducing mortality.

5. ACKNOWLEDGEMENT

All praise for almighty Allah, who guide us in the darkness and all respects are for his Holy Prophet Muhammad (P.B.U.H) who enables us to recognize our creator.

I would like to say thanks and to present my heartiest gratitude to my Lovıng Parents, grandparents (late). They always gave me strength whenever i lost my courage especially in the field of my studies. Without their encouragement it would not have been possible for me to complete this research.

I feel highly obligated and want to express my cordial and heartiest gratitude to my respected supervisors Ms Yasmın Zıa and Ms Asma Zeb for her advice, help, guidance, suggestion and motivating criticism throughout my research work. I will always begrateful to her.

REFERENCES

1. Ardeshna, D.R., Bob-Manuel, T., Nanda, A., Sharma, A., Skelton IV, W.P., Skelton, M. and Khouzam, R.N. (2018). Asian-Indians: a review of coronary artery disease in this understudied cohort in the United States. *Annals of Translational Medicine*, 6(1), 12.
2. Iqbal, M.P. (2006). Hyper homocysteinemia and coronary artery disease in Pakistan. *Journal of Pakistan Medical Association*, 56(6), 282-285.
3. Khan, M.S., Jafary, F.H., Faruqui, A.M., Rasool, S.I., Hatcher, J., Chaturvedi, N. and Jafar, T.H. (2007). High prevalence of lack of knowledge of symptoms of acute myocardial infarction inPakistan and its contribution to delayed presentationto the hospital. *BMC Public Health*, 7(1), 1-8.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 331-334

EXPORTS AND IMPORTS OF PAKISTAN IN TEXTILE HOSEPIPING, SIMILAR TEXTILE TUBING, WITH OR WITHOUT LINING, ARMOUR OR ACCESSORIES OF OTHER MATERIALS (2010-2019)

Muhammad Ali[1], Ibrahim[1] and **Danish Hassan[2§]**

[1] Department of Textile & Clothing, National Textile University, Karachi, Pakistan
Email: ali.ikhtiar406@gmail.com; ibrahim.khan9718@gmail.com

[2] Department of Applied Sciences, National Textile University Karachi, Pakistan. Email: danishhassan@ntu.edu.pk

[§] Corresponding author

ABSTRACT

Pakistan is an agricultural country and a great producer of cotton among the different countries. The exports and imports of Pakistan in textile tubing, hosepiping and other such textile materials have been discussed in this paper using of statistical tools like central tendency and coefficient of variation.

The results show the imports of Pakistan in textile hosepiping, armor and other materials for past ten years (2010-2019) was found greater than exports and the imports are approximately normal distributed year by year as compared to exports.

1. INTRODUCTION

Textile hosepiping is the flexible piping in which water or other materials can be transferred through it. It can be covered with textile material with different weave. Fabric hose is essentially circulated over pipe.

Textile tubing is the tube cover with textile materials. Armor and other accessories are also made from textile materials [1].

There is a need of use latest marketing strategies for the production along with the quality of work produce [3].

2. METHODOLOGY

There are three methods that we have used in this research work for the last 10 years export and import time series data. The time series data is collected from [5].

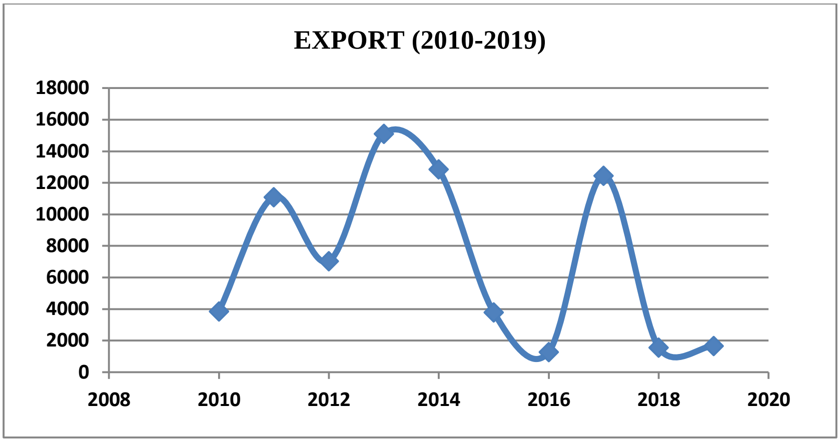
- Central Tendency
- Coefficent of Variation (CV)
- Visual Statistics

The central tendency shows the normality of the any data. In this technique we can observe that the time series data how much data is normal about its mean. Our results indicated the import is more normal than the export.

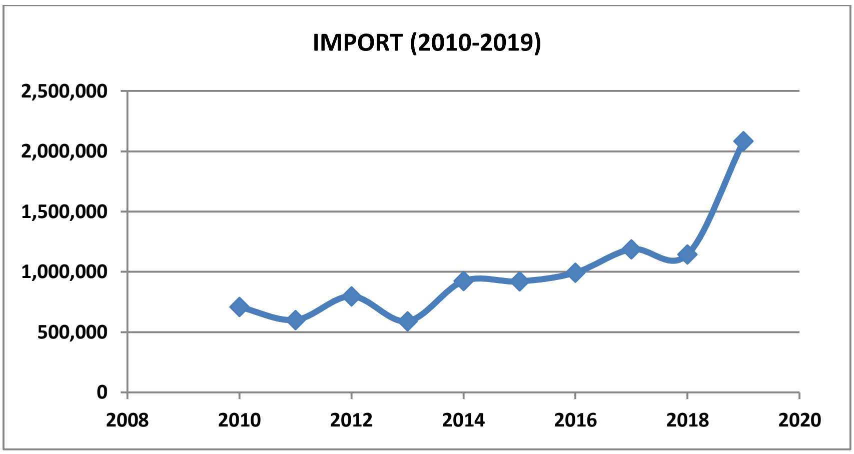
The coefficient of variation (CV) is the statistical technique that represents the reliability of the data sets. Decreasing the coefficient of variation increasing will be the reliability. The CV of import is observed less than the export and hence reflects the reliability that is expected to continue to the long term [4].

Visual statistics is the comparison of two data sets (Export & Import) graphically. The line graph of import exhibiting the increasing trend and export depict the decrease with high fluctuation [4].

The visual statistics are as under:



(a)



(b)

Figure 1: (a) Indicates the High Variation in Textile Related Export and (b) Represents the Increasing of Textile Related Import.

In both the figures, (a) indicates the decrease in export whereas (b) depicts the increasing of import in the last 10 years of textile related product (mentioned in sec.1). The graph of import shows that it is expected to continue in future.

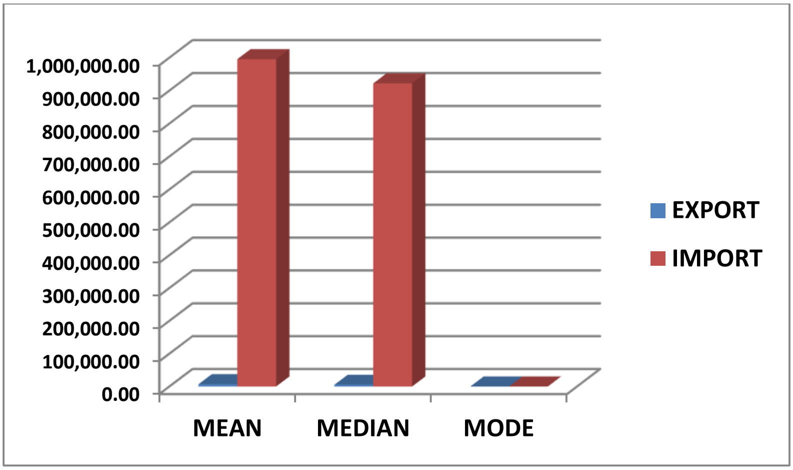


Figure 2: The Figure Depicts the Central Tendency of Export and Import (2010-2019)

The central tendency of import is looking approximately normal as compare to the export. This kind of behavior of import is expected to continue in future of textile related production.

3. COEFFICIENT OF VARIATION (CV)

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. Higher the coefficient of variation greater will be the level of dispersion around the mean. It is generally expressed as a percentage. The lower the value of the coefficient of variation, the more smooth the estimate [4].

$$\textbf{Coefficient of variation } (\mathbf{CV}) = \frac{\boldsymbol{\sigma}}{\boldsymbol{\mu}} \times \mathbf{100} \tag{1}$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x - \mu)^2}{n}} \tag{3}$$

Our results shows that CV of import ($CV_{IMPORT} = 41\%$) is less than CV of export ($CV_{EXPORT} = 72\%$) of last 10 years selected time series data. So it means that the import is normally distributed and export is highly variation and found to be less reliable and smooth as compare to import.

$CV_{IMPORT} < CV_{EXPORT}$.

4. RESULT AND DISCUSSION

By comparing exports and imports values in US ($) of Pakistan textile hosepiping, lining, armor, and accessories of other materials with respect to central tendency method there is a major difference in export and import value, so it's means that Pakistan import value in US ($) is greater than export according to the previous 10 years' data, so we can see that Pakistan import in textile hosepiping, lining, armor and accessories of other materials are much more as compared to Pakistan export.

The imports values of central tendency are normal distributed as compared to exports values and the value of coefficient of variation of imports is less than the value of exports.

The increasing of import from export reflects the negative balance of textile related trade which is a trade deficit of Pakistan.

5. COMMENTS AND CONCLUSION

Pakistan needs to increase in the quality production of textile hopepiping, tubing and other textile related materials. We conclude from the results obtained that Pakistan exports and imports in textile hosepiping, textile tubing, armour or accessories of other materials were compared by using a statistical methods and found negative balance in trade. Textile industry of Pakistan needs to produce quality products to balance the import and export.

6. ACKNOWLEDGEMENT

All praises and thanks are to almighty Allah the most merciful, most beneficent. The source of knowledge and wisdom endowed to main kind, those blessings enabled us to complete this article. We offer our humble thanks from the core of our heart to the Holy Prophet MUHAMMAD ﷺ who is forever model of guidance and knowledge for humanity.

We are thankful to our honorable Dr. Danish Hassan, Asst. Prof. (Mathematics) National Textile University Karachi Campus. Those healthy discussions encouraged us to take up this article and to work on this.

REFERENCES

1. Bacchetta, M. and Bora, B. (2003). *Industrial tariff liberalization and the Doha development agenda (No. 1).* WTO Discussion Paper.
2. Hose-Pipe - Encyclopedia (theodora.com)
3. Khan, S. (2016). *A Study on the authenticity of indigenious Art used in brass Anitque and Artifacts-A case study of Brass City "Moradabad"* (Doctoral dissertation, Pondicherry University).
4. Murphy, T., Norris, K.P. and Tippett, L.H.C. (1973). *Statistical methods for textile technologists*. Textile Institute.
5. TrendEconomy (2020). *Pakistan | Imports and Exports | World | Textile hosepiping and similar textile tubing, with or without lining, armour or accessories of other materials | Value (US$) and Value Growth, YoY (%) | 2009 – 2020*. Retrieved from https://trendeconomy.com/data/h2/Pakistan/5909

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 335-338

STATISTICAL ANALYSIS OF NET PROFIT OF SAPPHIRE TEXTILE MILL LAHORE

Muhammad Umair Zafar[1], Ali Raza[1], Hafiz Abdul Rehman Qaiser[1], Muhammad Ali[1], Ibrahim[1], Umair Alam[1], Arsalan Ahmed[1], Hamza Khan[1] and **Danish Hassan[2]**

[1] Department of Textile & Clothing, National Textile University, Karachi, Pakistan
Email: umairzafar9886@gmail.com; ali.ikhtiar406@gmail.com
ibrahim.khan9718@gmail.com; umairalam@ntu.edu.pk
arsalan_ahmed@ntu.edu.pk; hamzakhan@ntu.edu.pk

[2] Department of Applied Sciences, National Textile University Karachi, Pakistan. Email: danishhassan@ntu.edu.pk

ABSTRACT

Profit is the amount of money your business makes. Gross profit gives your business revenue minus how much you spend on the particular goods. Our business's other expenses are not included in that amount that you consume to make a good. In short gross profit is your company's profit before subtracting expenses. Net profit is your business's revenue after subtracting all operating, interest, and tax expenses, in addition to deducting your cost of goods sold. To calculate net profit, you must know your company's gross profit. Your business's net profit is known as a net loss if the number is negative. Here in this article we are going to find out how much sapphire textile mill Lahore earned in last seven years (2014-2020).

1. INTRODUCTION

The textile sector of Pakistan contributes nearly one-fourth of industrial value-added, contributes 8.5% of GDP of Pakistan and provides 40 percent (about 19 million) of the country employments to un-employed workers. Karachi, Faisalabad and Lahore are the cities having majority of top textile industries of Pakistan. Sapphire textile mill is a Pakistani textile industry which is located in Lahore, Pakistan. Its products are exported to over 35 countries around the globe and it employs more than 16,000 people.

In 2012, the company had a turnover of over US$800 million, with an asset base of over US$500 million. It has the spinning capacity of 350,000 spindles, with a production capacity of 65,000 kg of yarn per month from 6/s to 120/s, Knits dyeing and finishing and apparels unit has a capacity of knitting 500 tons of greige fabric and 300 tons of dyed and finished fabric and producing 430,000 garments per month (Qayum, 2020).

The company has also diversified into the power generation and dairy sectors. Sapphire Electric Company is a 234 Megawatt combined cycle plant in Muridke (city in Pakistan). Sapphire Dairies Private Limited operates a large mechanized dairy farm based on 100 acres near Manga, Lahore, with a herd size of 3000 and a target of 10,000 milking cows (300,000 liters per day) by 2020. We are going to find the Net Profit of last five years of

Sapphire Textile mills Lahore. All the data that I am going to use has been provided on the website of mill. I will be applying some rules of statistics on that data which we are studying in our courses.

2. METHODOLOGY

By getting this data we will apply some rules of statistics on it. The last seven years data is as follow:

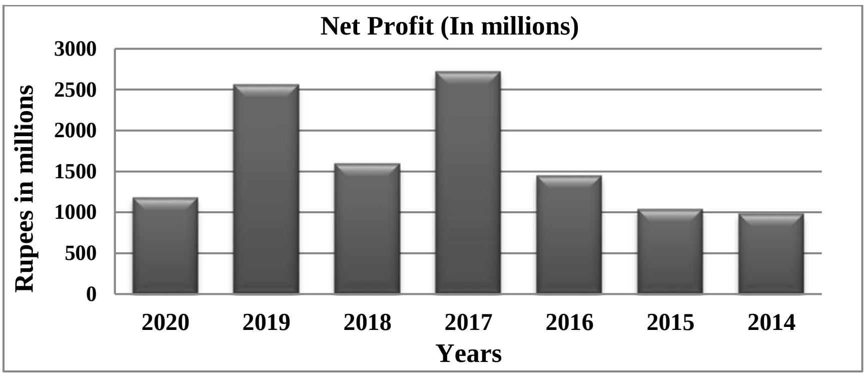
Annual Report 2020							
Seven years growth at a glance							
Rupees in Million							
Years	**2020**	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**
Profit After Tax	1179.1	2559.44	1595.06	2721.75	1448.21	1034.15	983.40

(Ibrar, 2020)

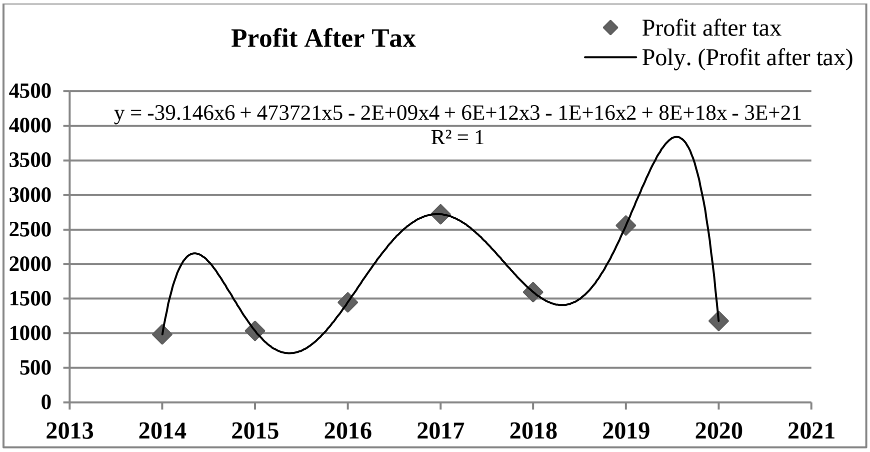
As the data has been provided regarding last seven years growth in net profit of the industry. We will show this data first on histogram and then further statistical techniques will be applied.

Histogram:

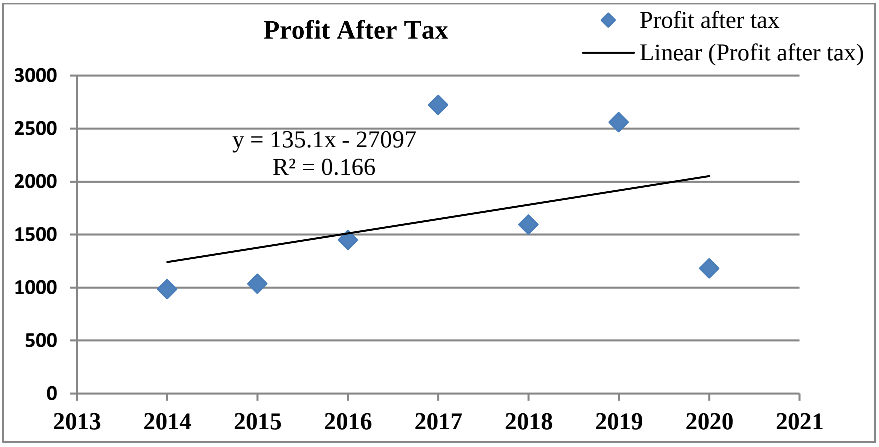
Histogram showed us that in which year the net profit went higher and in which year the net profit remained lower. So according to histogram we observed all the data of last seven years on a graph. We can say that having a glance at this graph and we will get to know the values.



This is the equation of polynomials which has a value near 1. The value varies with the change in the data. More the value closer to 1, the more reliable and stable our data is. So it can be seen from the graph that except sudden breakdowns happened to be seen in 2018 and 2020, the graph is showing us a positive trend. In 2018 General Elections took place and in 2020 COVID-19 rose.



We can explain it with another graph as well. We can see that the straight line shows the positive trend in data. This graph shows us an increasing trend in the net profit on mill.



Standard Deviation

In order to get standard deviation on the following data need to find mean first of this ungrouped data.

$$\mu = 1645.87$$

Now we need to find the standard deviation.

$$\sigma = 661.677\ Millions$$

Now CV (coefficient of variation) needed to be found. We know that = 40.2%.

3. RESULTS & DISCUSSIONS

As we are analyzing the last five years of data of sapphire textile mills Lahore. So according to the results of statistical tools and analysis, this is what we found:

- Histogram showed us that in which year the net profit went higher and in which year the net profit remained lower. So according to histogram we observed all the data of last seven years on a graph. Other graphs showed us an increasing and positive trend in Sapphire textile mills net profit. In 2018 General Elections were conducted and government changed and in 2020 COVID-19 appeared. In these two years a sudden breakdown happened to be seen but except that industry is on the right track in terms of income.
- The Standard Deviation is a measure of how much spread the numbers are. So, we applied this on our data and we got to know a certain value. As we found the standard deviation for population so we got a value that falls near the above given values.

4. COMMENTS & CONCLUSIONS

This activity will give us a huge boost in our practical lives. So according to all our calculations, we concluded that how we can apply statistical tools and formulas in our daily lives. As we have applied these things in our relevant fields so we have analyzed that how much statistics is involved in it. By this we got so much idea that how we can calculate the probable profit for upcoming years. We can find how much our profit money is dispersed then last few years. We can have a glance and we may know the trend of our data in last several years on histogram. So overall this has been a brilliant activity regarding our fields and studies.

5. ACKNOWLEDGEMENT

This activity will give us a huge boost in our practical lives. Our problem-solving techniques will help us out in our whole lives. So all we learned is just because of our faculty members who are always there whenever we need them. A special thanks to our subject teacher Dr. Danish Hassan, Dr. M. Qamar Khan (HOD Department of Textile & Clothing) and Dr. Khalid Pasha (Director Campus) of National Textile University Karachi Campus who motivated us to participate in this activity with all our heart involved.

REFERENCES

1. Ibrar, N. (2020). *http://www.sapphire.com.pk/stml/annualreports.htm*. Retrieved December 27, 2020, from http://www.sapphire.com.pk
2. Qayyum, A. (2020). *Sapphire Textile*. Retrieved December 28, 2020, from https://en.wikipedia.org/wiki/Sapphire_Textile
3. Sapphire Textile Mills Limited (2020). *Annual Reports 2008-09 to 2019-20*. Retrieved December 27, 2020, from http://www.sapphire.com.pk/stml/annualreports.htm

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 339-369

FACTORS ASSOCIATED WITH CHILD LABOUR IN PUNJAB PAKISTAN

Aqsa Abid, Anum Batool Naqvi and **Naila Alam**
Department of Statistics, Kinnaird College for Women, Lahore, Pakistan
Email: aqsaabid1023@gmail.com
anu.naqvi56@gmail.com
naila.alam1969@gmail.com

ABSTRACT

Child labor has emerged as a serious, widespread and growing issue in several components of the world. Child labor is defined as the participation of school-age children (5–15 years) within the labor force that is work for wage or in family enterprises to earn a living for themselves or to support household income. There’s not a well-documentation of the extent of participation of those children. However, the difficulty of child labor is incredibly complicated as a result of, on one hand it's going to take the child out of school and adversely impact human capital accumulation and lifetime earnings, on the other hand, labor work a child does is a significant risk coping mechanism, which can be necessary in assuaging the financial condition of a household within the short-term. Pakistan incorporates a high range of children taking part in economic activities and contributing significantly to household income. 5–14 years of children are playing a large vary of business activities. Some are serving to their parents in house-keeping, some are marketing newspapers or cigarettes within the streets, whereas some are working in formal and informal sectors of market. There is a worrying rise in child laborers and child abuse cases in Pakistan. Data from the Government’s 2012–2013 National Labor Force Survey (NNLFS) indicates that a high percentage 13.0% (2,449,480) of total working population is child laborers.

Data from MICS 2014 Punjab indicate a high percentage 16.0% of the total working population is child laborers. Developed countries have linked trade with child labor through the Harkin Bill and also the Uruguay round of multilateral Trade Agreement (1994), that banned the market factory-made or mined product created in whole or in part by children but fifteen years old-time. Pakistan is also facing restrictions on a number of its exports due to allegations of child labor. However, Pakistan has enacted the employment of children Act of 1991 that has banned employment of children below the age of fourteen years. The matter of child labor persists even within the presence of trade sanctions and legislation. According to report of UNICEF, the reasons are lack of proper law enforcement, negligence of parents and lack of awareness among the children and the society. Lack of empirical data and deep analysis of the related factors make it difficult to assess the severity of the issue. There is a shortage of quantitative work rather qualitative to explore and identify the factors associated with child labor in Punjab. Therefore, focused research is required that aimed at examining the factors related to child labor in Pakistan and can help in assessing the true factors and in decision making. This study is to explore different demographic variables of households that are associated with child labour in Punjab Pakistan.

1. INTRODUCTION

Child Labour

In line with International Labour Organization, any action apart from study or play, paid or unpaid, that's carried out by someone below the age of fifteen (14 in certain countries) is known as child labour. Asia contains a large number of children who are serving as child labourers. In Asia and Africa, over 200 million children are involved in some kind of child labour and about eight million are involved in dangerous and hazardous child labour. Developed countries have linked trade with child labour through the Harkin Bill and also the Uruguay round of multilateral Trade Agreement (1994) that banned the trade of factory made or mined product created by children younger than 15 years. Pakistan is one of the country facing restrictions on a number of its exports due to allegations on child labour. However, Pakistan has legitimate the employment of children (Act of 1991) and has banned employment of children below the age of fourteen years.

There are many factors associated with child labour in Pakistan, some are cultural among four provinces of Pakistan causing child labour (Jafri & Rashid, 1997). In Pakistan children of 5-14 years are playing a huge role in business, participating in financial exercises and contributing to family income. Some are serving to their family in house-keeping and some are showcasing papers or cigarettes inside the lanes. Different factors involved with child labour according to the culture of each province of Pakistan. (Awan et al. 2011).

Factors Affecting Child Labour

It is very hard to quantify the size of child labour in Pakistan. There are different components influencing child labour, growing step by step in Pakistan on a higher rate. Some of them are discussed here:

- Poverty
- Inadequate family income
- Illiteracy
- Family Background
- Broken families

Poverty

Poverty has its adverse effects on children in every aspect of life. It affects the development and learning abilities of children. Many of them even cannot afford to access the schools owing to poverty and are unable to continue their studies. The poor families have to work to meet the basic needs and therefore the children are engaged into child labour from very early age. Poverty is directly related with child labour. It is a major factor associated with child labour and without reducing poverty child labour ratio cannot be controlled (Fida, 2017).

Inadequate income of the family

Due to the low income of families children are forced to go out on roads to earn money. Children in homes with inadequate family income are less likely to go to colleges or schools than those whose families which are better off (Lenore, 1980).

Illiteracy

In many of the studies, education of the household head is directly related to child labour. Illiteracy is a social issue in the general public since individuals do not have enough education and therefore they engaged their children in child labour from a very early age (Maher, 2018).

Family background

Some children have family background in which they are taught to work from the start and so they are deprived from the opportunity to go to attend school. They are engaged in work from the younger age (Lenore, 1980).

Broken families

One of the factor associated with child labour is marital status of the household head. The children from the broken families are forced to earn money for their living (Asha, 2017).

2. LITERATURE REVIEW

Reviews on the attitudes, knowledge, and factor affecting the child labour are given in the following section. Studies included around the world like Bangladesh, Egypt, Geneva, India, Malaysia, Turkey, USA and Pakistan.

Ibrahim et al. (2019) has recently reviewed existing studies on child labour around the world. The 1050 studies on Pub Med and 833 studies on Science Direct were recovered. The studies conducted overall were cross-sectional. The result showed that although there was no cause and effect and there were higher health issues in working children than in the general population. Almost all studies concluded that factors related to child labour are insufficient educational chances, poverty and gender difference.

Nagar and Roy (2016) conducted a study in India to explore causes of child labour. The paper was related to descriptive studies both primary and secondary data. Number of child labour were present in Uttar Pradesh, Andhra Pradesh, Madhya Pradesh, Bihar and Maharashtra. Agriculture was the largest sector in India where many children work to support their families, many of children were forced to do work due to their family unemployment and parental education. These may be the major factors associated with child labour in India.

Siddiqi and Patrions (2014) performed a qualitative study and added that Child labour is frequent in rural areas. Schooling significantly contributes to child labour. Due to lack of education or distance of schools parents force their children to work and earn. Child labour is most targeted in Asia and Africa, that accounts over 90% child employment. In Asian country, 10 % of all employees are between the ages of ten and fourteen years. African nation has 12 million child workers. Many children have to work to get themselves educated so that they may avoid the child labour. Child labour in developing countries plays a decent role in economic growth of the state.

Zaidi et al. (2013) conducted a cross sectional study to explore causes of child labour in Pakistan. Seven hundred children were surveyed to get data. Surveyed areas included Commercial market, PirWidhaee, Kamran Market, Tipu Road Automobile Workshops, Airport road, Sawan, Murree Road and some residential areas of district Rawalpindi. It was concluded that most child labourers had a large number of family size which included 5-8 members and were male who had no education who were living in a family whose monthly income was very less as it was below 8000 Pakistani Rupee. The main reason of work for the child labourers was low family income.

Celik and Baybuga (2009) had conducted a study about the children that worked on the street and were from a mixed population. The participants involved in the study were selected from Ankara, Turkey. Children working on the streets were exposed to numerous risks including sexual, emotional and physical abuse. It was discovered that 2.3% were physically punished, 2% were economically used, 1.7% were forced to be involved in inappropriate work and 3% were compelled to perform various harmful activities. It was concluded that Children who worked on the streets needed information and training to build up themselves and needed the mindfulness about various types of abuse.

3. MATERIAL AND METHODS

Data

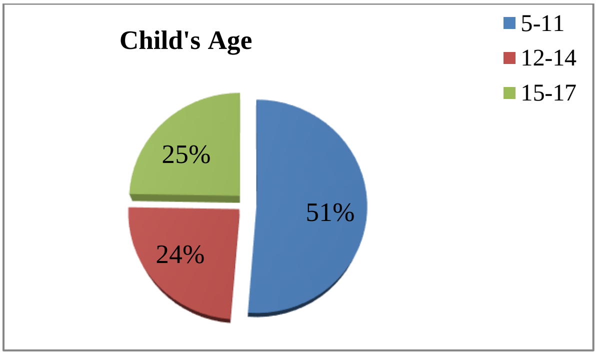
In order to attain the objectives of the study, data from Multiple Indicator Cluster Survey (MICS) Punjab 2014 is used. To investigate the factors associated with child labour in Pakistan, data is further cleaned and scrutinized. The data of MICS 2018 is also used to compare the statistics of two MICS surveys (2014 and 2018).

4. RESULTS

Descriptive Statistics for demographic variables of Households:

Pie Chart and frequency distribution are obtained to get statistics of different demographic variables of Households**.**

Frequency Distribution and Pie Chart for Child's Age:

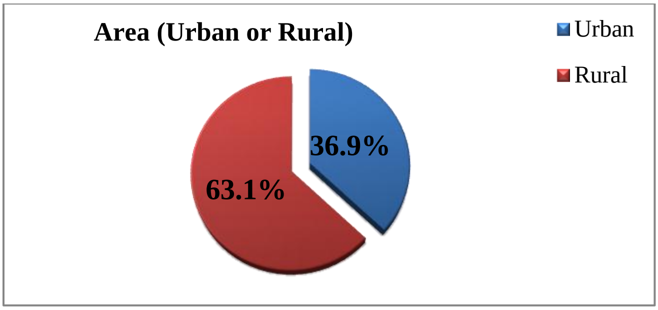


Pie Chart for Childs Age

Child's Age		
	Frequency	**Percent**
5-11	11562	51.0
12-14	5414	24.0
15-17	5575	25.0
Total	22556	100.0

Out of the total surveyed children 51% children were from the age between 5-11 years while 24% child laborers were from the age between 12-14 years.

Frequency Distribution and Pie Chart for Area (Urban or Rural):

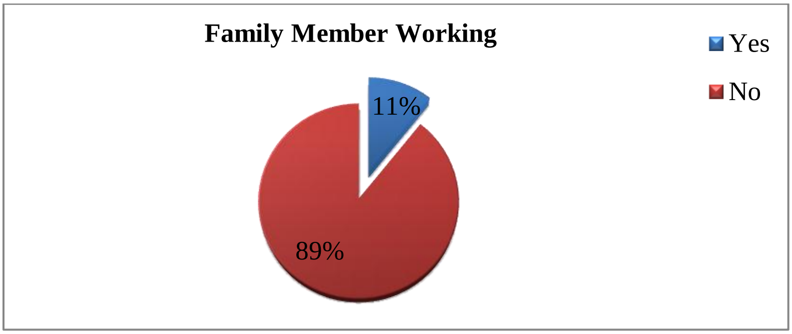


Pie Chart for Area (Urban or Rural)

Area (Urban or Rural)		
	Frequency	**Percent**
Urban Area	8334	36.9
Rural Area	14222	63.1
Total	22556	100.0

The frequency table and pie chart shows that 63.1% of the Households were from Rural Area while 36.9% were from Urban Area.

Frequency Distribution and Pie Chart for number of Family Members Working:

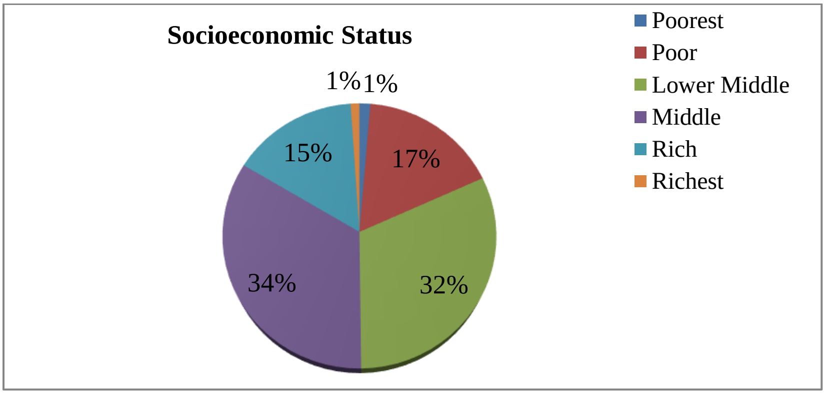


Pie Chart for Family Member Working

Family Member Working		
	Frequency	**Percent**
Yes	2430	11.0
No	20112	89.0
Total	22556	100.0

Out of 22556 respondents 2430 (11%) responded that their family members were working outside the village, city or country

Frequency Distribution and Pie Chart for Socioeconomic Status of household:

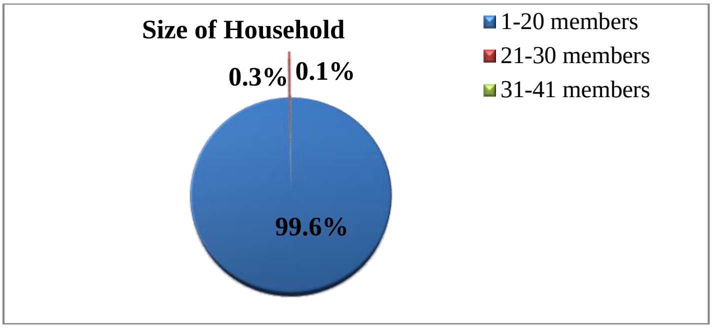


Pie Chart for Socioeconomic Status

Socioeconomic Status		
	Frequency	**Percent**
Poorest	296	1.0
Poor	3784	17.0
Lower Middle	7039	32.0
Middle	7528	34.0
Rich	3440	15.0
Richest	238	1.0
Total	22556	100.0

It can be observed from the above frequency table and pie chart that most of the respondents belonged (66%) to middle class or lower middle according to MICS 2014, Punjab.

Frequency Distribution and Pie Chart for Family Size

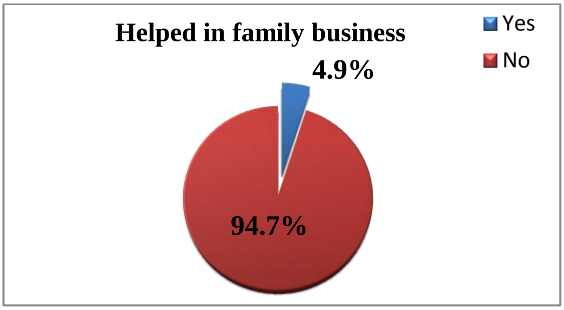


Pie Chart for Number of Household Members

Number of Household Members		
	Frequency	**Percent**
1-20	22475	99.6
21-30	67	0.3
31-41	12	0.1
Total	22556	100.0

From the pie chart it can be observed that most of the household members were 20 or less than 20 with the percentage of 99.6%.

Frequency Distribution and Pie Chart for Children who helped in family business:

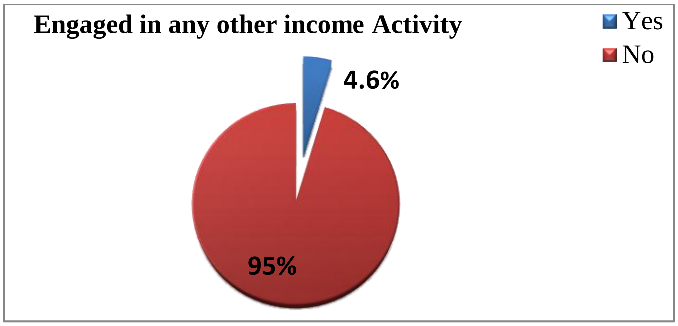


Pie Chart for Helped in Family Business

Helped in Family Business		
	Frequency	**Percent**
Yes	1105	4.9
No	21362	94.7
Total	22556	100.0

Among those who were engaged in child labor 94.7% were not helping in the Family Business while 4.9% were helping in the Family Business.

Frequency Distribution and Pie Chart for Children who were engaged in any other income related activity:

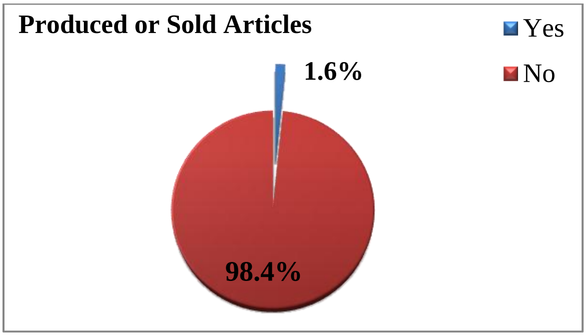


Pie Chart for children involved in any other income related activity

Involved in any other activity for income		
	Frequency	**Percent**
Yes	1033	4.6
No	21432	95.0
Total	22556	100.0

Out of the total 95% child laborers were not engaged in other activity for income while 4.6% were engaged.

Frequency Distribution and Pie Chart for Children who produced or Sold Articles

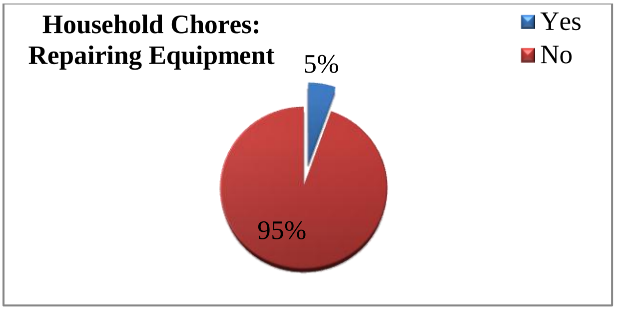


Pie Chart for Produced or Sold Articles

Produced or Sold Articles		
	Frequency	**Percent**
Yes	360	1.6
No	22103	98.4
Total	22556	100.0

Out of the total children who were engaged in child labor 98.4% didn't not Producing or Sold the things while 1.6% produced or sold the articles.

Frequency Distribution and Pie Chart for Children who were engaged in household chores:

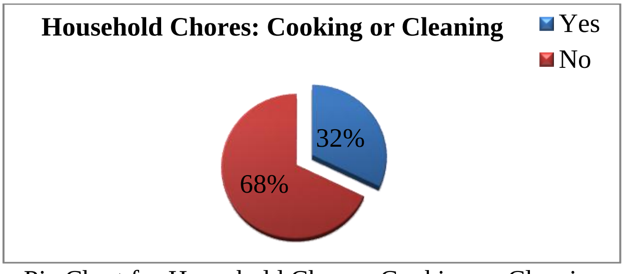


Pie Chart for Household Chores: Repairing Equipment

Household Chores: Repairing Equipment		
	Frequency	**Percent**
Yes	1221	5.0
No	21239	95.0
Total	22556	100.0

From the above pie chart we can see that 5% of the children were engaged in repairing equipment whereas 95% were not of MICS 2014.

Frequency Distribution and Pie Chart for Children who were engaged in cooking or cleaning:

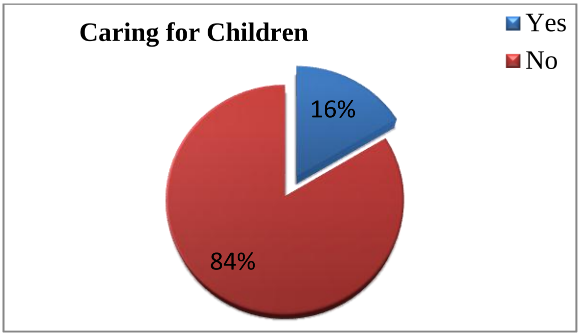


Pie Chart for Household Chores: Cooking or Cleaning

Household Chores: Cooking or Cleaning		
	Frequency	**Percent**
Yes	7231	32.0
No	15233	68.0
Total	22556	100.0

From the above pie chart it can observed that 32% of the children were engaged in cooking or cleaning whereas 68% were not.

Frequency Distribution and Pie Chart for Children who were caring for children:



Pie Chart for children who cared for Children

Household Chores: Caring for Children		
	Frequency	**Percent**
Yes	3688	16.0
No	18764	84.0
Total	22556	100.0

From the above pie chart we can observe that 16% of the children were caring for children whereas 84% were not.

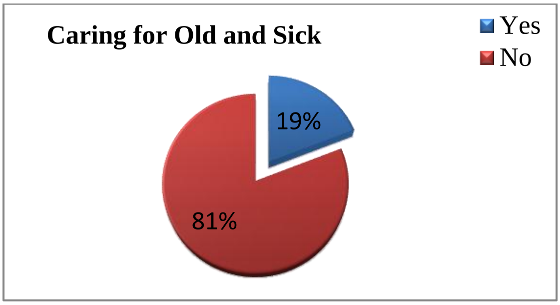
Frequency Distribution and Pie Chart for Children who were engaged in washing clothes:

Pie Chart for Household Chores: Washing Clothes

Household Chores: Washing Clothes		
	Frequency	**Percent**
Yes	4994	22.0
No	17464	78.0
Total	22556	100.0

It can be observed from the above pie chart and frequency distribution that 78% children were not washing clothes and 22% were engaged in washing clothes.

Frequency Distribution and Pie Chart for Children who cared for Old or sick:

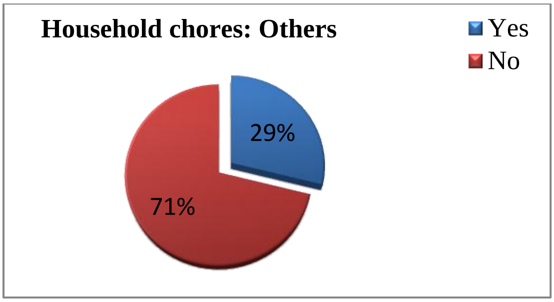


Pie Chart for Household Chores: Caring for Old and Sick

Household Chores: Caring for Old and Sick		
	Frequency	**Percent**
Yes	4276	19.0
No	18179	81.0
Total	22556	100.0

From the above pie chart it can be observed that 19% children had to care for old and sick persons at home according to MICS (2014) Punjab.

Frequency Distribution and Pie Chart for Children who were engaged in other household chores:

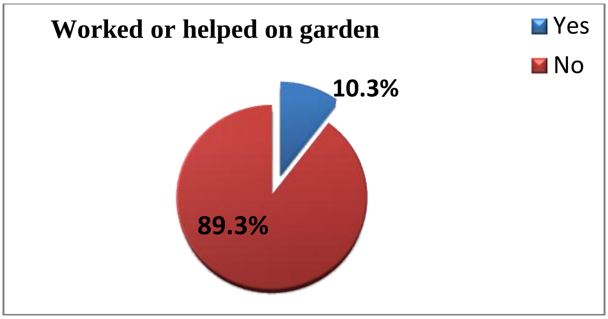


Pie Chart for Household Chores: Others

Household Chores: Others		
	Frequency	**Percent**
Yes	6465	29.0
No	15939	71.0
Total	22556	100.0

From the above pie chart we can observed that 29% children were engaged some others household chores.

Frequency Distribution and Pie Chart for the children who worked or helped on garden:

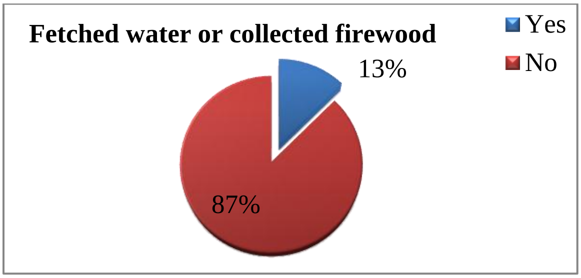


Pie Chart for Worked or helped on garden

Worked or helped on garden		
	Frequency	**Percent**
Yes	2318	10.3
No	20150	89.3
Total	22556	100.0

Out of the total, 89.3% children were not working or helping on garden while 10.3% worked or helped on garden.

Frequency Distribution and Pie Chart for Children who fetched water or collected firewood

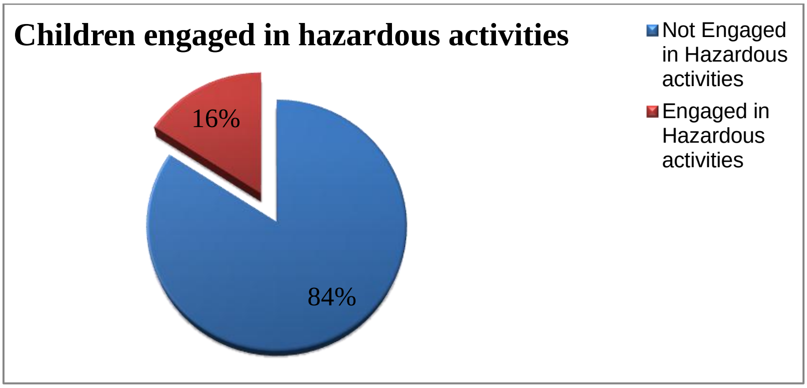


Pie Chart for Fetched water or collected firewood

Fetched water or collected firewood		
	Frequency	**Percent**
Yes	2808	13.0
No	19652	87.0
Total	22556	100.0

We had concluded from the above that 87.0% of children were not able to fetched water or collect firewood whereas 13.0% of children were able to fetched water or collect firewood

Frequency Distribution and Pie Chart for the children who were involved in hazardous working conditions:

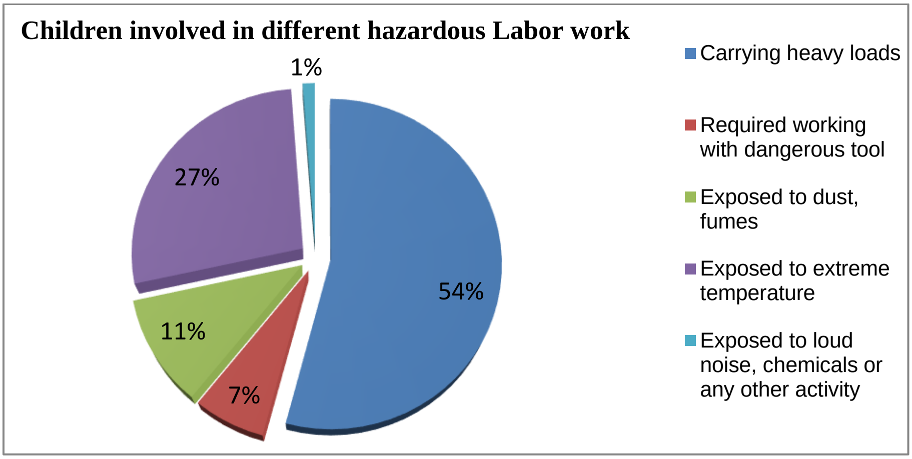


Pie Chart of Children engaged in hazardous activities

Children engaged in hazardous activities		
	Frequency	**Percent**
Not Engaged in Hazardous activities	18984	84.0
Engaged in Hazardous activities	3572	16.0
Total	22556	100.0

The above table and chart show that among total children 16% of the children were engaged in harmful labor activities whereas 84% were not.

Frequency Distribution and Pie Chart for the proportion of children who were involved in different hazardous working condition:

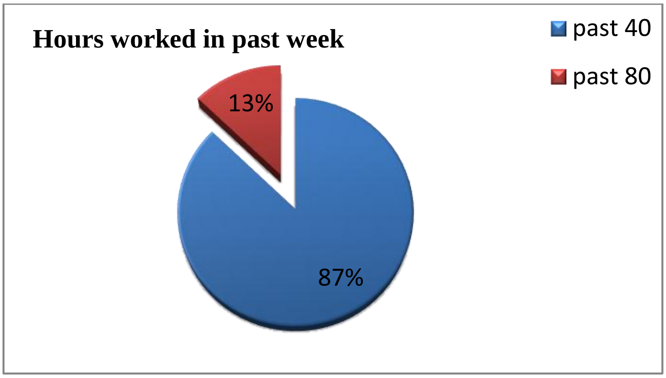


.Pie Chart of Children involved in different hazardous Labor work

Children involved in different hazardous Labor work		
	Frequency	**Percent**
Carrying heavy loads	1699	54%
Required working with dangerous tool	208	7%
Exposed to dust, fumes	339 (%)	11%
Exposed to extreme temperature	855(%)	27%
Exposed to loud noise, chemicals or any other activity	37(1%)	1%
Total	3572	100

The above frequency distribution and pie chart show that among children who were engaged in any type of hazardous activities, 1699 (54%) of the children were those who were involved in activities carrying heavy loads, 7% were those who were required to work with dangerous tools, 11% were exposed to dust or fumes, 27% were exposed to extreme temperatures and 1% were those who had to work at heights, work with chemical, loud noise or any other harmful activity.

Frequency Distribution and Pie Chart for Hours worked past week



.Pie Chart for Hours worked past week

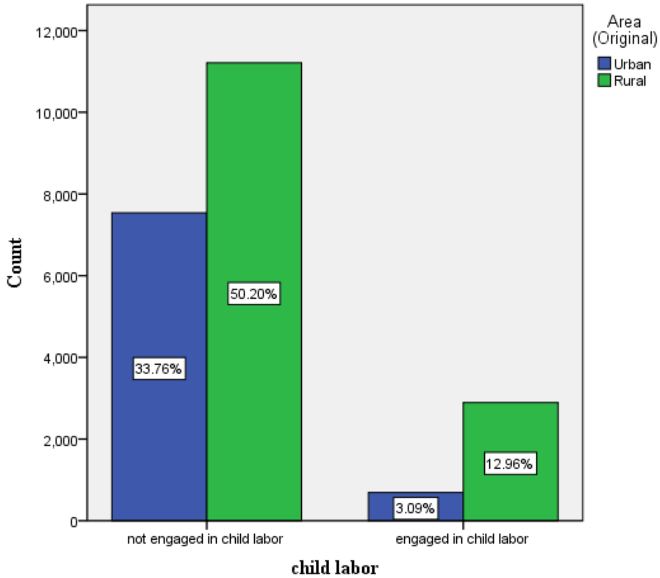
Hours worked past week		
	Frequency	**Percent**
0 to 40	3021	87.2
40 to 80	444	12.8
Total	3465	100.0

It can be observed that 87.2% the child laborers had worked 40 and fewer hours in a week and 12.8% had worked past 40 to 80 hours in a week.

Crosstab and association between child labour and demographic variables of households:

In this section association between different characteristics of households is checked with child labour using chi square test of independence.

Association between child labour and Area (Urban, Rural)



Bar chart of Child Labour and Area

Child Labour and Area		
Area	**Not Engaged in Child Labour**	**Engaged in Child Labour**
Urban	7539 (33.75%)	690 (3.09%)
Rural	11211 (50.20%)	2894 (12.96%)
Total	18750	3584

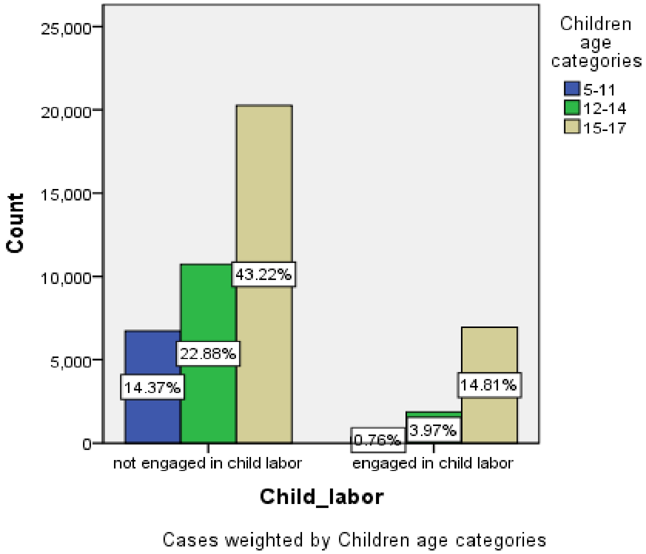
It can be observed from the above table that from urban area 3.09% of the children were engaged in child labor whereas 12.96% of the children who were engaged in child labor were from rural areas.

Chi-Square Tests of Child Labor and Area

Chi-Square Test	
Statistic	**P-value**
567.836	0.000

The above shows that the p-value for chi square test is less than 0.05 so we conclude that there is association between child labor and area.

Association between Child Labor and age of child



Bar Chart of Child Labor and Children age

Child Labor and Children age		
Child age	**Not Engaged in Child Labor**	**Engaged in Child Labor**
5-11	6740 (14.37%)	355 (0.76%)
12-14	5364 (22.88%)	930 (3.97%)
15-17	6755 (43.22%)	2315 (14.81%)
Total	18859	3600

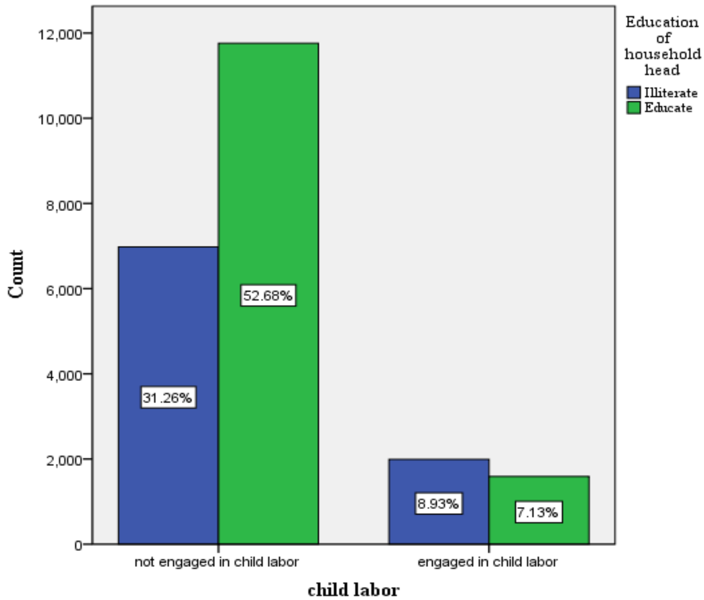
The above shows that among the children who were engaged in child labor 0.76% were of 5-11 years, 3.97% were from 12 to 14 years and 14.81% were from 15 to 17 years.

Chi-Square Tests of Child Labour and Child age

Chi-Square Test	
Statistic	**P-value**
1246.660	0.000

The p-value for chi square test of association is less than 0.05, therefore we conclude that there is association between child labour and the age of children.

Association between Child labour and Education of Household head



Bar Chart of Child Labour and Education of household head

Child Labour and Education		
Education	**Not Engaged in Child Labour**	**Engaged in Child Labour**
Illiterate	6980 (31.26%)	1993(8.93%)
Educated	11762 (52.68%)	1591 (7.13%)
Total	18742	3584

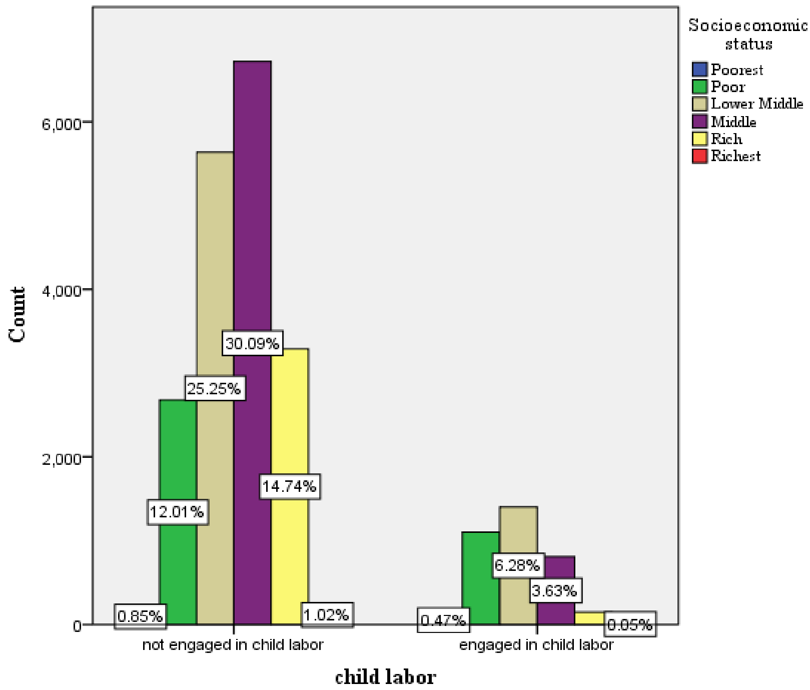
Above table shows that the 8.93% of the children who were engaged in child labour were from the households whose head education was non/pre-school whereas 7.13% of the children were from those households whose heads were educated.

Chi-Square Tests of Child Labour and Education of household head

Chi-Square Test	
Statistic	**P-value**
422.176	0.000

The p-value of chi square test is less than 0.05 indicating that there is association between child Labour and education of household head

Association between Child labour and Socio-economic status



Bar Chart of Child Labour and Socioeconomic status

Child Labour and Socioeconomic status		
Socioeconomic status	**Not Engaged in Child Labour**	**Engaged in Child Labour**
Poorest	190 (0.85%)	106 (0.47%)
Poor	2681 (12.01%)	1103 (4.94%)
Lower Middle	5635 (25.25%)	1402 (6.28%)
Middle	6717 (30.09%)	810 (3.63%)
Rich	3289 (14.74%)	149 (0.67%)
Richest	227 (1.02%)	11 (0.05%)
Total	18739	3581

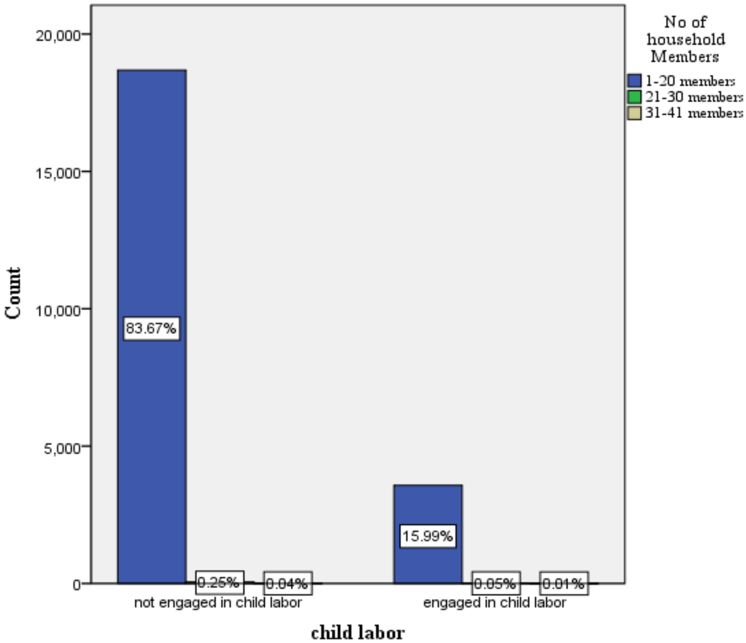
Table above table shows that the children who were engaged in child labour 5.41% were from poor or poorest families and 6.28% were from lower middle class whereas from the richest only 0.05% were engaged in child labor.

Chi-Square Tests of Child Labour and Socioeconomic status

Chi-Square Test	
Statistic	**P-value**
1175.942	0.000

The p-value for test is less than 0.05. Therefore, we conclude that there is association between child labour and socioeconomic status of the household.

Association between Child labour and Number of Household Members



Bar Chart of Child Labour and Number of Household Members

Child Labour and Number of Household Members		
Household Members	**Not Engaged in Child Labour**	**Engaged in Child Labour**
1-20	18685 (83.67%)	3571 (15.99%)
21-30	55 (0.25%)	11 (0.05%)
31-41	8(0.04%)	2 (0.01%)
Total	18748	3584

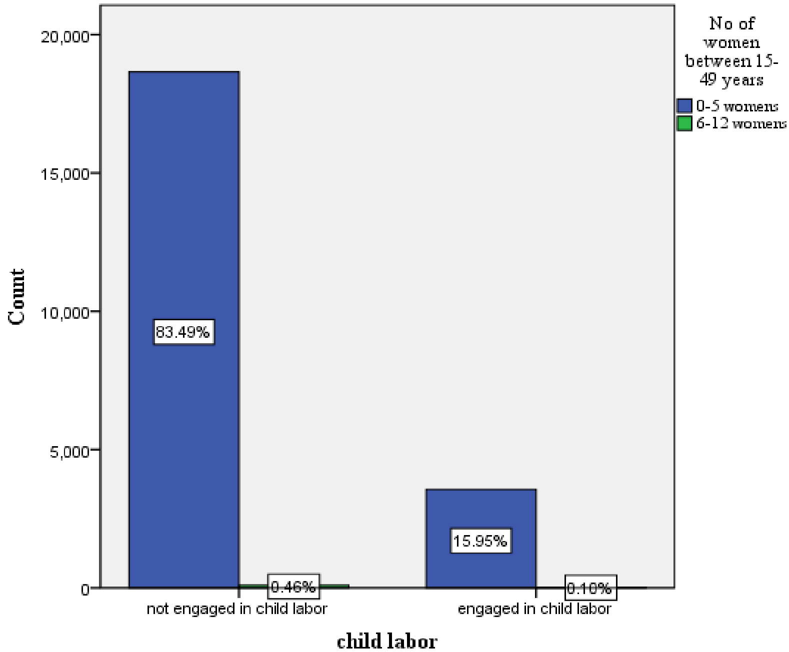
The above table shows that from the households who had one to twenty members, the 15.99% of the children were engaged in child labour. It can be observed that the households with 21 to 30 members, 0.05% of the children were engaged in child labour.

Chi-Square Tests of Child Labour and Number of Household Members

Chi-Square Test	
Statistic	**P-value**
0.135	0.905

Table shows that the p-value is greater than 0.05, so there is no association between child labour and number of households

Association between Child labour and Number of Women between 15-49



Bar Chart of Child Labour and Number of Women between15-49

Child Labour and Number of Women between 15-49		
Number of Women between 15-49	**Not Engaged in Child Labour**	**Engaged in Child Labour**
0-5	18646 (83.49%)	3561 (15.95%)
6-12	102 (0.46%)	23 (0.10%)
Total	18748	3584

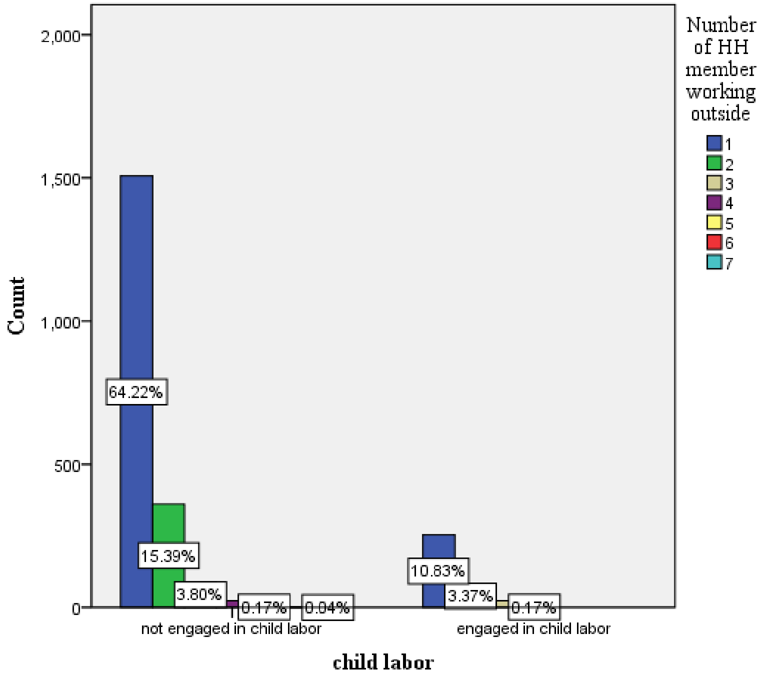
The above table shows that from households which had 0 to 5 number of women between age 15 to 49, 15.95% of the children were engaged in child labour whereas 0.10% were from the households those had 6 to 12 women.

Chi-Square Tests of Child Labour and Number of Women between 15-49

Chi-Square Test	
Statistic	**P-value**
0.516	0.473

The p-value is greater than 0.05 indicating that there is no association between child labour and number of women between 15-49

Association between Child Labour and Number of Households members working:



Bar Charts of Child Labour and Number of household member working

Child Labour and Number of household member working outside		
Number of members	**Not Engaged in Child Labour**	**Engaged in Child Labour**
1	1506 (64.24%)	254 (10.83%)
2	361 (15.39%)	79 (3.37%)
3	89 (3.80%)	23 (1.99%)
More than 4	29 (0.21%)	4 (0.17%)
Total	2042	369

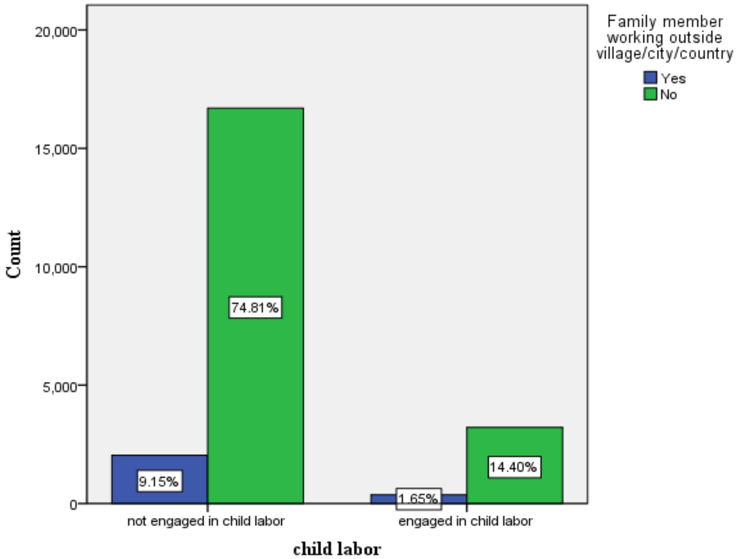
The above shows that 10.83% of the children who were engaged in child labour were from the households whose only one member was working outside. It can be observed that the household from which more than 4 household members were working 0.17% of the children were engaged in child labour and 0.21% were not engaged in the child labour. Out of 369 children who were engaged in child labour 14.2% were from those households whom 1 or 2 family members were working outside.

Chi-Square Tests of Child Labour and number of household member working outside

Chi-Square Test	
Statistic	**P-value**
7.013	0.428

The above table shows that the p-value for chi square test of association is greater than 0.05, so we conclude that there is no association between child labour and number of family members working outside.

Association between Child labour and Family Member working outside village/city/country



Bar chart of Child Labour and Family Member Working Outside Village/City/Country

Child Labour and Family Member Working Outside Village/City/Country		
Family Member Working Outside	**Not Engaged in Child Labour**	**Engaged in Child Labour**
Yes	2042 (9.15%)	369 (1.65%)
No	16702 (74.81%)	3214 (14.40%)
Total	18750	3584

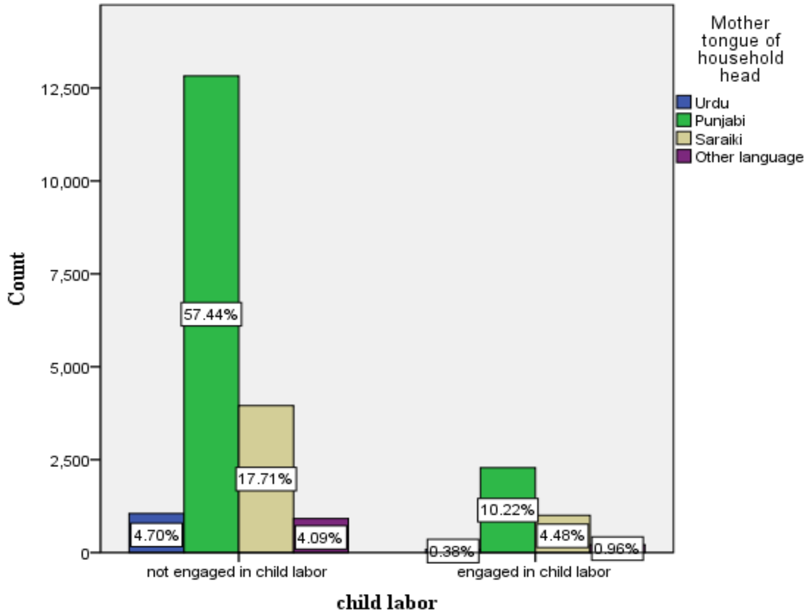
Table shows that 1.65% of the children who were engaged in child labour were those whose family members were working outside (village, City or Country). It can be observed that 14.40% of the children who were engaged in child labour were those whose family member were not working outside (village, City or Country).

Chi-Square Tests of Child Labour and Family Member Working Outside Village/City/Country

Chi-Square Test	
Statistic	**P-value**
1.124	0.570

The p-value is greater than 0.05 which is indicating that there is no association between Child Labour and Family Member Working Outside Village/City/Country

Association between Child labour and Mother tongue of household head



Bar Chart of Child Labour and Mother tongue of household head

Child Labour and Mother tongue of household head		
Mother tongue	**Not Engaged in Child Labour**	**Engaged in Child Labour**
Urdu	1050 (4.70%)	85 (0.38%)
Punjabi	12826 (57.44%)	2282 (10.22%)
Saraiki	3955 (17.71%)	1001 (4.48%)
Other Language	914 (4.09%)	215 (0.96%)
Total	18750	3584

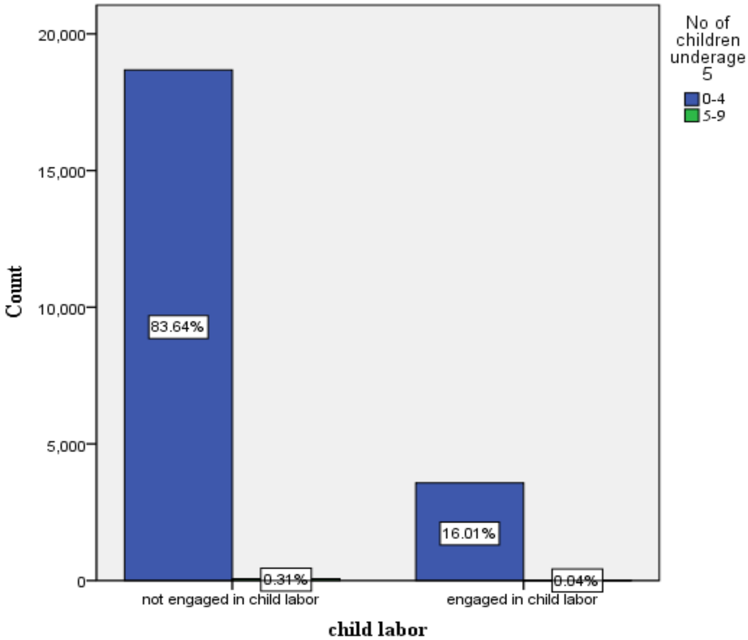
The above table shows that 10.22% of the children who were engaged in child labour were from the household whose mother tongue was Punjabi. It can be observed that 0.38% of the children who were engaged in child labour were from the household whose mother tongue was Urdu and the 4.48 % were those whose mother tongue was Saraiki.

Chi-Square Tests of Child Labour and Mother tongue of household head

Chi-Square Test	
Statistic	**P-value**
142.568	0.000

Above Table shows that the p-value of chi square test of association is less than 0.05 indicating that there is association between child labour and mother tongue of household head.

Association between Child Labour and Number of Children under age 5



Bar Chart of Child Labour and Number of Children under age 5

Child Labour and Number of Children under age 5		
Number of Children under age 5	**Not Engaged in Child Labour**	**Engaged in Child Labour**
0-4	18679 (83.64%)	3575 (16.01%)
5-9	69 (0.31%)	9(0.04%)
Total	18748	3584

Table shows that the 16.01% who were engaged in child labour were from the households which had 0 to 4 number of children under age 5. It can be observed that the 0.04% of the children who were engaged in child labour had 5 to 9 children of under age 5.

Chi-Square Tests of Child Labour and Number of Children under age 5

Chi-Square Test	
Statistic	**P-value**
1.182	0.277

The above table shows an insignificant p-value at 0.05 level of significance, indicating that there is no association between child labour and children under age 5.

Factor Analysis

In this section factor analysis is used to explore main dimensions of child labour and child abuse in Punjab according to MICS 2014 Punjab.

Factor Analysis for items related to Child labor

The total 21 items that were used by MICS 2014 to quantify child labour in Punjab Pakistan are compressed to explore main factors of child labour. The items having loadings < 0.34 on the corresponding factors are ignored. The five main factors are explored using factor analysis whose Eigen values are greater than 1.

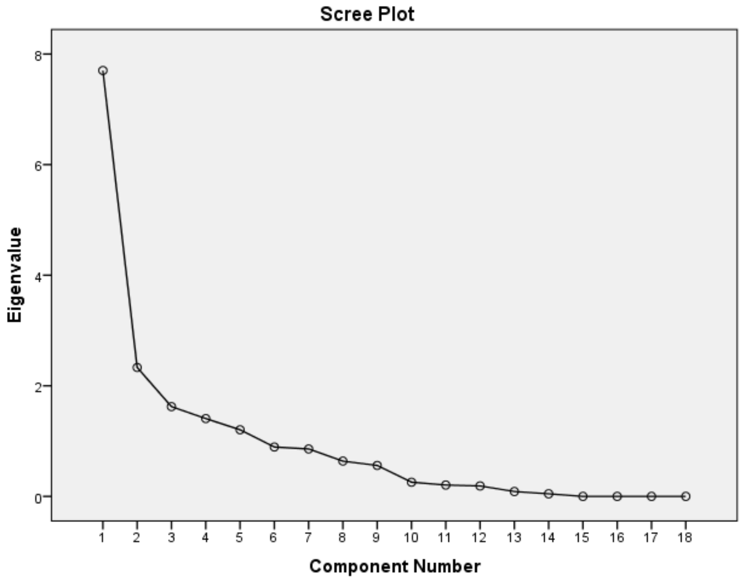
KMO and Bartlett's Test of Child Labour	
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.851
Bartlett's Test of Sphericity	0.000

The above Table shows the value of KMO test comes out to be 0.851 and the p-value of Bartlett's test of Sphericity comes out to be 0.000 which show that items are correlated and the sample size is adequate to run factor analysis.

Total variance explained

Components	Eigen Values	Cumulative%
1	7.703	40.542
2	2.372	53.025
3	1.622	61.564
4	1.412	68.998
5	1.229	75.468

Table shows that from 21 items related to child labor, 5 main factors are extracted which have Eigen values greater than 1. These 5 factors are able to explain 75.468% of the total variation.



Scree plot for factor analysis of child labour

The above graph is showing that the elbow of scree plot at 5, is an indicator of the selection of factors having Eigen values greater than 1.

Factor Analysis Child Labor		
Factors	**Variables**	**Loadings**
F1	Activities required carrying heavy loads	0.980
	Activities required working with dangerous tools or heavy machinery	0.957
	Description of work: exposed to dust, fumes, or gas	0.995
	Description of work: exposed to extreme temperatures or humidity	0.995
	Description of work: exposed to loud noise or vibration	0.995
	Description of work: required to work at heights	0.995
	Description of work: required to work with chemicals	0.995
	Description of work: exposed to other	0.973
F2	Household chores: shopping	0.767
	Household chores: cooking or cleaning	0.739
	Household chores: washing clothes	0.766
F3	Fetched water or collected firewood	0.348
	Household chores: caring for old or sick	0.716
	Household chores: caring for children	0.767
F4	Worked or helped on garden	0.780
	Helped in Family business	0.876
	Household chores: repairing equipment	0.475
F5	Engaged in any other activity for income Household chores:	0.948
	other	0.863

First column of the above table is showing 5 constructs that are explored by factor analysis, 2nd column is showing items which have significant loadings on the corresponding constructs and 3rd column show the item loadings.

Naming the variables

Factors	Variables
F1	Hazardous activities
F2	Household Chores
F3	Caring for sick and children
F4	Family business
F5	Other activities

The above table is showing name of the constructs according to the associated items.

5. DISCUSSION

The MICS 2014 data was collected for a randomly selected child of age of 5-17 years from each of the surveyed household. The statistics show that among selected children 45.5% were of 4-10 years and 54.5% were 11-17 years old. The 36.9% of the children belonged to urban area whereas 63.1% children were from rural area, 89% of children were those whose family member worked outside (village/city/country) and 68% of the children mother tongue was Punjabi. The results show that according to MICS 2014 Punjab, 16% of the children were engaged in any type of child labour. According to MICS 2018, the 13.4% of the children are engaged in child labour that shows a reduction in the percentage of child labourers.

From the final report on findings of MICS-2018 we see that only 7% children of age 5-17 are involved in child labour. The 21.4% of children were involved in economic activities according to MICS 2014 which include paid or unpaid work for someone who is not a member of the household, work for family farm or business whereas in MICS-2018 report we see that 29% of children were involved in economic activities.

22.75% of children were involved in Household chores which include activities such as cooking, cleaning or caring for children, as well as collecting firewood or fetching water. Whereas in MICS-2018 report we see that 11% of children were involved in household chores.

According to MICS 2014, 15.83 % of children were working in hazardous condition whereas in MICS 2018 report 16% of children are working in harmful economic condition which indicates a slight rise in hard child labor in Punjab, Pakistan.

According to the statistical results the associated factors with child labor are Socio economic status (0.000) of household. The Education of household head (0.000), Area (0.000), Mother tongue of household (0.000), and Age of children are statistically significantly associated with child labour according to MICS (2014) Punjab data.

The results are supported by Zaidi et al. (2013) who reported that child labourers are enforced to work due to poverty and illiteracy.

Results are supported by Hyder and Malik (2007). They conducted a study on children and stated that outsized family, joblessness, illiteracy and poverty are the main factors for child labour but in this study no association of family size was found with child labour.

The Area (0.05) has a strong association with child labour that indicates that most of the child labourers were from rural areas in line with Siddique and Patrions (2014) who reported that Child labour is frequent in rural areas.

Results are supported by Hyder and Malik (2007). They conducted a study on children and stated that outsized family, joblessness, illiteracy and poverty are the main factors for child labour but in this study no association of family size was found with child labour.

The Area (0.05) has a strong association with child labour that indicates that most of the child labourers were from rural areas in line with Siddique and Patrions (2014) who reported that Child labour is frequent in rural areas.

The child labour has a statistically significant association with child age that indicates that older the child, more likely to be involved in child labour.

The other variables such as “number of women” (0.756), “family member working outside the country/village/city” (0.570) and number of children under age 5 have no association with child labour in Punjab.

6. CONCLUSION

This study is conducted to explore main factors that are associated with child labour in Punjab, Pakistan. The MICS 2014, Punjab data is used for this purpose. The results show that “number of household members”, Socio economic status of household, Education of household head, Urban and rural area, Mother tongue and age of children are associated with child labour according to MICS (2014) Punjab. The results show that the chances of experiencing physical punishment were more in children who were engaged in any form of child labour.

It is observed that Area (urban or rural), mother tongue of household head, Education of household head’ age of children, number of household members, number of children under age 5 and socioeconomic status of household are also associated with child abuse. The perception or belief of household head about brought up of children is a significant predictor of child abuse as those parents or caretakers who think that physical punishment is necessary to be brought up a child properly are more intended to corporal punishments. The other significant predictors of child abuse are number of working hours, child’s age, mother tongue of household and Education of household head.

7. LIMITATIONS OF STUDY

- The statistical analysis to find relationship between variables was done using MICS 2014 as the latest data of MICS 2017-2018 was not available.
- The data on gender (children) was not available therefore it was not possible to make comparison on gender basis.

- The data was not available on attending school of children; therefore it was not possible to make comparison of the children who were attending school and those who were not.
- As the data was missing, therefore the results are based on the reduced sample, obtained after excluding the missing cases.

8. RECOMMENDATIONS

The study reveal that poverty, illiteracy and family size are positively associated factors with both child abuse and child labour. Therefore, it is not possible to reduce child labour and child abuse without focusing on these associated factors.

REFERENCES

1. Azhar, U. (2015). Child Labour and Violation of the Rights of the Child in Pakistan. *Journal of European Studies*, 31(1).
2. Celik, S.S. (2009). Verbal, physical and sexual abuse among children working on the street. *The Australian Journal of Advanced Nursing*, 26(4), 14-22.
3. Dryden, K. M., Dryden Kyrsha, M., & Guide, N. A. R. (2009). Child Abuse and Neglect: *A Resource Guide*.
4. Fry, D. (2012). Child maltreatment: Prevalence, incidence and consequences in the East Asia and Pacific Region: A Systematic Review of Research, *Strengthening Child Protection Systems Series: No 1*, UNICEF.
5. International Labour Organization (2017). *Global estimates of child labour: Results and trends*, 2012–2016, ILO, Geneva.
6. Hyder, A.A. and Malik, F.A. (2007). Violence against children: a challenge for public health in Pakistan. *Journal of Health, Population, and Nutrition*, 25(2), 168-178.
7. Hair Jr., J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2010). *Multivariate Data Analysis: A Global Perspective*. 7th Edition, Pearson Education, Upper Saddle River.
8. Ike, C.A. and Twumasi-Ankrah, K. (1999). Child abuse and child labour across culture: Implications for research, prevention and policy implementation. *Journal of Social Development in Africa*, 14(2), 109-118.
9. Ibrahim, A., Abdalla, S.M., Jafer, M., Abdelgadir, J. and de Vries, N. (2019). Child labor and health: a systematic literature review of the impacts of child labor on child's health in low-and middle-income countries. Journal of Public Health, 41(1), 18-26.
10. Kazmi, S.M.A. (2015). *How Does Socio-economic Factors Force Children Into Child Labour? A Case Study of Sahiwal District, Punjab, Pakistan*. Sustainable Development Policy Institute.
11. Khalid, U. and Shahnaz, L. (2004). Socio Economic Conditions of Child Labourers in Pakistan: Evidence from the Labour Force Survey, *Lahore Journal of Economics*, 9(1), 85-105.
12. Mahmod, N.K.A.N., Salleh, M.C.M., Muhammad, A.A. and Mohd, A. (2016). A Study on Child Labour as a Form of Child Abuse in Malaysia. *International Journal of Social Science and Humanity*, 6(7), 525-530.
13. Mazhar, Z.A. (2008). *Child Labour in Pakistan: Including the Excluded* (Doctoral dissertation, Erasmus University).

14. Nafees, A.A., Khan, K.S., Fatmi, Z. and Aslam, M. (2012). Situation analysis of child labour in Karachi, Pakistan: a qualitative study. *Journal of Pakistan Medical Association*, 62(10), 1075-1082.
15. Runyan, D. (2002). Child abuse and neglect by parents and other caregivers (Chapter 3), *World Report on Violence and Health*, 59-86.
16. Siddique, F. and Patrinos, H.A. (1995). *Child labor: issues, causes and interventions*. Education and Social Policy Department, Human Resources Development and Operations Policy, the World Bank.
17. University of Bristol (2016). Significant new study shows importance of help for childhood sexual abuse victims. *ScienceDaily*. Retrieved September 24, 2018 from www.sciencedaily.com/releases/2016/02/160222111050.html
18. UNICEF (2018) *Multiple Indicator Cluster Survey 2017*-18. Bureau of Statistics Punjab, Lahore.
19. Wallat, J.R. (2017). *Crossing the Line: Is Corporal Punishment Child Abuse*? Retrieved from Sophia, the St. Catherine University repository website: https://sophia.stkate.edu/msw_papers/806
20. Zaidi, N., Javed, N. and Khan, M.S. (2013). Child Labour in Pakistan-A tip of an iceberg. *Journal of Postgraduate Medical Institute (Peshawar-Pakistan)*, 27(3), 285-289.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 371-380

A STUDY ON THE PREVALENCE OF COUGH IN THE PUNJAB PROVINCE OF PAKISTAN

Saleha Naghmi Habibullah, Naila Alam, Fatima Jamil, Aqsa Sheikh, Iqra Manzoor, Laiba Subhani and **Rozeena Gulab**
Kinnaird College for Women, Lahore, Pakistan
Email: saleha.habibullah@kinnaird.edu.pk
naila.alam@kinnaird.edu.pk
fatima.jamil@kinnaird.edu.pk

ABSTRACT

A cough, also known as tussis, is a voluntary or involuntary act that clears the throat and breathing passage of foreign particles, microbes, irritants, fluids, and mucus; it is a rapid expulsion of air from the lungs [1]. A cough can be due to a respiratory tract infection such as the common cold, pneumonia, acute bronchitis, tuberculosis or COVID-19. In a multitude of cases, acute coughs are due to the common cold [2]. According to the World Health Organization, respiratory tract infections are among the most important human health problems because of their high incidence and consequent economic costs. The majority of respiratory infections are confined to the upper respiratory tract and most of these are simple coughs or colds. Drug therapy for viral common colds produces few measurable benefits [3]. Punjab being the most populous province of Pakistan, in this research paper, we explore the prevalence of cough in the Punjab over a period of eight years from 2003 to 2014 by analyzing data pertaining to cough collected during the Multiple Indicator Cluster Surveys (MICS) that were carried out in the province during this period. Analyses have been carried out for both male and female persons belonging to various age-groups, and comparisons have been made between the situations of the urban and the rural people as well as between the situations of people belonging to North/Central Punjab and those belonging to South Punjab. The purpose of this in-depth analysis is to assist the relevant health departments in formulating cough-reduction strategies specific to the various categories of people created by virtue of the analysis. The paper ends with some recommendations for a reduction in the incidence of cough not only in the Punjab but also in the entire country.

Analyses include urban/rural comparison, comparison between North/Central and South Punjab as well as comparisons between the rich and the poor and between the educated and the uneducated. Results of the analyses reveal important differences between various socio-economic classes of people and, as such, it is clear that steps aimed at reducing the prevalence of this infectious disease need to be in accordance with the requirements of the people belonging to the different classes.

KEYWORDS

Cough, WHO, MICS.

1. INTRODUCTION

Good health and well-being is one of the seventeen important Sustainable Development Goals adopted by the world in 2015.

Sustainable Development Goals

Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. “Good heath means to be fit physically and emotionally as well as being illness-free”.

The milestones of the SDG 3 are:

- Improving health for all and decreasing health disparities
- Supporting health through a life-course approach and empowering citizens
- Undertaking Europe's major disease burdens of NCDs and communicable diseases
- Strengthening people-centered health systems and public health capacity, including response capacity and preparedness for dealing with emergencies
- Producing supportive environments and spirited communities (Dyakova M. et al., 2017).

Pakistan being a developing country and Punjab being the most populous province of Pakistan, in this research paper, we explore the prevalence of cough in the Punjab over a period from year 2003 to 2014 by analyzing data pertaining to cough collected during the Multiple Indicator Cluster Surveys (MICS) that were carried out in the province during this period.

A cough can be due to a respiratory tract infection such as the common cold, pneumonia, acute bronchitis, tuberculosis or COVID-19. In a multitude of cases, acute coughs are due to the common cold. According to the World Health Organization, respiratory tract infections are among the most important human health problems because of their high incidence and consequent economic costs.

Analyses have been carried out for both male and female persons belonging to various age-groups, and comparisons have been made between the situations of the urban and the rural people as well as between the situations of people belonging to North/Central Punjab and those belonging to South Punjab.

The purpose of this in-depth analysis is to **assist the relevant health departments** in formulating cough-reduction strategies specific to the various categories of people created by virtue of the analysis. Results of the analyses reveal important differences between various socio-economic classes of people and, as such, it is clear that steps aimed at reducing the prevalence of this infectious disease need to be in accordance with the requirements of the people belonging to the different classes.

2. LITERATURE REVIEW

It has been stated by Asma, Adnan, Hasnain, Shaper, Safee and Shahzad-ul-Hussan (2019) that the population above the 40 years of age is more on a risk to get infected. The most populated areas of the Punjab are on higher risk to get infected.

Faryatt and Bhuwanee (2015) stated in their study that The estimates originate from the analysis provide convincing advices that investments in health require to focus not only on direct service release but also on overall health-systems strengthening.

Strong health systems will be significant to achievement of SDG 3.

3. METHODOLOGY

The files pertaining to MICS 2003-2004, MICS 2007-2008, MICS 2011 and MICS 2014 data has been acquired from the Bureau of statistics, Punjab. Techniques of univariate and bivariate have been applied to analyzing the data. The MICS was originally developed in response to the World Summit for Children to measure progress towards an internationally agreed set of mid-decade goals. The first round of MICS was conducted around 1995 in more than 60 countries. The five rounds of MICS Punjab have been completed from 2003-2018.

4. DATA ANALYSIS

Area and Gender wise comparison of incidence of Cough in the Punjab province.

4.1 Analysis of MICS 2003-2004 Data

Rural, Urban or City Location * cough response Cross-tabulation

Count

		cough response		Total
		cough	no cough	
Area	Rural	1216986	49288296	50505282
		2.4%	97.6%	100%
	Other Urban	232163	9884316	10116479
		2.29%	97.71%	100%
	Major City	136993	9930596	10067589
		1.36%	98.64%	100%
Total		1586142	69103208	70689350
		2.24%	97.76%	100%

Sex of HH members * cough response Cross-tabulation

			cough response		Total
			cough	no cough	
Sex of HH members	male	Count	829482	35557756	36387238
		% within Sex of HH members	2.3%	97.7%	100.0%
	female	Count	756660	33543741	34300401
		% within Sex of HH members	2.2%	97.8%	100.0%
Total		Count	1586142	69101497	70687639
		% within Sex of HH members	2.2%	97.8%	100.0%

4.2 Analysis of MICS 2007-2008 Data

- Area wise comparison of incidence of Cough in The Punjab
- Gender wise Comparison with reference to Cough

Area * Had cough for last 3 weeks or more? Cross-tabulation

Count

		Had cough for last 3 weeks or more?			Total
		Yes	No	Don't know	
Area	Rural	8523	379623	1658	389804
		2.19%	97.39%	0.42%	100%
	All Urban	4460	199249	1338	205047
		2.18%	97.17%	0.65%	100%
Total		12983	578872	2996	594851
		2.18%	97.31%	0.50%	100%

Sex * Had cough for last 3 weeks or more? Cross-tabulation

			Had cough for last 3 weeks or more?		Total
			Yes	No	
Sex	Male	Count	6793	295954	302747
		% within Sex	2.2%	97.8%	100.0%
	Female	Count	6190	282918	289108
		% within Sex	2.1%	97.9%	100.0%
Total		Count	12983	578872	591855
		% within Sex	2.2%	97.8%	100.0%

Interpretation:

The table reveals that, in the MICS 2007-08 samples, the proportion of people affected by cough for last three weeks is slightly higher in male than in the female.

4.3 Analysis of MICS 2011 Data

- Area Wise Comparison of Incidence of Cough in The Punjab
- Gender wise Comparison with reference to Cough

Area of residence * Had been having Cough and Fever for last three weeks? Cross-tabulation

Count

		Had been having Cough and Fever for last three weeks?		Total
		Yes	No	
Area of residence	Rural	8796 2.46%	348403 97.54%	357199 100%
	All Urban	3492 1.45%	236866 98.55%	240358 100%
Total		12288 2.05%	585269 97.95%	597557 100%

Sex * Had been having Cough and Fever for last three weeks? Cross-tabulation

			Had been having Cough and Fever for last three weeks?		Total
			Yes	No	
Sex	Male	Count	6517	294709	301226
		% within Sex	2.2%	97.8%	100.0%
	Female	Count	5771	290560	296331
		% within Sex	1.9%	98.1%	100.0%
Total		Count	12288	585269	597557
		% within Sex	2.1%	97.9%	100.0%

4.4 Analysis of MICS 2014 Data

- Area wise Comparison of Incidence of Cough in The Punjab
- Gender wise Comparison with reference to Cough

Area * Had cough and fever for last 3 weeks Cross-tabulation

			Had cough and fever for last 3 weeks		Total
			Yes	No	
Area	Rural	Count	5447	151281	156728
		% within Area	3.5%	96.5%	100.0%
	All Urban	Count	2518	86907	89425
		% within Area	2.8%	97.2%	100.0%
Total		Count	7965	238188	246153
		% within Area	3.2%	96.8%	100.0%

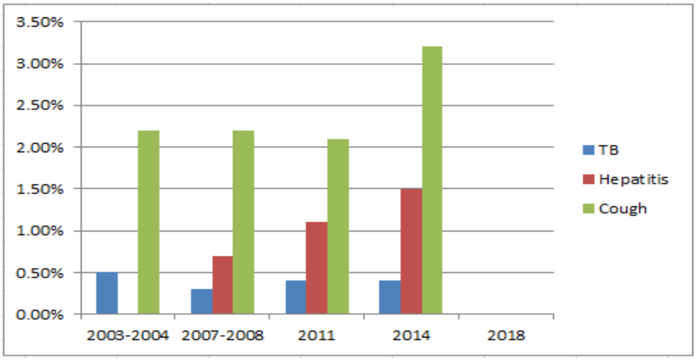
Sex * Had cough and fever for last 3 weeks Cross-tabulation

			Had cough and fever for last 3 weeks		Total
			Yes	No	
Sex	Male	Count	4113	120404	124517
		% within Sex	3.3%	96.7%	100.0%
	Female	Count	3852	117784	121636
		% within Sex	3.2%	96.8%	100.0%
Total		Count	7965	238188	246153
		% within Sex	3.2%	96.8%	100.0%

4.5 Comparison over Time

Table: Statistics for the persons having, Cough from 2003-2018:

MICS					
Variable	**2003-2004**	**2007-2008**	**2011**	**2014**	**2018**
Cough	2.2%	2.2%	2.1%	3.2%	-



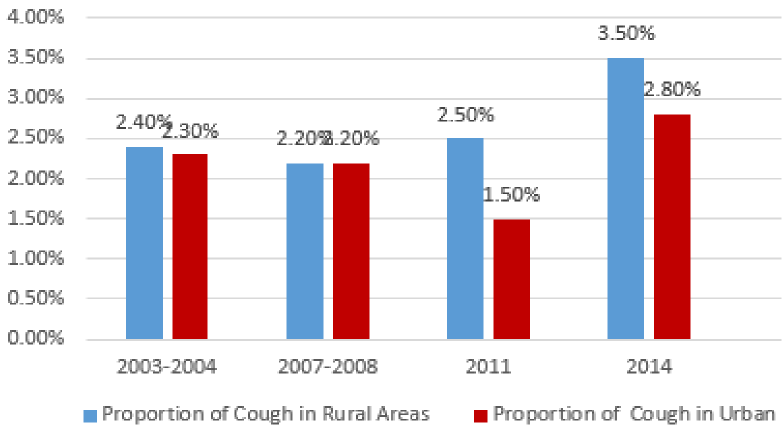
From the above table and multiple bar chart, we can observe that percentages of cough has increased in 12 years from 2003-14. Surprisingly, we found that MICS 2018 data files do not contain any variables pertaining to the incidence of Cough.

4.6 Area, gender and Income wise comparison with reference to health indicator cough in Punjab Pakistan over 12 years from 2003-2014.

Area wise comparison of the population having Cough

years	Proportion of Cough in **Rural Areas**	Proportion of Cough in **Urban Areas**	**difference**	**Combined proportion** of Urban and Rural areas who had Cough
2003-2004	2.4%	2.3%	2.1%	2.2%
2007-2008	2.2%	2.2%	0.0%	2.2%
2011	2.5%	1.5%	1.0%	2.1%
2014	**3.5%**	**2.8%**	0.7%	**3.2%**
2018	-	-	-	-

- The indicator of Cough was not included in the 2018 MICS data.



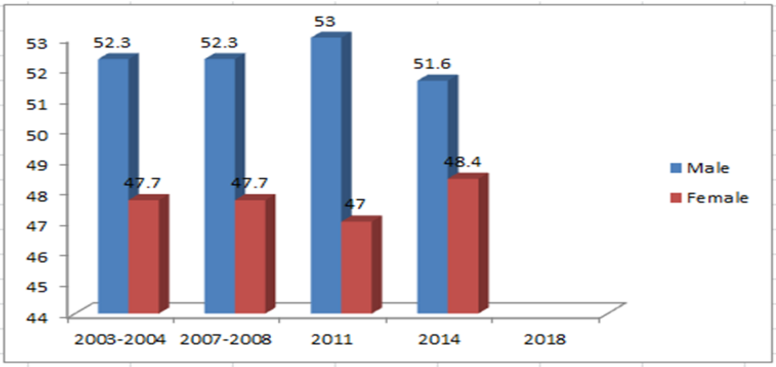
Interpretation:

In 2003-2004 the proportion of persons who diagnosed with cough in past one year was observed 2.4% in rural area that was higher as compared to the urban area (2.3%) with the difference of 2.1%. It was hopeful that the proportions were similar by 2.2% in both rural and urban areas of Punjab according to MICS 2007-2008. But in 2011 arise of 2.5% was observed in rural areas that was higher as compared to the urban area (1.5%) with the difference of 1.0%. In 2014 the proportion of persons diagnosed with cough in the past one year 3.5% in urban areas but the number of cases is decreased in rural areas by 2.8%.

Unfortunately the indicators to measure prevalence of cough who were not included in 2018, the data collected by other surveys indicates a worrying rise in the cases of cough patients in Punjab Pakistan.

If we compared the statistics of cough patients over the span of 15 years no improvements can be seen in rural areas and for urban areas. The statistic shows the condition is not satisfactory for the prevention of cough.

4.7 Gender wise comparison as having Cough in the Punjab Using the Previous Five Multiple Indicator Cluster Surveys:



It can be observed that the situation of the Punjab regarding having Cough across gender over the 15 years using five MICS surveys. It can be observed that there is no substantial difference between the percentages of female and male from 2003-2018.

4.8 Income wise comparison as having Cough in the Punjab Using the Previous Five Multiple Indicator Cluster Surveys:

- **Income wise comparison of the population having Cough:**

years	Proportion of persons diagnosed with Cough in **lowest quintile**	Proportion of persons diagnosed with Cough in **second quintile**	Proportion of persons diagnosed with Cough in **middle quintile**	Proportion of persons diagnosed with Cough in **fourth quintile**	Proportion of persons diagnosed with Cough in **highest quintile**	Proportion of persons diagnosed with Cough **combined in all wealth quintiles**
2003-2004	-	-	-	-	-	-
2007-2008	3.1%	2.5%	2.2%	1.8%	1.4%	2.2%
2011	3.1%	2.1%	2.1%	1.6%	1.0%	2.1%
2014	**4.2%**	**3.8%**	**3.5%**	**2.7%**	**2.0%**	**3.2%**
2018	-	-	-	-	-	-

- The indicator of Cough was not included in the 2018 MICS data.

It can be observed that the situation of the Punjab regarding having Cough across different income groups over the 15 years using five MICS surveys. It can be observed that there is substantial increase in the incidence of cough in the year 2014 between the percentages of different wealth index quintiles from 2003-2018.

5. COMMENTS AND CONCLUSION

After doing comparison, we conclude that with reference to SDG3, there is an increase in the incidence of cough which is a cause for alarm. All in all, although not much improvement is seen in the health sector during the twelve-year period considered in this study.

There is a need for strategic decision-making, policy-formulation and implementation of plans for the improvement of the health sector in the Punjab.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge the Bureau of Statistics, Punjab for providing the raw data files pertaining to MICS 2003-2004, MICS 2007-2008, MICS 2011, MICS 2014 and MICS 2018.

REFERENCES

1. Ahsan, A., Khan, A.Z., Javed, H., Mirza, S., Chaudhary, S.U. and Shahzad-ul-Hussan, S. (2019). Estimation of hepatitis C prevalence in the Punjab province of Pakistan: A retrospective study on general population. *PloS one*, 14(4), e0214435.
2. Bureau of Statistics (2003-04). *Multiple Indicator Cluster Survey*, Punjab, Planning and Development Department, Government of the Punjab.
3. Bureau of Statistics (2007-08). *Multiple Indicator Cluster Survey*, Punjab, Planning and Development Department, Government of the Punjab.
4. Bureau of Statistics (2014). *Multiple Indicator Cluster Survey*, Punjab, Planning and Development Department, Government of the Punjab.
5. Bureau of Statistics (2017-18). *Multiple Indicator Cluster Survey*, Punjab, Planning and Development Department, Government of the Punjab.
6. Hosseinpoor, A.R., Bergen, N., Schlotheuber, A. and Grove, J. (2018). Measuring Health Inequalities in the context of Sustainable Development Goals. *Bulletin of the World Health Organization*, 96(9), 654-659.
7. Shah, M.M. (2008). *Sustainable Development Encyclopedia of Ecology*. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/sustainable-development
8. Wikipedia. *Multiple Indicator Cluster Surveys*. Retrieved on July 31, 2020: https://en.wikipedia.org/wiki/Multiple_Indicator_Cluster_Surveys#:~:text=The%20Multiple%20Indicator%20Cluster%20Surveys,situation%20of%20children%20and%20women.
9. Unicef. *Multiple Indicator Cluster Surveys*. Retrieved on July 31, 2020: https://mics.unicef.org/faq
10. Tomas, H., Savatva, J., Bedrich, M. (2016). Sustainable Development Goals: A need for relevant indicators. *Journal & Book of Ecological Indicator*, 60, 506-573.
11. United Nations (2017). SDG Indicators Data collection Information & Focal points. Retrieved from: https://unstats.un.org/sdgs/dataContacts?selectIndicator=1.1.1+Proportion+of+population+below+the+international+poverty+line%2C+by+sex%2C+age%2C+employment+status+and+geographical+location+%28urban%2Frural%29&selectAgency=
12. Wikipedia (2017). Retrieved from: https://en.m.wikipedia.org/wiki/
13. World Bank (2012). *Press Books*. Retrieved from: https://slideplayer.com/slide/11697394/
14. World Health Organization (2015). *Environment Public Health* - Unit 3. Retrieved from: https://www.slideshare.net/jagdishbabu7/environment-public-health-unit-3
15. https://www.medicalnewstoday.com/articles/220349

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 381-391

ON THE SIA-NAKAGAMI-RATIO DISTRIBUTION AND ITS PROPERTIES

Nowara Farooq, Saleha Naghmi Habibullah and **Fatima Jamil**
Kinnaird College for Women, Lahore, Pakistan
Email: nowara57222@gmail.com
saleha.habibullah@kinnaird.edu.pk
fatima.jamil@kinnaird.edu.pk

ABSTRACT

In this paper, we derive the SIA-Nakagami-Ratio distribution which has two parameters and is more flexible than the SIU-Nakagami-Ratio distribution. Self-inversion of the density function is verified. Basic properties including the shape of the distribution, moments, moment generating function, quantile function, measures of central tendency, dispersion, skewness and kurtosis are presented. Hazard function and survival function are derived. The positively skewed shape of the distribution indicates its applicability to real-life data.

KEYWORDS

Self-Inverse distributions, Survival Function, Moment Generating Function, Quantile Function.

1. INTRODUCTION

Habibullah and Ahmad (2006) focused on inverted probability distributions that find applications in econometrics, biological sciences, life-testing, etc. and asserted that closure under inversion indicates that the multiplicative inverse of a continuous random variable X has the same probability function as that of the original random variable, permitting a possible change in parameter values. In this paper, an attempt was made to generate a class of distributions that are closed under inversion, and to obtain some statistical properties of this class of distributions. Jamil & Habibullah (2013) considered the distribution of the ratio X_1/X_2 where X_1 and X_2 are identically and independently distributed Nakagami random variables, thus obtaining a distribution that is self-inverse at unity ('SIU'). The pdf of the 'SIU-Nakagami-Ratio distribution' is given by

$$f(x)=\frac{2\Gamma(2m)}{\left(\Gamma(m)\right)^2}\frac{x^{2m-1}}{(x^2+1)^{2m}},0<x<\infty,m>0 \tag{1.1}$$

A number of properties of this single-parameter distribution were derived. SIA stands for "Self-Inverse at A" where A is the median of the distribution of a non-negative continuous random variable X for which X/A is distributed in exactly the same way as A/X. In this paper, we derive the SIA-Nakagami-Ratio distribution which is *more flexible* than the SIU-Nakagami-Ratio distribution due to the addition of the self-inversion parameter A. Basic properties of the distribution are derived. The positively skewed shape of the distribution attests to its applicability to real-world data.

2. LITERATURE REVIEW

In this section, we present a review of the literature regarding self-inverse distributions. Seshadri (1965) carried out a research study in which he explored the innate characteristics of non-negative random variables that have exactly the same distributions as their reciprocals. Saunders (1974) discussed the ξ –normal family of continuous distributions which can be regarded as a sub-class of the class of distributions invariant under the reciprocal transformation. Jones (2008) presented the concept of 'log-symmetry' defined as $\log Y - \log\theta \sim \log\theta - \log Y$ i.e. $Y/\theta \sim \theta/Y$ where Y is a positive continuous random variable and θ is the median of the distribution.

Jamil & Habibullah (2013) conducted a research study in which the ratio of two Nakagami random variables was used to obtain a new density function to which the authors assigned the name 'SIU-Nakagami-Ratio distribution'. The main properties of the distribution were derived. A simulation study based on 1000 samples of various sizes 'n' drawn from the SIU-Nakagami-Ratio distribution was carried out for different values of the parameter m. The distribution was applied to two real data-sets, one related to life-length and the other pertaining to successive failures of air-conditioning equipment in aircraft.

Habibullah and Jamil (2014) took up the self-inversion-based estimator of the parameter of the SIU-Nakagami-Ratio distribution and, by conducting a simulation study based on 1000 samples of various sizes drawn from this distribution, showed that the variance of this estimator is *less* than that of the estimator based on the simple mean.

Butt & Habibullah (2016) conducted a research-study on the SIA-log-student-t distribution. The property of being invariant under the reciprocal transformation was verified. Basic properties including the hazard function and the survival function were derived. Bowley's coefficient of skewness and percentile coefficient of kurtosis were derived by putting the values of its parameters.

Ali & Habibullah (2016) concentrated on the SIA-Log-Cauchy distribution. A simulation study was carried out based on different values of the scale parameter that showed that the coefficient of variation of a newly formulated SIA-estimator is less than the coefficient of variation of the corresponding non-SIA estimator implying that the SIA estimator is more efficient.

Habibullah (2017) focused on an interesting sub-class of the class of distributions of non-negative continuous random variables, those for which some researchers have used the nomenclature 'log-symmetric' distributions whereas some others have used the terminology 'Self-Inverse at A'. The author (2017) proposed that, since log-symmetric distributions are essentially SIA, the two nomenclatures be combined in order to adopt the nomenclature 'SIA log-symmetric distributions'. Applying the power-transformation to the log-symmetric density function $g(y)$, she (2017) obtained the class of 'SIA Log-Symmetric Power distributions' that possess greater flexibility than the SIA log-symmetric distributions due to the arbitrariness of the exponent.

3. DERIVATION OF THE SIA-NAKAGAMI-RATIO DISTRIBUTION

In this section, we drive the SIA-Nakagami-Ratio distribution. Applying the transformation $x = y/A$ to the pdf of the SIU-Nakagami-Ratio distribution given in equation (1.1), we have

$$g(y) = \frac{2\Gamma(2m)}{(\Gamma(m))^2}\left(\frac{y}{\mathrm{A}}\right)^{2m-1}\left(\left(\frac{y}{\mathrm{A}}\right)^2+1\right)^{-2m}\cdot\frac{1}{\mathrm{A}}$$

$$= \frac{2\mathrm{A}^{2\mathrm{m}}}{\mathrm{B}(m,m)}\,y^{2m-1}\left(y^2+\mathrm{A}^2\right)^{-2m},\ 0<y<\infty,\ m>0,\ \mathrm{A}>0$$

Since Y is a dummy variable, hence we can write

$$g(x) = \frac{2\mathrm{A}^{2\mathrm{m}}}{\mathrm{B}(m,m)}\,x^{2m-1}\left(x^2+\mathrm{A}^2\right)^{-2m},\ 0<x<\infty, m>0, \mathrm{A}>0 \tag{3.1}$$

Hence, the SIA-Nakagami-Ratio distribution is given by (3.1).

It is easy to show that the function $g(x)$ fulfills the properties of a probability density function.

4. SELF-INVERSION OF THE DENSITY FUNCTION

To verify the self-inversion of the density function of the distribution, we employ the functional equation given by

$$f(x) = \frac{\mathrm{A}^2}{x^2}\,f\left(\frac{\mathrm{A}^2}{x}\right) \tag{4.1}$$

Substituting from eq. (3.1) in the right-hand-side of eq. (4.1), we have

$$\text{R.H.S.} = \frac{\mathrm{A}^2}{x^2}\left(\frac{2\mathrm{A}^{2m}}{\mathrm{B}(m,m)}\left(\frac{\mathrm{A}^2}{x}\right)^{2m-1}\left(\left(\frac{\mathrm{A}^2}{x}\right)^2+\mathrm{A}^2\right)^{-2m}\right)$$

$$= \frac{2\mathrm{A}^{2m}}{\mathrm{B}(m,m)}\,x^{2m-1}\left(x^2+\mathrm{A}^2\right)^{-2m} = \text{L.H.S.}$$

5. SHAPE OF THE DISTRIBUTION

In this section, we discuss the shape of the SIA-Nakagami-Ratio distribution. Figure 1 contains the graph of the PDF of the SIA-Nakagami-Ratio distribution for different combinations of values of parameters m and A.

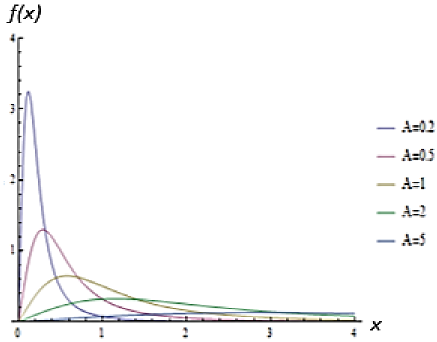


Figure 1(a): Graph of the PDF of the SIA-Nakagami-Ratio Distribution for m=1 and different Values of Parameter A

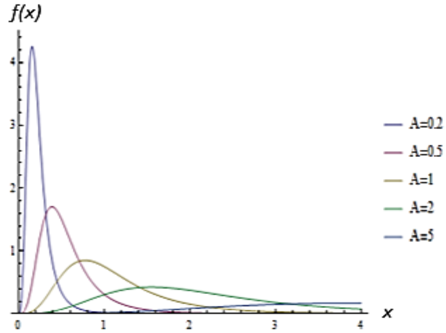


Figure 1(b): Graph of the PDF of the SIA-Nakagami-Ratio Distribution for m=2 and different Values of Parameter A

As is evident from Figure 1, the SIA-Nakagami-Ratio distribution is skewed to the right. We note that, for fixed *m*, the skewness decreases as A increases.

6. DISTRIBUTION FUNCTION, SURVIVAL FUNCTION AND HAZARD FUNCTION

The cumulative distribution function of the SIA-Nakagami-Ratio distribution is

$$\mathrm{G}(x)=\frac{x^{2m}}{\mathrm{A}^{2\mathrm{m}}m\mathrm{B}(m,m)}\,{}_2F_1\left[m,2m,1+m,-\frac{x^2}{\mathrm{A}^2}\right],\ 0<x<\infty,\ \mathrm{A}>0,\ m>0$$

where ${}_2F_1\left[m,2m,1+m,-\frac{x^2}{\mathrm{A}^2}\right]$ a Gaussian hypergeometric function.

The survival function of the SIA-Nakagami-Ratio distribution is given by

$$S(x)=1-G(x)$$
$$=\frac{\left(mB(m,m)-A^{-2m}x^{2m}\,{}_2F_1\left[m,2m,1+m,-\frac{x^2}{A^2}\right]\right)}{m\left(B(m,m)\right)}$$

As such, the hazard function of the SIA-Nakagami-Ratio distribution is given by

$$h(x)=\frac{2\mathrm{A}^{2\mathrm{m}}x^{2m-1}\left(x^2+\mathrm{A}^2\right)^{-2m}m}{m\mathrm{B}(m,m)-\mathrm{A}^{-2\mathrm{m}}x^{2m}\,{}_2F_1\left[m,2m,1+m,-\frac{x^2}{\mathrm{A}^2}\right]}$$

Figures 2, 3 and 4 contain, respectively, the graphs of the distribution function, the survival function and the hazard function of the SIA-Nakagami-Ratio distribution for different combinations of values of parameters m and A.

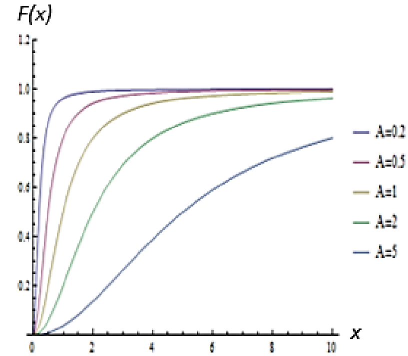


Figure 2(a): Graph of the CDF of the SIA-Nakagami-Ratio Distribution for *m*=1 and different Values of Parameter A

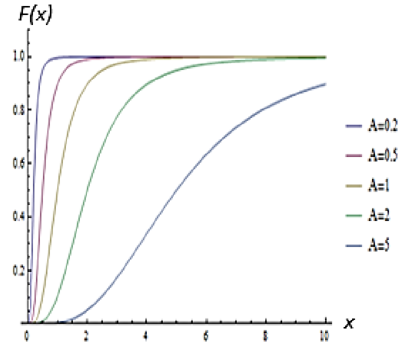


Figure 2(b): Graph of the CDF of the SIA-Nakagami-Ratio Distribution for *m*=2 and different Values of Parameter A

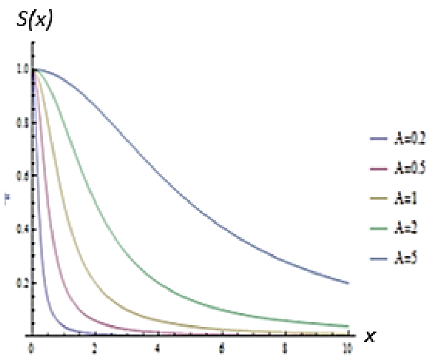


Figure 3(a): Graph of the Survival Function of the SIA-Nakagami-Ratio Distribution for *m*=1 and different Values of Parameter A

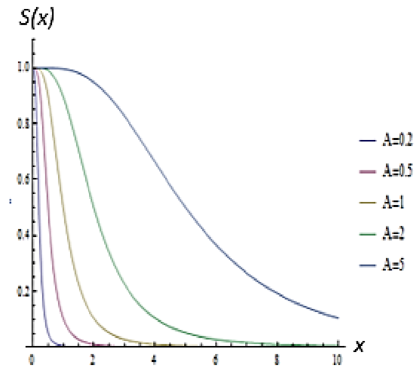


Figure 3(b): Graph of the Survival Function of the SIA-Nakagami-Ratio Distribution for *m*=2 and different Values of Parameter A

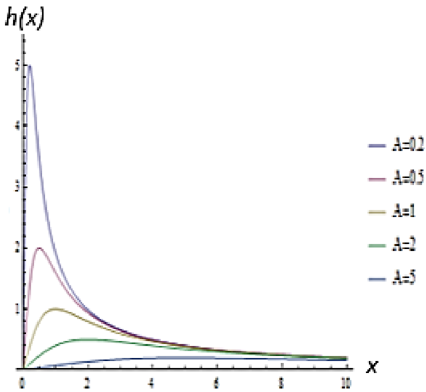


Figure 4(a): Graph of the Hazard Function of the SIA-Nakagami-Ratio Distribution for *m*=1 and different Values of Parameter A

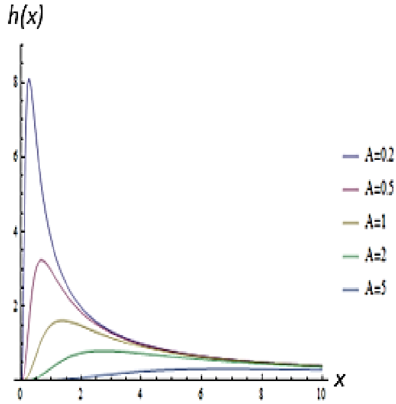


Figure 4(b): Graph of the Hazard Function of the SIA-Nakagami-Ratio Distribution for *m*=2 and different Values of Parameter A

7. MEASURES OF CENTRAL TENDENCY

In this section, we derive some measures of central tendency of the SIA-Nakagami-Ratio distribution in order to locate the general position of the distribution.

The mean of the SIA-Nakagami-Ratio distribution is

$$E(X)=\frac{\mathrm{A}\left(m-\frac{1}{2}\right)\left(\Gamma\left(m-\frac{1}{2}\right)\right)^{2}}{\left(\Gamma(m)\right)^{2}}\ ,m>\frac{1}{2} \tag{7.1}$$

The geometric mean of the SIA-Nakagami-Ratio distribution is turn out to be

$$\begin{aligned}\text{G.M.}&=\exp\left[\frac{2\mathrm{A}^{2\mathrm{m}}\Gamma(2m)}{\left(\Gamma(m)\right)^{2}}\left(\int_{0}^{\infty}\ln(x)\,x^{2m-1}\left(x^{2}+\mathrm{A}^{2}\right)^{-2m}dx\right)\right]\\&=\exp\left[-\frac{\mathrm{A}^{2\mathrm{m}}\Gamma(2m)}{\left(\Gamma(m)\right)^{2}}\frac{\mathrm{A}^{2\mathrm{m}}\mathrm{A}^{-4m}\left(\Gamma(m)\right)^{2}\ln\left(\mathrm{A}^{-2}\right)}{2\Gamma(2m)}\right]=\mathrm{A}\end{aligned} \tag{7.2}$$

The harmonic mean of the SIA-Nakagami-Ratio distribution is turn out to be

$$\text{H.M.}=\frac{\mathrm{A}\left(\Gamma(m)\right)^{2}}{\left(m-\frac{1}{2}\right)\left(\Gamma\left(m-\frac{1}{2}\right)\right)^{2}}\ ,m>\frac{1}{2} \tag{7.4}$$

It is very simple to verify that the arithmetic mean and the harmonic mean of the SIA-Nakagami-Ratio distribution have the following interesting relationship between them:

$$\frac{\text{H.M.}}{\text{A}} = \frac{\text{A}}{\text{A.M.}}$$

The median M of the SIA-Nakagami-Ratio distribution is obtained

$$\int_0^{\text{M}} g(x)dx = \frac{1}{2}$$

which yields

$$\text{M}^{2m} {}_2F_1\left(m, 2m, 1+m, -\frac{\text{M}^2}{\text{A}^2} \right) = \frac{1}{2} m\text{A}^{2m}\text{B}(m,m)$$

Putting $m = 1$, we obtain

$$\text{M} = \text{A} \tag{7.5}$$

To find the mode of the SIA-Nakagami-Ratio distribution, the first derivative of the density is obtained as follows:

$$g'(x) = \frac{2\text{A}^{2m}}{\text{B}(m,m)} \left(-4mx^{2m}\left(\text{A}^2 + x^2\right)^{-1-2m} + \left(-1+2m\right)x^{-2+2m}\left(\text{A}^2 + x^2\right)^{-2m} \right)$$

By putting $g'(x) = 0$, the mode of the distribution turns out to be.

$$\hat{x} = \text{A}\sqrt{\frac{(2m-1)}{(1+2m)}} \tag{7.6}$$

8. MOMENTS AND MOMENT GENERATING FUNCTION

In this section, the moments and moment ratios are derived. The r^{th} moment about origin of SIA-Nakagami-Ratio distribution are

$$\mu'_r = \frac{\text{A}^r\Gamma\left(m - \frac{r}{2}\right)\Gamma\left(m + \frac{r}{2}\right)}{\left(\Gamma(m)\right)^2}, \; m > \frac{r}{2}, \; r = 1, 2, 3, .. \tag{8.1}$$

Moving on to central moments, the moments about the mean of the SIA-Nakagami-Ratio distribution are

$$\mu_1 = 0$$

$$\mu_2 = \frac{A^2 m(\Gamma(m))^4 - A^2(m-1)\left(m-\frac{1}{2}\right)^2\left(\Gamma\left(m-\frac{1}{2}\right)\right)^4}{(m-1)(\Gamma(m))^4} \tag{8.2}$$

$$\mu_3 = \frac{A^3\left(m-\frac{1}{2}\right)\Gamma\left(m-\frac{1}{2}\right)}{(\Gamma(m))^2}\left(\left(m-\frac{3}{2}\right)\left(m+\frac{1}{2}\right) - \frac{3m\Gamma\left(m-\frac{1}{2}\right)}{(m-1)}\right)$$

$$+\frac{A^3\left(m-\frac{1}{2}\right)\Gamma\left(m-\frac{1}{2}\right)}{(\Gamma(m))^2}\left(\frac{2\left(m-\frac{1}{2}\right)^2\left(\Gamma\left(m-\frac{1}{2}\right)\right)^4}{(\Gamma(m))^4}\right) \tag{8.3}$$

$$\mu_4 = \frac{A^4}{(m-1)}\left(\frac{m(m+1)}{(m-2)} - \frac{4\Gamma\left(m-\frac{3}{2}\right)\left(m+\frac{1}{2}\right)\left(m-\frac{1}{2}\right)^2\left(\Gamma\left(m-\frac{1}{2}\right)\right)^3}{\Gamma(m-1)(\Gamma(m))^3}\right)$$

$$+\frac{A^4}{(m-1)}\left(\frac{6m\left(m-\frac{1}{2}\right)^2\left(\Gamma\left(m-\frac{1}{2}\right)\right)^4}{(\Gamma(m))^4}\right)$$

$$-\frac{A^4}{(m-1)}\left(\frac{3\left(m-\frac{1}{2}\right)^4\left(\Gamma\left(m-\frac{1}{2}\right)\right)^8}{\Gamma(m-1)(\Gamma(m))^7}\right) \tag{8.4}$$

The coefficient of variation (C.V.) turns out to be

$$\text{C.V.} = A^{-1}\left(m-\frac{1}{2}\right)^{-1}\left(\Gamma\left(m-\frac{1}{2}\right)\right)^{-2}\left(\frac{A^2 m(\Gamma(m))^4 - A^2\left(m-\frac{1}{2}\right)^2\left(\Gamma\left(m-\frac{1}{2}\right)\right)^4}{(m-1)}\right)^{\frac{1}{2}} \tag{8.5}$$

Moment ratios $\beta_1 \& \beta_2$ as well as the index of dispersion (γ) of SIA-Nakagami-Ratio distribution can easily be found using the moments μ_1', μ_2, μ_3 & μ_4 obtained above.

The moment generating function of the SIA-Nakagami-Ratio distribution is derived as follows:

$$\begin{aligned}
M_X(t) = E[e^{tx}] &= \int_0^\infty e^{tx} \frac{2\mathrm{A}^{2m}}{\mathrm{B}[m,m]} x^{2m-1}\left(x^2+\mathrm{A}^2\right)^{-2m} dx \\
&= {}_PF_Q\left[\{m\},\left\{\frac{1}{2},1-m\right\},-\frac{\mathrm{A}^2t^2}{4}\right] \\
&- \frac{\mathrm{A}^{2m}\pi(-t)^{2m}Csc(2m\pi)}{m\left(\Gamma(m)\right)^2}\,{}_PF_Q\left[\{2m\},\left\{\frac{1}{2}+m,1+m\right\},-\frac{\mathrm{A}^2t^2}{4}\right] \\
&+ \frac{\mathrm{A}2^{1-2m}\sqrt{\pi}t\Gamma\left(m-\frac{1}{2}\right)\Gamma(2m)}{m\left(\Gamma(m)\right)^3}\,{}_PF_Q\left[\left\{\frac{1}{2}+m\right\},\left\{\frac{3}{2},\frac{3}{2}-m\right\},-\frac{\mathrm{A}^2t^2}{4}\right] \qquad (8.6)
\end{aligned}$$

where $(1-m) \Rightarrow m<1$, $\left(m-\frac{1}{2}\right) \Rightarrow m>\frac{1}{2}$ and $\left(\frac{3}{2}-m\right) \Rightarrow m<\frac{3}{2}$.

9. QUANTILE FUNCTION

In this section, the quantile function and its related properties are derived. For $m = 1$, the quantile function of the SIA-Nakagami-Ratio distribution turns out to be

$$x_p = \mathrm{A}\sqrt{\frac{p}{(1-p)}} \qquad (9.1)$$

The lower, central and upper quartiles of the SIA-Nakagami-Ratio distribution are given by

$$Q_1 = \frac{1}{\sqrt{3}}\mathrm{A},\ Q_2 = \mathrm{A},\ Q_3 = \sqrt{3}\mathrm{A}$$

The expressions of the lower, central and upper deciles of the SIA-Nakagami-Ratio distribution are

$$D_1 = \frac{1}{3}\mathrm{A},\ D_5 = \mathrm{A},\ D_9 = 3\mathrm{A}$$

The lower, central and upper percentiles of the SIA-Nakagami-Ratio distribution

$$P_1 = \frac{1}{\sqrt{99}}\mathrm{A},\ P_{50} = \mathrm{A},\ P_{99} = \sqrt{99}\mathrm{A}$$

It is easy to see that

$$P_{50} = D_5 = Q_2 = \text{A}$$

and that the upper and lower quartiles, deciles and percentiles of the SIA-Nakagami-Ratio distribution fulfill the relationships

$$\frac{Q_1}{\text{A}} = \frac{\text{A}}{Q_3},\ \frac{D_1}{\text{A}} = \frac{\text{A}}{D_9},\ \frac{P_1}{\text{A}} = \frac{\text{A}}{P_{99}}.$$

Quantile-based measures of central tendency, dispersion, skewness and kurtosis are given below:

$$\text{Mid Quartile Range}\left(\text{M.Q.R.}\right) = \frac{Q_1 + Q_3}{2} = \frac{2\sqrt{3}\text{A}}{3} \tag{9.2}$$

$$\text{Inter Quartile Range }\left(\text{I.Q.R.}\right) = Q_3 - Q_1 = \frac{2\sqrt{3}\text{A}}{3} \tag{9.3}$$

$$\text{Semi-Interquartile Range }\left(\text{Q.D.}\right) = \frac{Q_3 - Q_1}{2} = \frac{\sqrt{3}\text{A}}{3} \tag{9.4}$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1}{2} \tag{9.5}$$

$$\text{Bowley's Coefficient of Skewness} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 2 - \sqrt{3} \tag{9.6}$$

$$\text{Percentile Coefficient of Kurtosis} = \frac{Q.D}{P_{90} - P_{10}} = \frac{8\sqrt{3}}{9} \tag{9.7}$$

10. CONCLUDING REMARKS

The SIU-Nakagami-Ratio distribution is moderately positively skewed indicating its applicability to real-world data-sets pertaining to non-negative random variables. In this paper, we derive the SIA version of the SIU-Nakagami-Ratio distribution in order to obtain a ***wider*** family of density functions through the introduction of the parameter A which makes the distribution applicable to right-skewed non-negative data-sets having different values of the median. The basic properties of the SIA-Nakagami-Ratio distribution are derived including measures of center, spread, skewness and kurtosis. Hazard and survival functions have also been obtained. In an upcoming study on the real-life application of the SIA-Nakagami-Ratio distribution, an attempt will be made to apply it to survival data pertaining to COVID-19.

REFERENCES

1. Ali, A. and Habibullah, S.N. (2016). On A Decile-Based SIA-Estimator of the Scale Parameter of the Log-Cauchy Distribution. In *Proceedings of the 14th International Conference on Statistical Sciences*, 29, 427-434.
2. Butt, A. and Habibullah, S.N. (2016). On the Derivation of the SIA-log-Student-t distribution for Achieving Better Fit, Proceedings of the *2nd International Multi-Disciplinary Conference (IMDC)*, University of Lahore, Gujrat Campus, Gujrat, Pakistan.
3. Habibullah, S.N. (2017). On the Power Generalization of Log-Symmetric SIA Distributions. *Journal of ISOSS*, 3(1), 45-54.
4. Habibullah, S.N. and Ahmad, M. (2006). On a New Class of Univariate Continuous Distributions that are Closed Under Inversion. *Pakistan Journal of Statistics and Operations Research*, 2, 151-158.
5. Habibullah, S.N. and Jamil, F. (2014). On the Variance of an Estimator of the Parameter of the SIU-Nakagami-Ratio Distribution. Presented at *Conference on Recent Advances in Mathematical Methods, Models & Applications (RAMMMA)* organized by the Centre for Mathematics & Statistical Sciences, Lahore School of Economics, Lahore, Pakistan on April 19-20, 2014.
6. Jamil, F. and Habibullah, S.N. (2013). On a Self-Inverse Probability Model Based on the Nakagami Distribution. Presented at *Kinnaird Multidisciplinary Research Conference (KMRC 2013)* held at Kinnaird College for Women, Lahore, Pakistan, (November 2013).
7. Jones, M.C. (2008). On reciprocal symmetry. *Journal of Statistical Planning and Inference*, 138(10), 3039-3043.
8. Saunders, S.C. (1974). A family of random variables closed under reciprocation. *Journal of the American Statistical Association*, 69(346), 533-539.
9. Habibullah, S.N., Memon, A.Z. and Ahmad, M. (2010). *On a Class of Distributions Closed under Inversion*, Lambert Academic Publishing (LAP), ISBN 978-3-8383-4868-1
10. Seshadri, V. (1965). On random variables which have the same distribution as their reciprocals. *Canadian Mathematical Bulletin*, 8(6), 819-824.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 393-408

ON SOME PROPERTIES OF A NEW PROBABILITY MODEL BASED ON LINEAR COMBINATION OF TWO SELF-INVERSE DENSITIES

Iqra Mushtaq, Saleha Naghmi Habibullah and **Fatima Jamil**
Kinnaird College for Women, Lahore, Pakistan
Email: iqramushtaq36@outlook.com
saleha.habibullah@kinnaird.edu.pk
fatima.jamil@kinnaird.edu.pk

ABSTRACT

In this paper, we derive a new density function based on the linear combination of two SIA distributions i.e. the SIA Log-Logistic Distribution and the SIA Inverse-Gamma Ratio distribution. Self-inversion of the density function is verified. Basic properties including the shape of the distribution, hazard function, moments, quantile function, measures of central tendency, dispersion, skewness and kurtosis are presented. The shape of the newly derived density function points to the viability of the distribution as a suitable model for real data pertaining to non-negative continuous random variables.

KEYWORDS

SIA Log-Logistic distribution, SIA Inverse-Gamma Ratio distribution**.** Linear Combination, Hazard function, Moments, Quantile Function

1. INTRODUCTION

Habibullah and Saunders (2011) introduced the term 'self-inverse at unity' for those probability density functions for which the pdf of the multiplicative inverse is identical to that of the original random variable. Habibullah (2012) adopted the abbreviation SIU to describe the term 'Self-Inverse at Unity'. Habibullah and Fatima (2015) introduced the term 'Self Inverse at A (SIA)' for a *more general* class of self-inverse distributions. Given a non-negative continuous random variable Y, this class contains distributions for which Y/A is distributed exactly as A/Y where A is the 50th percentile of the distribution.

In this research study, we obtain a new probability model based on the linear combination of two SIA density functions i.e. the SIA Log-Logistic Distribution and the SIA Inverse-Gamma Ratio distribution. We show that the newly obatained density also belongs to the class of self-inverse distributions. Fundamental properties such as the shape of the distribution, the hazard function, quantile function, moments. measures of centre, spread, skewness and kurtosis are derived.

The graphical form of the newly obtained probability model testifies to its usefulness for modeling real-life data.

2. LITERATURE REVIEW

In this section, we present a review of the literature with reference to (i) the class of self-inverse distributions and (ii) probability density functions based on linear combinations of random variables.

Seshadri (1965) focused on non-negative continuous random variables for which the distribution reciprocal is the same as that of the original random variable. One of the fundamental properties of this class of distributions is that the $(1-p)^{th}$ quantile is the multiplicative inverse of the p^{th} quantile, and the median is equal to unity. Habibullah and Ahmad (2006) considered inverted probability distributions and focused on those inverted distributions that are identical to the original distributions. The authors (2006) regarded such distributions as being Strictly Closed Under Inversion (SCUI) and attempted to generate a class of such distributions. Some statistical properties of this class were obtained. Jones (2008) concentrated on the class of distributions of non-negative continuous random variables X for which the distribution of X/median is the same as the distribution of median/X. He (2008) adopted the term 'log-symmetric' for such distributions. Habibullah et al. (2010) formulated two forms of differential equations for the generation of SCUI density functions. Habibullah and Saunders (2011) presented the term 'self-inverse at unity' for this type of distributions and Habibullah (2012) advocated the abbreviation SIU to describe the term 'Self-Inverse a Unity'. Vanegas and Paula (2016) studied the main statistical properties of the class of log-symmetric distributions. The authors (2016) obtained quantile-based measures of location, dispersion, skewness and kurtosis for the log-symmetric class that are applicable in the case of asymmetric and heavy-tailed distributions. The advantageousness of the log-symmetric class of distributions was illustrated through a statistical analysis of a real dataset, in which the performance of the log-symmetric class was examined as compared with that of some competitive and very flexible distributions.

Focusing on distributions based on linear combinations, Nadarajah (2007) showed that the distributions of linear combinations, products and ratios of random variables arise in many areas of engineering. Nadarajah and Kotz (2007) considered Laplace and logistic random variables, X and Y respectively, distributed independently of each other, and derived the exact distribution of the linear combination $\alpha X + \beta Y$. Extensive tabulations of the associated percentage points obtained by inverting the derived distribution are also given in this paper. Marques et al. (2015) asserted that the distribution of linear combinations of independent Gumbel random variables is of great interest for modeling risk and extremes in the most different areas of application. Near-exact approximations for the distribution of linear combination of independent Gumbel random variables were developed based on a shifted generalized near-integer gamma distribution and on the distribution of the difference of two independent generalized integer gamma distributions. The proposed approximations were illustrated on applied problems in various fields.

3. DERIVATION OF A NEW SELF-INVERSE DISTRIBUTION

In this section, we derive a new density function through a linear combination of two self-inverse distributions i.e. the SIA Log-Logistic Distribution and the SIA Inverse-Gamma Ratio distribution.

The pdf of the SIA Log-Logistic Distribution is given by

$$f_1(x) = bA^b x^{b-1}\left(x^b + A^b\right)^{-2}, \quad 0 < x < \infty, \, A > 0, b > 0 \tag{3.1}$$

The median is A and the mean is given by

$$E(X) = \frac{A(\pi / b)}{\sin(\pi / b)}, b > 1$$

On the other hand, the pdf of the SIA Inverse-Gamma Ratio distribution is given by

$$f_2(x) = \frac{A^c}{B(c,c)} x^{c-1}(x + A)^{-2c}, 0 < x < \infty, c > 0, A > 0 \tag{3.2}$$

The median of this distribution too is A whereas the mean is given by

$$E(X) = \frac{Ac}{c-1}, \, c > 1$$

Taking linear combination of the two pdfs, we obtain

$$\begin{aligned} f(x) &= dbA^b x^{b-1}\left(x^b + A^b\right)^{-2} + (1-d)\frac{A^c}{B(c,c)} x^{c-1}(x + A)^{-2c} \\ &= \frac{1}{B(c,c)}\left[dbB(c,c)A^b x^{b-1}\left(x^b + A^b\right)^{-2} + (1-d)A^c x^{c-1}(x + A)^{-2c}\right], \\ &\qquad 0 < x < \infty, \; A > 0, b > 0, c > 0, 0 < d < 1 \end{aligned} \tag{3.3}$$

It is easy to show that area under the curve of the four-parameter density function (3.3) is unity.

Next, to show that the newly derived density function is also self-inverse, we consider the functional equation of an SIA distribution given by

$$f(x) = \frac{A^2}{x^2} f\left(\frac{A^2}{x}\right) \tag{3.4}$$

Substituting from eq. (3.3) in the right-hand-side of eq. (3.4), we have

$$\text{R.H.S.}=\frac{A^2}{x^2 B(c,c)}\left[\begin{array}{r} bdB(c,c)A^b\left(\frac{A^2}{x}\right)^{b-1}\left(\left(\frac{A^2}{x}\right)^b+A^b\right)^{-2} \\ +(1-d)A^c\left(\frac{A^2}{x}\right)^{c-1}\left(\frac{A^2}{x}+A\right)^{-2c} \end{array}\right]$$

$$=\frac{1}{B(c,c)}\left[\begin{array}{r} bdB(c,c)A^{b-2+2}x^{b+c-c-1}\left(A^b+x^b\right)^{-2} \\ +(1-d)A^{c-2+2}x^{b+c-b-1}(A+x)^{-2c} \end{array}\right]$$

$$=\frac{1}{B(c,c)}\left[\begin{array}{c} bdB(c,c)A^b x^{b-1}\left(A^b+x^b\right)^{-2}+ \\ (1-d)A^c x^{c-1}(A+x)^{-2c} \end{array}\right]=f(x)=\text{L.H.S.}$$

Hence, it is clear that the density is Self-Inverse at A.

4. SHAPE OF THE NEWLY DERIVED DISTRIBUTION

For any newly derived distribution, one of the first points that one needs to consider is the shape of the distribution. Figures 4.1(a) and 4.1(b) display the shape of the density function based on linear combination of the SIA Log-Logistic and the SIA Inverse-Gamma Ratio distribution. The positively skewed shape attests the applicability of the distribution to data-sets of non-negative variables such as life-length, income, etc.

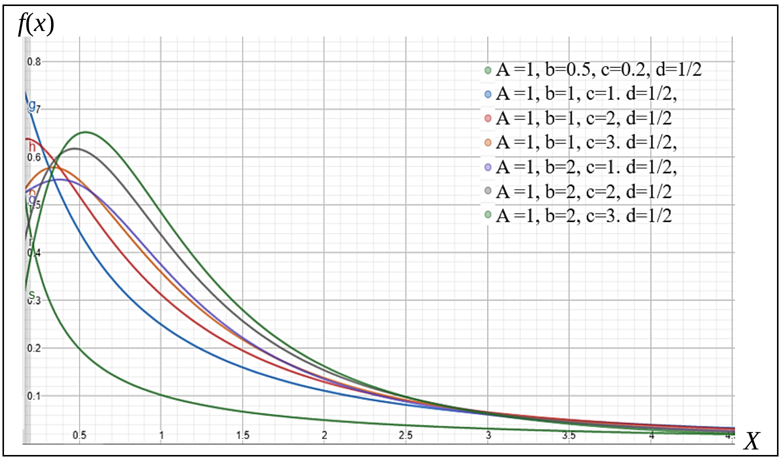


Figure 4.1(a) Shape of the PDF when A is Fixed and b,c,d are having Different Values

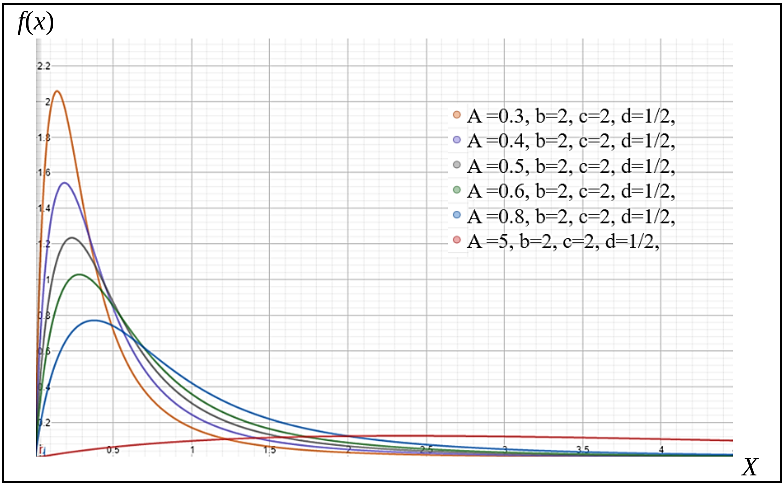


Figure 4.1(b) Shape of the PDF when b, c, d are Fixed and A is having Different Values

5. CUMULATIVE DISTRIBUTION FUNCTION & ITS SHAPE

The CDF of the newly obtained distribution is given by

$$F(x) = \int_0^x f(x)dx = \left(\frac{d\left(x^b\right)}{\left(A^b + x^b\right)} \right) + \left(\frac{(1-d)A^{-c}x^c \, {}_2F_1\left[c, 2c, 1+c, -\frac{x}{A}\right]}{cB\left(c,c\right)} \right)$$

$$= \frac{cB\left(c,c\right)dx^b + \left(A^b + x^b\right)(1-d)A^{-c}x^c \, {}_2F_1\left[c, 2c, 1+c, -\frac{x}{A}\right]}{cB\left(c,c\right)\left(A^b + x^b\right)} \tag{5.1}$$

where ${}_2F_1\left[c, 2c, 1+c, -\frac{x}{\mathrm{A}}\right]$ is the hypergeometric function.

Figures 5.1(a) and 5.1(b) present the shape of the cumulative distribution function for various combinations of values of the four parameters.

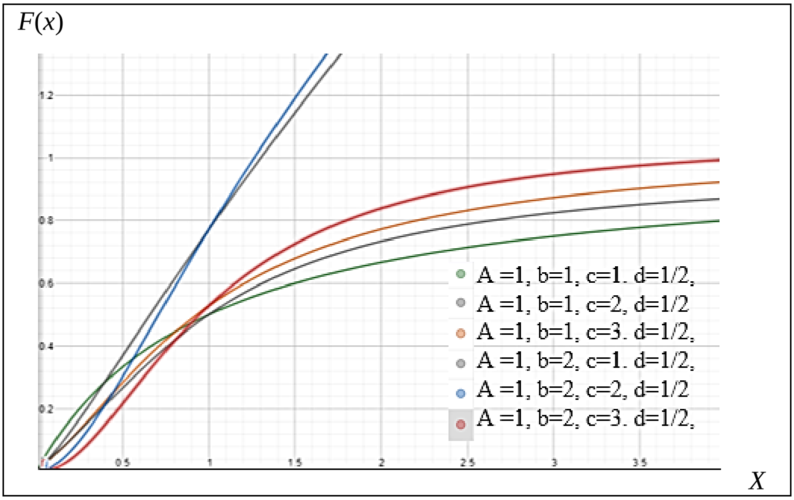


Figure 5.1(a) Shape of the CDF when A is Fixed and b,c,d are having Different Values

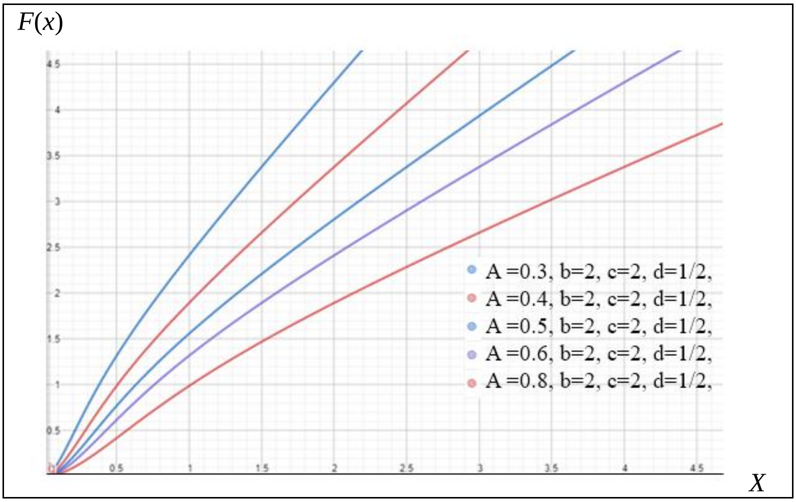


Figure 5.1(b) Graph of the CDF when b, c, d are Fixed and A is having Different Values

6. RELIABILITY-RELATED PROPERTIES OF THE NEWLY DERIVED DISTRIBUTION

In this section, we present some reliability-related properties of the newly derived density function.

The reliability function of the newly derived distribution is given by

$$R(x)=\frac{\begin{array}{l} c\mathrm{B}(c,c)\left(\mathrm{A}^{b}+x^{b}\right)-c\mathrm{B}(c,c)dx^{b} \\ \qquad +\left(\mathrm{A}^{b}+x^{b}\right)(1-d)\mathrm{A}^{-c}x^{c}\ {}_{2}F_{1}\left[c,2c,1+c,-\frac{x}{\mathrm{A}}\right] \end{array}}{c\mathrm{B}(c,c)\left(\mathrm{A}^{b}+x^{b}\right)} \tag{6.1}$$

The shape of the reliability function for various combinations of values of the four parameters is presented in Figures 6.1(a) and 6.1(b).

The hazard function of the newly derived distribution is given by

$$h(x)=\frac{c\left(\mathrm{A}^{b}+x^{b}\right)\left[db\mathrm{B}(c,c)\mathrm{A}^{b}x^{b-1}\left(x^{b}+\mathrm{A}^{b}\right)^{-2}+(1-d)\mathrm{A}^{c}x^{c-1}\left(x+\mathrm{A}\right)^{-2c}\right]}{\begin{array}{l} c\mathrm{B}(c,c)\left(\mathrm{A}^{b}+x^{b}\right)-c\mathrm{B}(c,c)dx^{b} \\ \qquad +\left(\mathrm{A}^{b}+x^{b}\right)(1-d)\mathrm{A}^{-c}x^{c}\ {}_{2}F_{1}\left[c,2c,1+c,-\frac{x}{\mathrm{A}}\right] \end{array}}$$

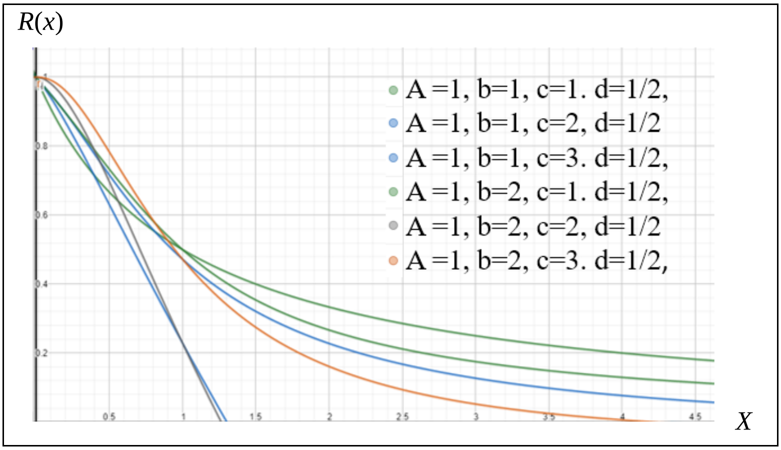


Figure 6.1(a) Shape of the Reliability Function when A is Fixed and b,c,d are having Different Values

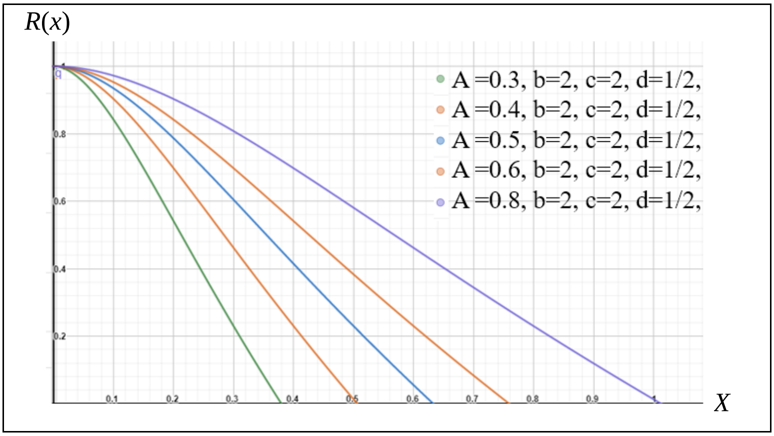


Figure 6.1(b) Shape of the Reliability Function when b, c, d are Fixed and A is having Different Values

The shape of the hazard function for various combinations of values of the four parameters is presented in Figures 6.2(a) and 6.2(b).

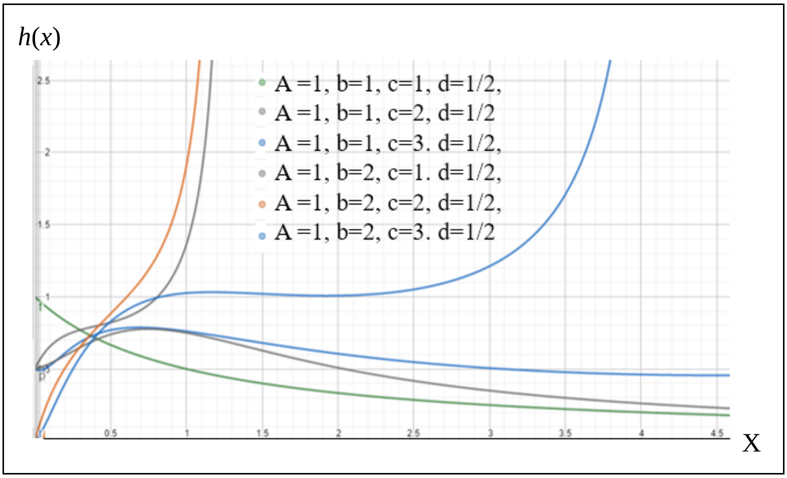


Figure 6.2(a) Shape of the Hazard Rate when A is Fixed and b,c,d are having Different Values

7. QUANTILE FUNCTION

In this section, we derive the quantile function of the newly derived distribution. We know that, in general,

$$F(x_p) = p \Rightarrow F^{-1}\left(F(x_p)\right) = F^{-1}(p) \Rightarrow x_p = F^{-1}(p)$$

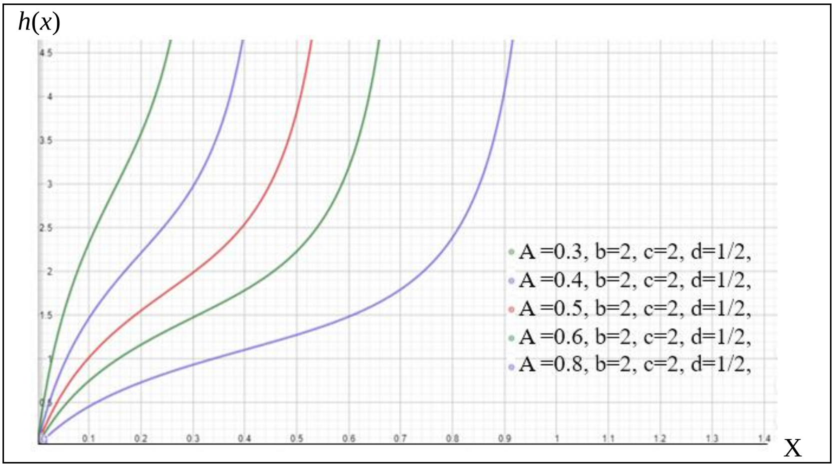


Figure 6.2(b) Shape of the Hazard Rate when b, c, d are Fixed and A is having Different Values

So, for density function (3.3), we have

$$\frac{dc\mathrm{B}(c,c)\,x_p^b + \left(\mathrm{A}^b + x_p^b\right)(1-d)\mathrm{A}^{-c}x_p^c\;{}_2F_1\left[c,2c,1+c,-\dfrac{x_p}{\mathrm{A}}\right]}{c\mathrm{B}(c,c)\left(\mathrm{A}^b + x_p^b\right)} = p$$

$$\Rightarrow dx_p^b + \left(\mathrm{A}^b + x_p^b\right)(1-d)\mathrm{A}^{-1}x_p\;{}_2F_1\left[1,2,2,-\frac{x_p}{\mathrm{A}}\right] = \left(\mathrm{A}^b + x_p^b\right)p$$

$$\Rightarrow x_p = p\left(\mathrm{A}+x_p\right)$$

As such, we obtain:

$$x_p = \frac{p\mathrm{A}}{(1-p)} \tag{7.1}$$

Hence, the lower, central and upper quartiles of density function (3.3) are given by

$$Q_1 = \frac{1}{3}\mathrm{A},\ Q_2 = \mathrm{A},\ Q_3 = 3\mathrm{A}$$

The lower, central and upper deciles of density function (3.3) are given by

$$D_1 = \frac{1}{9}\mathrm{A},\ D_5 = \mathrm{A},\ D_9 = 9\mathrm{A}$$

The lower, central and upper deciles of density function (3.3) are given by

$$P_1 = \frac{\mathrm{A}}{99},\ P_{50} = \mathrm{A},\ P_{99} = 99\mathrm{A}$$

It is obvious that $P_{50} = D_5 = Q_2 = \mathrm{A}$ and that the upper and lower quartiles, deciles and percentiles of density function (3.3) fulfill the relationships

$$\frac{Q_1}{\mathrm{A}}=\frac{\mathrm{A}}{Q_3},\ \frac{D_1}{\mathrm{A}}=\frac{\mathrm{A}}{D_9},\ \frac{P_1}{\mathrm{A}}=\frac{\mathrm{A}}{P_{99}}.$$

8. MEASURES OF CENTRAL TENDENCY AND DISPERSION

In this section, we present a few measures of central tendency and dispersion of the newly derived density function.

The expression of arithmetic mean is derived as

$$A.M.=E(X)=d\int_0^\infty \frac{xbA^b x^{b-1}}{\left(x^b+A^b\right)^2}dx+\left(1-d\right)\int_0^\infty \frac{xA^c x^{c-1}\left(x+A\right)^{-2c}}{B\left(c,c\right)}dx$$

$$=\frac{A\left(c-1\right)d\left(\pi/b\right)+A\left(c\right)\left(1-d\right)\sin\left(\pi/b\right)}{\left(c-1\right)\sin\left(\pi/b\right)} \tag{8.1}$$

The geometric mean is given by $\mathrm{G.M.}=\mathrm{e}^{\mathrm{E}(\ln x)}$. Therefore

$$\ln\left(G.M.\right)=E\left[\ln x\right]$$

$$=\frac{1}{B\left(c,c\right)}\int_0^\infty\left(\ln x\left[dbB\left(c,c\right)A^b x^{b-1}\left(x^b+A^b\right)^{-2}+(1-d)A^c x^{c-1}\left(x+A\right)^{-2c}\right]\right)dx$$

$$=dbA^b\left(-\frac{A^{-b}\log\left(A^{-b}\right)}{b^2}\right)+\frac{(1-d)A^c}{B\left(c,c\right)}\left(\frac{A^{-c}\log\left(A\right)\left(\Gamma\left(c\right)\right)^2}{\Gamma(2c)}\right)$$

$$=\frac{1}{b}\left[-d\log\left(A^{-b}\right)+b(1-d)\log\left(A\right)\right]=\frac{b}{b}\log\left(A\right)=\log\left(A\right)$$

implying that

$$G.M.=A \tag{8.2}$$

The Harmonic Mean is derived as follows:

$$H.M.=\left(E\left(\frac{1}{X}\right)\right)^{-1}=\left(\begin{array}{l} d\int_0^\infty \frac{1}{x}bA^b x^{b-1}\left(x^b+A^b\right)^{-2}dx \\ \qquad +\left(1-d\right)\int_0^\infty \frac{1}{x}\frac{A^c x^{c-1}\left(x+A\right)^{-2c}}{B\left(c,c\right)}dx\end{array}\right)^{-1}$$

$$=\left(\frac{A^{-1}d\left(\pi/b\right)}{\sin\left(\pi/b\right)}+\frac{A^{-1}c\left(1-d\right)}{\left(c-1\right)}\right)^{-1}$$

$$=\frac{A\left(c-1\right)\sin\left(\pi/b\right)}{\left(c-1\right)d\left(\pi/b\right)+c\left(1-d\right)\sin\left(\pi/b\right)} \tag{8.3}$$

It is easy to see that the harmonic and arithmetic means are related by the equation $\frac{A.M}{A}=\frac{A}{H.M}$

The Median of the newly derived density function (3.3) is obtained as follows:

$$\int_0^{\mathrm{M}} \frac{db\mathrm{B}(c,c)\mathrm{A}^b x^{b-1}\left(x^b+\mathrm{A}^b\right)^{-2}+(1-d)\mathrm{A}^c x^{c-1}\left(x+\mathrm{A}\right)^{-2c}}{\mathrm{B}(c,c)}dx=\frac{1}{2}$$

$$\Rightarrow \frac{d\mathrm{M}^b}{\left(\mathrm{A}^b+\mathrm{M}^b\right)}+\frac{(1-d)\mathrm{A}^{-c}\mathrm{M}^c\,{}_2F_1\left[c,2c,1+c,-\frac{\mathrm{M}}{\mathrm{A}}\right]}{c\mathrm{B}(c,c)}=\frac{1}{2}$$

$$\Rightarrow \frac{\mathrm{M}}{\left(\mathrm{A}+\mathrm{M}\right)}=\frac{1}{2}\Rightarrow \mathrm{M}=\mathrm{A} \tag{8.4}$$

We can see that the median and geometric means of the newly derived SIA distribution are the same.

Focusing on dispersion, the algebraic expression of the variance is derived as follows:

$$\mathrm{Var}(X)=E(X^2)-\left(E(X)\right)^2$$

where

$$E\left(X^2\right)=\int_0^{\infty} x^2 f(x)=bd\mathrm{A}^b\int_0^{\infty} x^{b+1}\left(x^b+\mathrm{A}^b\right)^{-2}dx+\frac{(1-d)\mathrm{A}^c}{B(c,c)}\int_0^{\infty} x^{c+1}\left(x+\mathrm{A}\right)^{-2c}dx$$

$$=\frac{\mathrm{A}^2(2\pi/b)(c-1)(c-2)d+\mathrm{A}^2(c)(c+1)(1-d)\sin(2\pi/b)}{\sin(2\pi/b)(c-1)(c-2)} \tag{8.5}$$

Hence

$$Var(X)=\left(\frac{A^2d(2\pi/b)}{\sin(2\pi/b)}+\frac{A^2(c)(c+1)(1-d)}{(c-1)(c-2)}\right)-\left(\frac{Ad(\pi/b)}{\sin(\pi/b)}+\frac{A(c)(1-d)}{(c-1)}\right)^2$$

$$=\frac{A^2d\left[(2\pi/b)\sin^2(\pi/b)-d(\pi/b)^2\sin(2\pi/b)\right]}{\sin(2\pi/b)\sin^2(\pi/b)}$$

$$+\frac{A^2c(1-d)\left[\sin(\pi/b)\left(c^2d-1+2c(1-d)\right)-2cd(\pi/b)\left(c^2-3c+2\right)\right]}{\sin(\pi/b)(c-1)^2(c-2)}$$

$$=\frac{1}{(c-2)(c-1)^2\sin(2\pi/b)\sin^2(\pi/b)}$$

$$\left\{\mathrm{A}^2d(c-1)^2(c-2)\left[(2\pi/b)\sin^2(\pi/b)-d(\pi/b)^2\sin(2\pi/b)\right]\right.$$

$$+\mathrm{A}^2c(1-d)\sin(2\pi/b)\sin(\pi/b)$$

$$\left.\left[\sin(\pi/b)\left(c^2d-1+2c(1-d)\right)-2cd(\pi/b)\left(c^2-3c+2\right)\right]\right\} \tag{8.6}$$

9. MOMENTS AND MOMENT RATIOS

In this section, we present the algebraic expressions of the raw moments, moments about the mean and moment-ratios for the newly derived probability density function.

The r^{th} moment of the distribution is given by

$$\mu_r' = d\int_0^\infty x^r b\mathrm{A}^b x^{b-1}\left(x^b + \mathrm{A}^b\right)^{-2} dx + (1-d)\int_0^\infty \frac{x^r \mathrm{A}^c x^{c-1}(x+\mathrm{A})^{-2c}}{\mathrm{B}(c,c)} dx$$

$$= d\left(\frac{\mathrm{A}^{-b+r+b}\pi r}{b\sin\left(\frac{\pi r}{b}\right)}\right) + (1-d)\left(\frac{\mathrm{A}^r\Gamma(c-r)\Gamma(c+r)}{(\Gamma(c))^2}\right)$$

Hence the expression to find the r^{th} moment about origin is

$$\mu_r' = \frac{\mathrm{A}^r(\pi r/b)(\Gamma(c))^2 d + \mathrm{A}^r(1-d)\Gamma(c-r)\Gamma(c+r)\sin(\pi r/b)}{\sin(\pi r/b)(\Gamma(c))^2} \tag{9.1}$$

Putting $r = 1$ in equation (9.1), we obtain the expression of the mean given in (8.1); putting $r = 2$ in equation (9.1), we obtain the expression of the mean given in (8.5).

Putting $r = 3$ in equation (9.1), we obtain

$$\mu_3' = \frac{\mathrm{A}^3 d(3\pi/b)(\Gamma(c))^2 + \mathrm{A}^3(1-d)\Gamma(c-3)\Gamma(c+3)\sin(3\pi/b)}{\sin(3\pi/b)(\Gamma(c))^2}$$

$$= \frac{\mathrm{A}^3(3\pi/b)(c-3)(c-2)(c-1)d + \mathrm{A}^3(1-d)(c)(c+1)(c+2)\sin(3\pi/b)}{\sin(3\pi/b)(c-3)(c-2)(c-1)} \tag{9.2}$$

Putting put $r = 4$ in equation (9.1), we obtain

$$\mu_4' = \frac{\mathrm{A}^4 d(4\pi/b)(\Gamma(c))^2 + \mathrm{A}^4(1-d)\Gamma(c-4)\Gamma(c+4)\sin(4\pi/b)}{\sin(4\pi/b)(\Gamma(c))^2}$$

$$= \frac{\mathrm{A}^4 d(4\pi/b)}{\sin(4\pi/b)} + \frac{\mathrm{A}^4(1-d)(c)(c+1)(c+2)(c+3)}{(c-4)(c-3)(c-2)(c-1)}$$

$$= \frac{\mathrm{A}^4(4\pi/b)(c-4)(c-3)(c-2)(c-1)d + \mathrm{A}^4(1-d)(c)(c+1)(c+2)(c+3)\sin(4\pi/b)}{\sin(4\pi/b)(c-4)(c-3)(c-2)(c-1)} \tag{9.3}$$

The first four moments about the mean are obtained by using the relationships between moments about the mean and moments about the origin:

For 1st moment about mean, we know that $\mu_1 = 0$; for 2nd moment about mean, we obtain the expression of the variance given in (8.6).

Now, for the 3rd moment about mean, by substitution in the relationship between the third moment about the mean and the raw moments. We obtain:

$$
\begin{aligned}
\mu_3 = {} & \frac{A^3(3\pi/b)(c-3)(c-2)(c-1)d + A^3(1-d)(c)(c+1)(c+2)\sin(3\pi/b)}{\sin(3\pi/b)(c-3)(c-2)(c-1)} \\
& -3\left(\frac{A^2(2\pi/b)(c-1)(c-2)d + A^2(c)(c+1)(1-d)\sin(2\pi/b)}{\sin(2\pi/b)(c-1)(c-2)}\right) \\
& \left(\frac{A(\pi/b)(c-1)d + A(c)(1-d)\sin(\pi/b)}{\sin(\pi/b)(c-1)}\right) \\
& +2\left(\frac{A(\pi/b)(c-1)d + A(c)(1-d)\sin(\pi/b)}{\sin(\pi/b)(c-1)}\right)^3
\end{aligned}
$$

Solving the above expression, the 3rd moment about mean turns out to be

$$
\begin{aligned}
\mu_3 = {} & \frac{1}{\sin(3\pi/b)\sin(2\pi/b)\sin^3(\pi/b)(c-1)^3(c-2)(c-3)} \\
& \left\{ \begin{aligned} & A^3(c-1)^2\sin(2\pi/b)\sin^3(\pi/b) \\ & \qquad \left[\begin{aligned} & (3\pi/b)(c-3)(c-2)(c-1)d \\ & \quad +(1-d)(c)(c+1)(c+2)\sin(3\pi/b) \end{aligned}\right] \end{aligned} \right\} \\
& -\left\{ \begin{aligned} & 3A^3(c-1)^2(c-3)d(\pi/b)\sin(3\pi/b)\sin^2(\pi/b) \\ & \qquad \left[d(c-1)(c-2)(2\pi/b) + c(c+1)(1-d)\sin(2\pi/b)\right] \end{aligned} \right\} \\
& -\left\{ \begin{aligned} & 3A^3c(c-1)(c-3)(1-d)\sin^3(\pi/b)\sin(3\pi/b) \\ & \qquad \left[c(c+1)(1-d)\sin(2\pi/b) + d(c-1)(c-2)(2\pi/b)\right] \end{aligned} \right\} \\
& +\left\{ \begin{aligned} & 2A^3(c-2)(c-3)\sin(3\pi/b)\sin(2\pi/b) \\ & \qquad \left[(\pi/b)^3(c-1)^3 d^3 + c^3(1-d)^3\sin^3(\pi/b)\right] \end{aligned} \right\} \\
& +\left\{ \begin{aligned} & 6A^3cd(c-1)(c-2)(c-3)(1-d)(\pi/b)\sin(\pi/b)\sin(3\pi/b)\sin(2\pi/b) \\ & \qquad \left[d(c-1)(\pi/b) + c(1-d)\sin(\pi/b)\right] \end{aligned} \right\}
\end{aligned}
$$

For the 4th moment about mean, by substitution in the relationship between the fourth moment about the mean and the raw moments we have:

$$\mu_4 = \frac{A^4(4\pi/b)(c-4)(c-3)(c-2)(c-1)d + A^4(1-d)(c)(c+1)(c+2)(c+3)\sin(4\pi/b)}{\sin(4\pi/b)(c-4)(c-3)(c-2)(c-1)}$$

$$-4\left(\frac{A^3(3\pi/b)(c-3)(c-2)(c-1)d + A^3(1-d)(c)(c+1)(c+2)\sin(3\pi/b)}{\sin(3\pi/b)(c-3)(c-2)(c-1)}\right)$$

$$\left(\frac{A(c-1)d(\pi/b)+A(c)(1-d)\sin(\pi/b)}{(c-1)\sin(\pi/b)}\right)$$

$$+6\left(\frac{A^2(2\pi/b)(c-1)(c-2)d + A^2(c)(c+1)(1-d)\sin(2\pi/b)}{\sin(2\pi/b)(c-1)(c-2)}\right)$$

$$\left(\frac{A(c-1)d(\pi/b)+A(c)(1-d)\sin(\pi/b)}{(c-1)\sin(\pi/b)}\right)^2$$

$$-\left(\frac{A(c-1)d(\pi/b)+A(c)(1-d)\sin(\pi/b)}{(c-1)\sin(\pi/b)}\right)^4$$

Solving the above expression, the 4th moment about mean is given by

$$\mu_4 = \frac{1}{\sin^4(\pi/b)\sin(4\pi/b)\sin(3\pi/b)\sin(2\pi/b)(c-4)(c-3)(c-2)(c-1)^4}$$

$$\times\Big[A^4\sin(2\pi/b)\sin(3\pi/b)\Big\{(c-1)^3\sin^3(\pi/b)\sin(\pi/b)$$

$$\big[(4\pi/b)(c-4)(c-3)(c-2)(c-1)d+(1-d)(c)(c+1)(c+2)(c+3)\sin(4\pi/b)\big]$$

$$-(c-2)(c-3)(c-4)\sin(4\pi/b)\big[(c-1)^4 d^4(\pi/b)^4 + c^4(1-d)^4\sin^4(\pi/b)\big]\Big\}$$

$$-4A^4(c-1)(c-4)\sin(\pi/b)\sin(2\pi/b)\sin(4\pi/b)$$

$$\Big\{(3\pi/b)(c-3)(c-2)(c-1)^2 d\sin^2(\pi/b)\big[(\pi/b)(c-1)d$$

$$+c(1-d)\sin(\pi/b)\big]+c(1-d)\sin(3\pi/b)$$

$$\big[(c+1)(c-1)(c+2)\sin^2(\pi/b)\big[d(c-1)(\pi/b)+c(1-d)\sin(\pi/b)\big]$$

$$+(c-2)(c-3)d(\pi/b)\big[(c-1)^2 d^2(\pi/b)^2+c^2(1-d)^2\sin^2(\pi/b)\big]\big]\Big\}$$

$$+6\mathrm{A}^4(c-1)^2(c-3)(c-4)\sin(3\pi/b)\sin(4\pi/b)$$
$$\Big\{(2\pi/b)(c-1)(c-2)d\sin^3(\pi/b)\Big[(c-1)^2d^2(\pi/b)^2+c^2(1-d)^2\sin^2(\pi/b)\Big]$$
$$+c(1-d)\sin^2(\pi/b)\sin(2\pi/b)\Big[(c+1)\sin(\pi/b)$$
$$\Big[(c-1)^2d^2(\pi/b)^2+c^2(1-d)^2\sin^2(\pi/b)\Big]-(c-2)d^2(\pi/b)^2c(1-d)\Big]\Big\}$$
$$+12\mathrm{A}^4c(c-1)^3(c-3)(c-4)d(1-d)(\pi/b)\sin(\pi/b)\sin^3(\pi/b)$$
$$\sin(3\pi/b)\sin(4\pi/b)\Big[(2\pi/b)(c-1)(c-2)d+c(c+1)(1-d)\sin(2\pi/b)\Big]\Big] \quad (9.5)$$

The moment ratio β_1 and β_2 can easily be obtained.

10. QUANTILE-BASED MEASURES OF CENTER, SPREAD, SKEWNESS AND KURTOSIS

In this section, we derive some quantile-based measures of central tendency, dispersion, skewness and kurtosis.

The Mid Quartile Range is given by

$$\mathrm{MQR}=\frac{Q_1+Q_3}{2}=\frac{5\mathrm{A}}{3} \quad (10.1)$$

The Inter-Quartile Range is given by

$$Q_3-Q_1=3\mathrm{A}-\frac{\mathrm{A}}{3}=\frac{8\mathrm{A}}{3} \quad (10.2)$$

The Bowley's Co-efficient of Skewness is given by

$$\frac{Q_1+Q_3-2Q_2}{Q_3-Q_1}=\frac{\frac{\mathrm{A}}{3}+3\mathrm{A}-2\mathrm{A}}{3\mathrm{A}-\frac{\mathrm{A}}{3}}=\frac{1}{2} \quad (10.3)$$

The Percentile Co-efficient of kurtosis is

$$\frac{Q_3-Q_1}{2(D_9-D_1)}=\frac{3\mathrm{A}-\frac{\mathrm{A}}{3}}{2\left(9\mathrm{A}-\frac{\mathrm{A}}{9}\right)}=\frac{3}{20} \quad (10.4)$$

It is interesting to note that the Bowley's Co-efficient of Skewness and the Percentile Co-efficient of kurtosis of density function (3.3) is independent of A.

11. CONCLUDING REMARKS

SIA log-symmetric distributions are of interest due to their capability of providing SIA-estimators of distribution parameters which are more efficient than the ordinary moment estimators. A simple way of obtaining new SIA log-symmetric density functions is to take a linear combination of two already existing SIA densities. In this paper, we derive a new SIA log-symmetric density function based on linear combination of the SIA Log-Logistic Distribution and the SIA Inverse-Gamma Ratio distribution. Some of the fundamental properties of the newly derived distribution such as shape of the density, hazard function, quantile function moments and moment-ratios have been obtained. Quantile-based measures such as mid-quartile range, inter-quartile range, Bowley's coefficient of skewness and percentile coefficient of kurtosis have also been derived. Real-life application of the newly derived distribution is to be explored in the near future.

REFERENCES

1. Habibullah, S.N. (2012). A Generalization of the Standard Half–Cauchy Distribution. *Proceedings of the Twelfth Islamic Countries Conference on Statistical Sciences (ICCS-12)*, Doha, Qatar, Vol. 23, 33-42.
2. Habibullah, S.N. and Ahmad, M. (2006). On a New Class of Univariate Continuous Distributions that are Closed Under Inversion. *Pakistan Journal of Statistics and Operations Research*, 2, 151-158.
3. Habibullah, S.N. and Fatima, S.S. (2015). On a Newly Developed Estimator for More Accurate Modeling with an Application to Civil Engineering, *12th International Conference on Applications of Statistics and Probability in Civil Engineering* (Vancouver, Canada)
4. Habibullah, S.N. and Saunders, S.C. (2011). A Role for Self-Inversion", *Proceedings of International Conference on Advanced Modeling and Simulation (ICAMS, Nov 28-30, 2011)* published by Department of Mechanical Engineering, College of Electrical and Mechanical Engineering, National University of Science and Technology (NUST), Islamabad, Pakistan, Copyright 2011, ISBN 978-869-8535-11-7.
5. Habibullah, S.N., Memon, A.Z. and Ahmad, M. (2010). *On a Class of Distributions Closed Under Inversion*, Lambert Academic Publishing (LAP), Heinrich-Böcking-Straße 6, 66121 Saarbrücken, Germany, ISBN: 978-3-8383-4868-1.
6. Jones, M.C. (2008). On Reciprocal Symmetry. *Journal of Statistical Planning and Inference*, 138(10), 3039-3043.
7. Marques, F.J., Coelho, C.A. and de Carvalho, M. (2015). On the distribution of linear combinations of independent Gumbel random variables. *Stat Comput.*, 25, 683-701.
8. Nadarajah, S. (2007). The linear combination, product and ratio of Laplace random variables. *Statistics*, 41:6, 535-545.
9. Nadarajah, S. & Kotz, S. (2007). On the Linear Combination of Laplace and Logistic Random Variables. *Journal of Applied Statistics*, 34:2, 185-194,
10. Seshadri, V. (1965). On Random Variables which have the same Distribution as their Reciprocals. *Canadian Mathematical Bulletin*, 8(6), 819-824.
11. Vanegas, L.H. and Paula, G.A. (2016). Log-symmetric distributions: Statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics,* 30(2), 196-220.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 409-414

APPLICATION OF SIMPLE REGRESSION ON WALL THICKNESS OF PIPE USING NOMINAL PRESSURE PN-16 AND PN-20

Muhammad Jawad Khan, Usman Zain-ul-Abdin, Danish Hassan, Hamza Khan, S. Saqib Shams, Muhammad Ayyoob, Arsalan Ahmed and **Muhammad Ali**
National Textile University, Karachi Campus, Pakistan
Email: mjawadkhan130@gmail.com usmanabdin5@gmail.com
danishhassan@ntu.edu.pk hamzakhan@ntu.edu.pk
saqibshams@ntu.edu.pk dr.mayyoob@ntu.edu.pk
arsalan_ahmed@ntu.edu.pk ali.ikhtiar406@gmail.com

ABSTRACT

The research work done in this paper demonstrates the application of mathematical models based on wall thickness (minimum & maximum) and pressure rating data under the nominal pressure (PN-16 & PN-20). For this purpose, different sizes (mm) of HDPE water supply pipe have been observed. Our simple regression models show the adequacy as coefficient of determination (R^2) is found greater than 95% for each model. With the help of obtained mathematical model we can observe the minimum and maximum expected values of PN-16 and PN-20 for different sizes of pipes. This kind of analysis can be beneficial for the polymer industry to understand and predict the correlation between wall thickness and nominal pressure.

KEYWORDS

Polymer, simple regression, wall thickness, nominal pressure.

1. INTRODUCTION

There are different standards for measuring the relationship between wall thickness and the pressure, these are:

- ASTM
- ISO
- DIN
- BS/PS [1].

The PN (16 & 20) are the testing standard in DIN Standard.

The purpose of applying the Simple linear regression on the pipe thickness is to overcome the problem like variable pipe thickness for PN (16 & 20). E.g. if a pipe has size 20 or 25*mm* with it minimum and maximum PN (16 &20) given then by applying Simple Linear Regression, we can find the different size and its pressure for different values like: 20.5*mm*, 21*mm* etc. [2].

NOMINAL PRESSURE (PN):

- A numerical designation relating to pressure that is a convenient round number for reference purposes.
- It is intended that all equipment of the same nominal size (DN) designated by the same PN number shall have the same mating dimensions appropriate to the type of end connections [3].

Detail:

As in project, simple linear regression is applied on wall thickness model of pipes of HDPE and PPRC.

- Pressure Nominal -PN:
- PE pipe are produced in different pressure grades (PN Grades), which indicates the pressure in bars the pipe can support with water at 20^{O}C.
- PN 2.5 – max pressure 2.5 bar
- PN 4 - max pressure 4 bar
- PN 6 - max pressure 6bar
- PN 10 - max pressure 10 bar
- PN 16 - max pressure 16bar

$1 bar = 10^5\ Pa\ (N/m^2) = 0.1\ N/mm^2 = 10.197\ kp/m^2 = 0.9869\ atm$ [4].

2. METHODOLOGY

Simple linear Regression

- **Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables [5].
- One variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted y, is regarded as the **response**, **outcome**, or **dependent** variable.

$$Y = a + bX \tag{1}$$

where Y is the dependent variable, X is the independent variable and $\boldsymbol{a}$, $\boldsymbol{b}$ are slope [6].

Table 1
Standard values of PN-16 & PN-20 on Different Sizes

Size (mm)	Outer Diameter (mm)	Wall Thickness			
		PN-16		PN-20	
		MIN	MAX	MIN	MAX
20	20.0~20.4	2.7	3.2	3.3	4
25	20.0~25.3	3.4	4.1	4.3	4.8
32	32.0~32.4	4.4	5.2	5.3	6.1
40	40.0~40.4	5.5	6.3	6.6	7.6
50	50.0~50.5	6.8	7.7	8.2	9.2
63	60.0~63.6	8.5	9.6	10.5	11.7
75	70.0~75.6	10.2	11.5	12.5	14.1
90	90.0~90.7	12.2	13.8	15.1	16.5
110	110.0~111.1	15.0	16.8	18.2	20.3

Table 2
Expected Values of PN-16 & PN-20 on Different Sizes

Size (mm)	Outer Diameter (mm)	Wall Thickness			
		PN-16		PN-20	
		MIN	MAX	MIN	MAX
23	23.0~23.4	3.1	3.7	3.8	4.5
36	36.0~36.4	4.9	5.7	6.0	6.8
45	45.0~45.3	6.1	7.0	7.5	8.5
54	54.0~54.4	7.3	8.4	9.0	10.1
66	66.0~66.5	8.9	10.2	10.9	12.3
78	78.0~78.6	10.6	12.0	12.9	14.5
85	85.0~85.4	11.6	13.0	14.1	15.7
99	99.0~99.3	13.5	15.1	16.4	18.3
112	112.0~113.1	15.2	17.1	18.6	20.6

The Table 1 represents the standard values of PN-16 and PN-20 whereas Table 2 shows the simulated values using fitted adequate Mathematical models.

3. RESULTS AND DISCUSSIONS

The data sets were taken from the testing lab department of Polymer engineering National Textile University Karachi Campus. We have applied the simple linear regression models on the PN-16 and PN-20 minimum and maximum data sets of different sizes of pipe. The purpose of utilizing the Mathematical models on the selected data was to obtained the significant models that can give the expected values. Using the obtained models we can simulate the different PN values of any size of the pipe.

This technique can be useful for the polymer industry to understand the different PN values before testing in the labouraties. The obtained Mathematical models are as under.

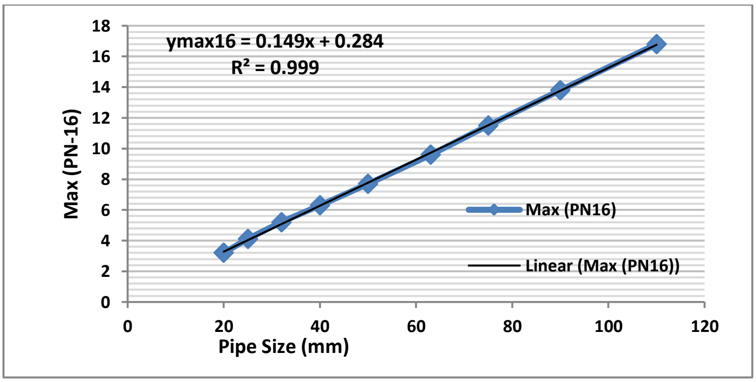
$$\boldsymbol{y_{min16} = 0.1359x + 0.0101} \quad (2)$$

$$\mathbf{y_{max16} = 0.1499x + 0.2849} \quad (3)$$

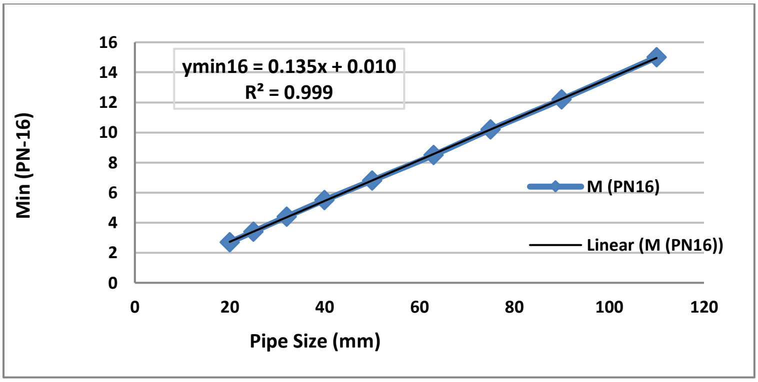
$$y_{min20} = 0.166x + 0.01 \quad (4)$$

$$y_{max20} = 0.181x + 0.294 \quad (5)$$

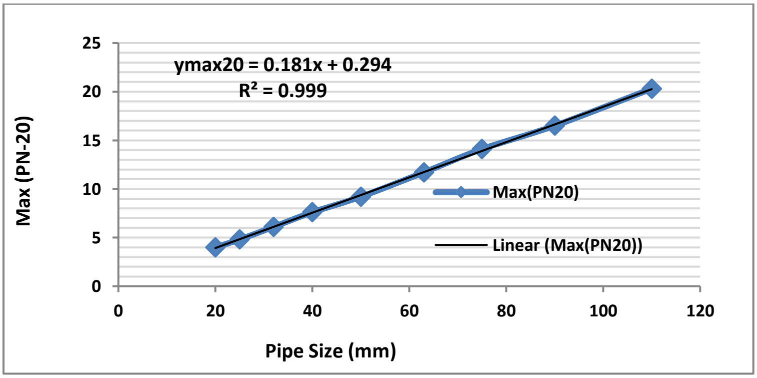
The Figures 1((a)-(d)) depicts the adequacy of the model, since coefficient of determination (R^2) shows the highest value.



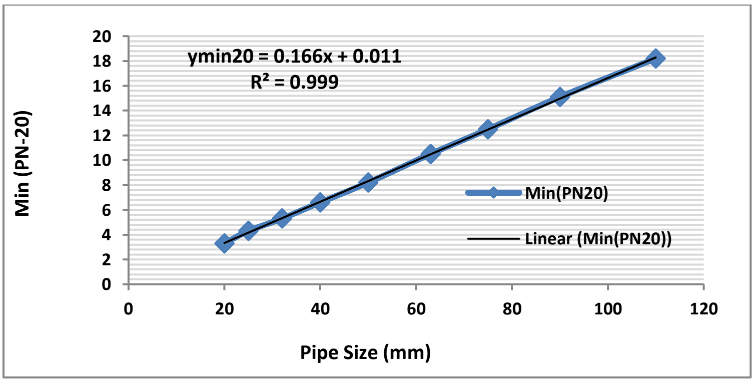
a



b



c



d

Figure 1: The Figs. (a) and (b) depicts the model fitting on PN-16 maximum and minimum respectively. Similarly Figs. (c) and (d) indicates the significant fitting of model on maximum and minimum of wall thickness of PN-20.

4. COMMENTS AND CONCLUSION

The research work done in this paper demonstrated the application of Mathematical models on wall thickness (minimum & maximum) data sets under the nominal pressure (PN-16 & PN-20). For this purpose, different sizes (*mm*) of HDPE water supply pipe have been observed. Our simple regression models show the adequacy as coefficient of determination (R^2) is found greater than 95 % for each model. The study can be beneficial for the polymer industry to observe and understand the expected values before testing for any PN test of any size.

ACKNOWLEDGEMENT

The authors are thankful to the staff (Mr. Tariq Jamal & Mr. M. Umair Israr) of polymer testing lab National Textile University Karachi Campus for providing the data.

REFERENCES

1. Khan, S. (2016). *A Study on the Authenticity of Indigenious Art used in Brass anitque and Artifacts-A Case Study of BRASS City "Moradabad"* (Doctoral dissertation, Pondicherry University).
2. Meliani, M.H., Bouledroua, O., Azari, Z., Sorour, A., Merah, N. and Pluvinage, G. (2017). The inspections, standards and repairing methods for pipeline with composite: a review and case study. In *International Conference on New Trends in Fatigue and Fracture* (pp. 147-156). Springer, Cham.
3. Johnstone, J.E. (2009). Facility Safety: Avoiding the 10 Most Common Design Errors. In *SPE Americas E&P Environmental and Safety Conference*. Society of Petroleum Engineers.
4. Erck, R.A. and Ajayi, O.O. (2001). *Analysis of sliding wear rate variation with nominal contact pressure* (No. ANL/ET/CP-104067). Argonne National Lab., IL (US).
5. Khan, A., Hassan, D. and Raza, S.M. (2018). Sunspots Activity Cycles Using Linear and Non-Linear Technique. *Journal of Information Communication Technologies and Robotic Applications*, 7(1), 45-54.
6. Mooi, E., Sarstedt, M. and Mooi-Reci, I. (2018). Regression analysis. In *Market Research* (pp. 215-263). Springer, Singapore.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 415-425

FINANCIAL RISK ANALYSIS

Yüksel Akay Ünvan
T.C. Ankara Yıldırım Beyazıt University, Turkey
Email: akay.unvan@gmail.com

ABSTRACT

Financial risk is the possibility of losing money on an investment or business venture. Financial risks include credit risk, liquidity risk, and operational risk etc. There are various types of financial risk analysis techniques for evaluating risk. For example, Value at Risk (VaR) is a metric that estimates the financial risk. This statistical technique used to measure the amount of risk (loss) that could happen in an investment in a defined period. VaR gives the probability of losing more than a given amount. So, risk managers use VaR to control the level of risk exposure. The aim of this presentation is to highlight and illustrate the use of some techniques for risk analysis in finance.

1. INTRODUCTION

Financial risks are uncertainties associated with any form of financing, including market risk, liquidity risk, credit risk, business risk, investment risk, and operational risk. Financial data analysis (business intelligence (BI)) helps companies like an early warning system to minimize default risks. BI is also useful for better decision (Kou et al., 2014).

Financial risk management is monitoring financial risks and managing their effect. In other words, financial risk is the risk that a company may default on its debt payments. The problem of the debt, default or prosperity of the company are one of the ways of the risk management. A prediction of corporate default is also an element of the risk management. Financial risk management is basically a process to deal with the unexpected losses (Valášková et al., 2018). In order to manage financial risk, it is a necessity for the firms to make organizational decisions about the risks they are exposed. Hedging is a financial term used to describe a risk management strategy aiming to limit or to counterbalance the probability of loss due to fluctuations in the prices of commodities, securities or currencies (Dalgıç, 2013).

Risk management in the finance can be defined as the process of identification, analysis, and acceptance or mitigation of uncertainty in investment decisions. Actually, risk management occurs when an investor or fund manager analyzes and attempts to measure the potential for losses in an investment, such as a moral hazard, and then takes the appropriate action (or inaction) given the fund's investment objectives and risk tolerance (Kenton, 2021).

Financial organizations are open to risks and they can not continue working without taking risks. Risk causes a great deal of potential damage and inconvenience for the enterprise stakeholders. So it is vital for organizations to establish proactive strategies to manage or prevent risks. Therefore it becomes important for financial organizations to model risks using historical data in order to gain insight into the risk patterns, so that it falls under their acceptable thresholds (Srinivasan and Kamalakannan, 2018).

2. TYPES OF FINANCIAL RISK

In this section the most common financial risk types are explaıned.

2.1 Market Risk

Market risk is the risk of loss caused by movements in the level or volatility of market prices. Market risk can be classified into absolute risk (measured in the relevant currency) and relative risk (measured against a benchmark index). While absolute risk is based on the volatility of total returns; relative risk measures risk in terms of tracking error, or deviation from the index. Market risk is diversified into directional and non-directional risks. Directional risks include risks regarding the direction of movements in financial variables (stock prices, interest rates, exchange rates and commodity prices). Non-directional risks include other risks than these. Market risk is controlled by the limits on notionals, exposures, VAR measures and independent audit of risk managers (Jorion, 2007).

Market risk refers to the risk that a financial instrument's cash flows or fair value may fluctuate owing to market price changes. Market risk includes currency risk, interest rate risk and other price risks (Kozol, 2020). There are four major sources of market risk called equity price risk, interest rate risk, foreign exchange risk, and commodity price risk (Dalgıç, 2013).

2.2 Credit Risk

Credit risk can be defined as the debtor's failure to fulfill its obligations under the terms of the agreement. With financial liberalization, increased competition among financial institutions, expansion of credit markets and the introduction of new financial instruments, managing credit risk has become complicated for banks. The first step in the effective credit risk management process is to define the credit risks that the bank is exposed to, measure these risks and use new techniques to protect against credit risks **(**Oktay and Temel, 2007).

Credit risk is the risk of losses owing to the fact that counterparties may be unable to fulfill their contractual obligations. Losses owing to credit risk, however, can occur before the actual default. It should be defined as the potential loss in mark-to market value that may be incurred owing to the occurrence of a credit event. A credit event occurs when there is a change in the counterparty's ability to perform its obligations. So changes in market prices of debt owing to changes in credit ratings or in the market's perception of default also can be viewed as credit risk, creating some overlap between credit risk and market risk. Credit risk also includes sovereign risk. This occurs, for instance, when countries impose foreign-exchange controls that make it impossible for counterparties to honor their obligations. Credit risk is controlled by credit limits on notionals, current and potential exposures, and increasingly, credit enhancement features such as requiring collateral or marking to market. The new methods to quantify market risk are now being extended to credit risk. Credit risk analysis is a multi-criteria decision problem since the decision taken depends on various factors that define a customer (Jorion, 2007).

2.3 Liquidity Risk

Liquidity risk is the risk that occurs when an organization cannot meet its contractual obligations on the due date at a reasonable cost. Liquidity risk arises either when an institution fails to respond to the changes in the market conditions which will result as the decreased ability of the institution to liquidate its assets quickly and with minimal loss, or when an institution fails to manage unanticipated changes in funding sources. Those explications show that the liquidity risk is related to the mix of assets and liabilities and it can arise in any situation where assets and liabilities are not matched (Dalgıç, 2013).

In other words, when liquidity is defined as cash and cash equivalents owned by a business, then the liquidity risk can be stated as the lack of cash and cash equivalents in the business or the inability to sell assets at a reasonable price specifically for the purpose of generating cash; or as the risk of not finding funding at reasonable costs from the market (Gülhan, 2018).

2.4 Operational Risk

In a general definition operational risk is the risk of damage caused by disruptions experienced in the operating process of the institution. According to the internationally accepted definition adopted by the Basel Committee, operational risk is the risk of damage caused by inadequate and unsuccessful internal processes, people and systems, or external events. Operational risk is the oldest and most fundamental risk faced by financial institutions. According to one view, operational risks are also on the basis of other risks (market, credit risk, etc.). It is stated that the cause and trigger of these risks are operational risks. For example, although the non-payment of the loan on maturity is defined as credit risk, the loss incurred if a staff approves a loan that he/she should not approve according to the procedures. It is considered as an operational risk. This situation makes operational risk management more important for the management of other risks that organizations are exposed to (Can, 2003).

3. LITERATURE REVIEW

Kou et al. (2014), introduce an MCDM-based approach which rank a selection of popular clustering algorithms in the financial risk analysis domain. This is an experimental study which is formed for verifying the proposed approach by handling 3 MCDM methods, 6 clustering algorithms, and 11 cluster validity indices over 3 real-life credit risk and bankruptcy risk data sets. The results obtained from the study show how much effective the MCDM methods about interpreting clustering algorithms and reveal that the repeated-bisection method provides good 2-way clustering solutions on the selected financial risk data sets.

Beasley et al. (2015) have conducted a study on different perspectives of risk managers and executives. They found out that risk assessment is a critical process for an organization. The study used data collected from executives over a 3 year period to evaluate different kinds of risks and revealed 20 different kinds of risk under 3 main categories which can be stated as macro-economic risks, strategic risks and operational risks. The risk of customer acquisition and retention by the organization, which is so meaningful for an organization about determining customer behavior, ranked in the fourth place. The study also finds out that customer characteristics have a great effect on an organization's credit risk.

Darwish and Abdelghany (2016) use a fuzzy logic model to identify credit risk in the Egyptian commercial banks. The authors have explored some characteristics of banks such as profitability, debt paying ability, operation ability and liquidity in order to make predictions about the credit risk.

Valášková et al. (2018) evaluated the financial risks of Slovak entities in order to reveal the significant factors and determinants affecting the prosperity of Slovak companies. They used the data of Slovak enterprises, obtained from annual financial reports covering the year 2015 and then calculated financial ratios of profitability, activity, liquidity and indebtedness that may affect the financial health of the company were applied in the regression analysis. The statistically significant determinants which affects the future financial development of the company are determined by using the multiple regression analysis and the regression model of the bankruptcy prediction. The study especially handled the significant economic risk factors by implementing multiple regression. The results obtained from the application point out that the most significant predictors are net return on capital, cash ratio, quick ratio, current ratio, net working capital, RE/TA ratio, current debt ratio, financial debt ratio and current assets turnover changing the decision about the future company default. That is, all these factors have importance while managing financial risks and identifying the profitability and prosperity of the company.

Nugroho et al. (2018) explored the financial risk management in Islamic social organizations. The research implemented the qualitative descriptive by using manager self-assessment questioner in Islamic social enterprises. The research considers 24 respondents from the total 100 respondents in Jakarta and Java Island. The results show that the financial risk management has an average score above 80 which means it has implemented well in Islamic social enterprises.

Dang et al. (2020) studied the factors which affects an enterprise's financial risk in the Vietnam stock market. The panel data analysis was implemented in the study considering a sample of 524 non-financial listed enterprises on the Vietnam stock market between the years of 2009 to 2019 and also the Generalized Least Square (GLS) method is used. Moreover, Alexander Bathory model was applicated in order to measure financial risk. The independent variables of the study were debt structure, solvency, profitability, operational ability, capital structure while the firm size, firm age, growth rate were taken as control variables. The results obtained by the study reveals that some of the variables such as liability structure ratio, quick ratio, return on assets, total asset turnover, accounts receivable turnover, net assets ratio and fixed assets ratio have importance for preventing financial risk.

Santos et al. (2017) explored the relationship between the financial risk management and the value creation to the shareholder. The study considered the non-financial Brazilian firms. The data used in the study was consisted of 1794 firm-year observations belonging the years between 2006 and 2014. The panel data and the Generalized Methods of Moments (GMM) analyses were implemented on the data and the results revealed that the firms which used derivatives did not get any value during the years considered.

Belas et al. (2018) explored some factors which are significant on the intensity of financial risk. In addition, they aimed to compare the financial risk perception of SME entrepreneur groups according to their motivation to start a business. In the study, they used Z-score and regression methods by conducting a survey on the participant group of 1,141 business owners. Evaluating the attitude towards financial risk, the authors used CRM (Accurate risk management by Entrepreneurs), and CFR (Credit risk impact during a crisis) and SFS (Adequate Funding for SMEs) factors. The data obtained showed that the entrepreneurs who saw their business as their mission answered the option of managing financial risk in their companies more frequently than entrepreneurs who established their businesses for money. The study also revealed that entrepreneurs who started their businesses because of money felt the effects of the crisis on the financial risks of their companies.

Zhang and Jiang (2019) use the Copula function to improve the ability of the Chinese financial industry to counter risks. The authors define Copula function as a class of functions that based on marginal probability density function to create joint distribution and is convenient to implement in finance risk analysis. The results of the study reveal that there is an upper tail correlation between the two indexes as well as a lower one, and the correlation between the upper tail and the lower tail is high. The findings confirm that the results obtained are meaningful and applicable to the financial world.

Olah et al. (2019) explored the economic and financial risk of the sources in V4 SMEs (Visegrad Group which consists of Czech Republic, Hungary, Poland and Slovakia) and Serbia. A survey was conducted in which 2110 SMEs from the countries specified within the scope of the study participated. The survey included questions about the importance of risks in organizations and risk management perspective. The general non-hierarchical log-linear model with a Z-value was implemented in the application part. As a result of the study, differences between V4 countries and Serbia were revealed. For example, Serbia is more open to the sources of financial risk compared to V4 countries. In addition, the analyzed data also showed that insufficient profit is most dangerous risk source which can influence all of the countries.

Yang et al. (2019) developed a supply chain financial risk management model to increase resilience and flexibility against risk. This model is defined under the internet finance model. The developed model is created based on multivariate piecewise regression analysis. Data science analysis method was also used in the study. Moreover, in order to evaluate the risks of Internet supply chain the fuzzy decision method is implemented, while the Simunic model is preferred to establish the correlation model of risk management in Internet supply chain. The piecewise sample regression analysis has been used to increase the risk management capability of the Internet supply chain. The study results revealed that the model used is reliable for financial risk management of the supply chain.

Acerbi, Nordiot, and Carlo Sirtori (2018) examine the characteristics of the Expected Fall in terms of financial risk management. In some cases where VaR cannot distinguish portfolios with different risk levels, this measure of choice has been come up to be more advantageous than VaR. They also state the main difference of this variable from VAR is that it is a coherent measure of risk while VaR is not as they show with a counterexample. They consider ES to be an important alternative to VaR for financial Risk Management purposes, where the warnings applicable to VaR will also be removed (precision of estimation, reliability of estimates, transparency in hypotheses). They also argue that ES is a reliable tool for assessing relative riskiness without restrictions on applicability.

Terpezan (2009) states Value at Risk (VaR) as the measurement method of market risk, and defines it as one of risk measurement methods in banking sector. After defining basic properties of VAR, the most used VAR calculating methods such as Parametric VaR, Historical Simulation and Monte Carlo Simulation are explained. At the end of the study, she concluded that VAR is a reliable risk measurement model which is often preferred in the banking sector and each calculation method also proves the importance of the method.

Uylangco (2015) also studied Value-at-Risk (VaR) method for risk measurement in Australian banks. The study considered the data of Global Financial Crisis (GFC) period and explored the effects of methodology and parameter selection about supporting a bank crisis. The author pointed out that the VaR methodology supported by Basel II was frequently criticized for failing to capture downside risk during the GFC. He stated that the results of the study indicated that 1-year parametric and historical models produced better VaR measurements than models with longer time applications. In the study, VaR models created using Monte Carlo simulations and Arma Garch model were also compared. It was emphasized that the findings obtained as a result of the study are compatible with the revised Basel II VaR methodology.

4. VALUE AT RISK (VAR)

VaR is a method of assessing risk that uses statistical techniques. VaR summarizes the worst loss that will not be exceeded with a given level of confidence. VaR provides a summary measure of market risk. For instance, a bank might say that the daily VaR of its trading portfolio is $100 million at the 95 percent confidence level. So, there is only 5 chance in a 100, under normal market conditions, for a loss greater than $100 million to occur. This number summarizes the bank’s exposure to market risk, as well as the probability of an adverse move. Shareholders and managers then can decide whether they feel comfortable with this level of risk. If the answer is no, the process that led to the computation of VaR can be used to decide where to trim the risk (Jorion, 2007).

Value at Risk as a method to quantify risk is utilized extensively by financial institutions. It is a simple method to calculate risk, since it only requires two parameters, namely confidence level and holding time period. By employing these two parameters, it yields the maximum expected percentage loss possible over the given horizon at the given confidence level (Şencal, 2007).

After defining the VAR method, the VAR types will be explained below.

4.1 Incremental VaR

Value at Risk (VaR), which we have witnessed its applications with an increasing trend in recent years, is frequently preferred as an important tool for the design of portfolio strategies. Incremental VaR (IVaR), the increased impact on VaR of adding a new instrument to an existing portfolio, is a tool for making hedging decisions. Approximate formulas suitable for small portfolio composition changes have been proposed in the literature, as calculating full IVaR can be troublesome and computationally costly. According to the most commonly used formula, IVaR is approximately equal to the current

VaR multiplied by the beta coefficient (with respect to the current portfolio) of the candidate asset.

VaR refers to the maximum amount we are likely to lose at some specific confidence level q%. Let X be the random current portfolio return with mean μz and finite varaince $\sigma x2$; Thus, the relative VaR is defined as

$$Pr(X - \mu z <= -VaR) = q\% \tag{1}$$

Clearly, the smaller its VaR, the more preferable the portfolio. In order to design profitable portfolio strategies, a logical step is to calculate the VaR of the portfolio either including the candidate asset or not. The Incremental VaR (IVaR) associated with a position in asset Y is, for a fixed value q%:

IVaR = VaR (with the asset Y) – VaR {without the asset Y) I

If IVaR is positive, the candidate asset entry upgrades the portfolio risk, vice versa if it is negative (Tibiletti, 2001).

The incremental VaR comes from two sources. The first source is the increment of the asset value of the portfolio. Since the P&L is the product of rate of return and the asset value at the beginning of the period under consideration, the increase of the asset value amplifies the P&L of a portfolio. When an additional asset is added into an existing portfolio, it tends to reduce the incremental VaR due to the increase of total asset value in the portfolio. The second source is the diversification of the portfolio. When a new asset is added to an existing portfolio to form a new portfolio and the asset value of the new portfolio is scaled to the same level as the one before the new asset is added, diversification reduces the size of the P&L of the new portfolio. The second source tends to increase the incremental VaR. The decomposition of incremental VaR makes it interesting to study which source is stronger for the current portfolio (Wang, 2002).

4.2 Marjinal VAR

In order to capture the effect of change in dollar exposure of assets on the overall risk of the portfolio, the VaR of the individual assets is not adequate. In this context it is more pertinent to account for the contribution of each asset to the overall risk of the portfolio. Accordingly, Marginal Value-at-Risk (MVaR) is more amenable when it comes to identifying the additional portfolio risk, consequent to every unit currency of additional exposure to that asset (Investopedia 2019).

MVaR for an asset in the portfolio, measures the sensitivity of the portfolio VaR to an increase in the dollar exposure of the asset. Accordingly, if VaRP is the portfolio VaR, and V_i is the exposure of the ith asset in dollar terms, then MVaR for the asset i is defined as

$$MVaR_i = {\partial(VaR_p)}\Big/{\partial V_i} \tag{2}$$

(Jain & Chakrabarty, 2020)

4.3 Conditional VaR

Conditional Value at Risk (CVaR) attempts to address the shortcomings of the VaR model, which is a statistical technique used to measure the level of financial risk within a firm or an investment portfolio over a specific time frame. While VaR represents a worst-case loss associated with a probability and a time horizon, CVaR is the expected loss if that worst-case threshold is ever crossed. CVaR, in other words, quantifies the expected losses that occur beyond the VaR breakpoint (Chen, 2020).

Usually, one collects additional information $\mathcal{J}$ (i.e., past observed returns, economical exogenous covariates, etc.), and to take into account such relevant information $\mathcal{J}$, the conditional VaR (CVaR) of the variable Y_t (being Y_t the risk or loss variable which can be the negative logarithm of returns at time t) is defined, for a fixed level p, as the value $vp(t)$ such that $P(Yt > vp(t)|\mathcal{J}) = p$. The conditional ES (CES) is defined as $\mu p(t) = E[Y_t|Yt > vp(t), \mathcal{J}]$. Most studies estimate CVaR through quantile estimation. (Ferraty and Quintela-Del-Río, 2016).

4.4 VaR Calculation Methods

VaR calculation methods which are mostly used in literature and financial environment are Historical Simulation (HS), Monte Carlo Simulation (MCS) and Variance Covariance (V-C) methods. These three methods will be described briefly below.

Firstly, the Historical Simulation method which is also known as “nonparametric VaR” indicates the distribution of profit and loss ranges at a given confidence level by examining the impact of backdated data on the current portfolio. This method is considered to be more understandable and easy to apply than others, and can even be described as a simplified version of Monte Carlo Simulation. The idea behind the VaR calculation using the HS method is to simulate the VaR using the historical return distribution of the securities in the portfolio, assuming we have preserved the current portfolio since the beginning of historical data. In order for this to be implemented, the past returns of the assets in the portfolio must first be obtained for a certain period of time. It is assumed that the historical returns of the assets that make up the portfolio are kept from the beginning of the historical returns in order to calculate the hypothetical returns of the current portfolio. Assuming that the distribution of portfolio returns reflects the future, the respective confidence level of this distribution gives the expected VaR of the portfolio (Dalbudak et al., 2017).

Variance-Covariance method is a parametric method and calculates the volatility and correlations of portfolio returns. For this purpose, the method uses the V-C matrix of portfolio returns based on past time series and so it takes its name from the matrix used. The reason why the method is characterized as parametric is that the result to be obtained with the model depends on the correct estimation of the distribution parameter and the formation of the appropriate return distribution of the assets (Gökgöz, 2006).

MCS Method is a frequently used method for VaR calculations and is also preferred in quantitative finance. MCS method is a very powerful and comprehensive method when used correctly in terms of measuring market risk. The method is a form of VaR calculation in which the computer is used extensively. It is usually used with a normal distribution, but it can also work with different distribution assumptions. In addition, the model risk that arises in other methods is almost completely eliminated in this method. In fact, it can be said that MCS Method is a mixture of V-C Method and HS Method. The variance covariance matrix of historical returns is required in both MCS and V-C methods. However, the MCS approach is not satisfied with this, and it creates a new correlated series based on the variance covariance matrix. The next process is the same as the HS method. If the time interval used in the method is the same as the time used to create the variance-covariance matrix and the portfolio exhibits a linear behaviour, the results for MCS and HS are more or less the same. However, if the portfolio exhibits non-linear behavior (for reasons such as options), the results will be different. (Dalbudak et al., 2017).

5. COMMENTS AND CONCLUSION

The financial world, which grows and diversifies day by day, is becoming more and more open to risks. Financial risk management appears with a greater importance for institutions with the effect of increasing external factors and competitive environment. This situation turns financial risk measurement methods into an attractive research and application area. The VAR method, which is used with an increasing interest both in the literature and in the financial world, is among these methods.

Survival in the global financial environment depends on accurate risk measurement calculations by establishing an effective financial risk management. At this point, institutions have to choose the most effective and functional method for themselves by correctly defining the individual characteristics specific to their institutions as well as global factors. Otherwise, it will be vulnerable to risk dynamics that are getting stronger with each passing day, causing their perpetuals to be threatened**.** So the aim of this paper is to emphasize and illustrate the use of some techniques for risk analysis in finance by noting the importance of financial risk management. Thus, it will be reminded once again how critical the science of statistics is also for financial risk management, which is in the field of application.

REFERENCES

1. Acerbi, C., Nordio, C. and Sirtori, C. (2001). Expected Shortfall as a Tool for Financial Risk Management. Preprint. Available at: https://arxiv.org/pdf/cond-mat/0102304.pdf
2. Beasley, M., Branson, B. and Pagach, D. (2015). An examination of the assessment of top risks on the horizon: Evidence from executives and risk professionals. *The Journal of Enterprise Risk Management*, 1, 1-22.
3. Belás, J., Dvorský, J., Kubálek, J. and Smrčka, L. (2018). Important factors of financial risk in the SME segment. *Journal of International Studies*. 11(1), 80-92.
4. Can, E. (2003). Operational Risk and Management, Proficiency Study of Capital Markets Board. Available at: https://www.spk.gov.tr/SiteApps/Yayin/YayinGoster/401.
5. Chen, J. (2020). Conditional value at risk (CVaR). *Investopedia. Dostupné na:< https://www.investopedia.com/terms/c/conditional_value_at_risk.asp*.
6. Dalbudak, Z.I., Murat, A.T.A.N. and Yilmaz, V. (2017). Comparison of Value at Risk Methods: Application of ISE 30. *Journal of Business Economics and Finance*, 6(3), 254-263.
7. Dalgıç, N. (2013). A Research on Derivatives Usage by Registered Corporations of Istanbul Stock Exchange, *Master of Thesis*, Yeditepe University.

8. Dang, H.T., Phan, D.T., Nguyen, H.T. and Hoang, L.H.T. (2020). Factors Affecting Financial Risk: Evidence from Listed Enterprises in Vietnam. *The Journal of Asian Finance, Economics and Business*, 7(9), 11-18.
9. Darwish, N.R. and Abdelghany, A.S. (2016). A Fuzzy logic model for credit risk rating of Egyptian commercial banks. *International Journal of Computer Science and Information Security*, 14, 11-19.
10. Ferraty, F. and Quintela-Del-Río, A. (2016). Conditional VaR and expected shortfall: a new functional approach. *Econometric Reviews*, 35(2), 263-292.
11. Gökgöz, E. (2006). *Value at Risk VaR and Portfolio Optimization*, Ankara: Capital Markets Board Publishing, No:190.
12. Gülhan, O. (2018). *Liquidity Risk and Liquidity in Banking Regulations Applications on the Turkish Banking Sector*, PhD Tesis, Başkent University.
13. Jain, S. and Chakrabarty, S.P. (2020). Does Marginal VaR Lead to Improved Performance of Managed Portfolios: A Study of S&P BSE 100 and S&P BSE 200. *Asia-Pacific Financial Markets*, 27(2), 291-323.
14. Jorion, P. (2007). *Value at Risk: The New Benchmark for Managing Financial Risk*, McGraw Hill, USA.
15. Kenton, W. (2021). *Risk Management in Finance*, available at https://www.investopedia.com/terms/r/riskmanagement.asp.
16. Kou, G., Peng, Y. and Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1-12.
17. Kozol, E. (2020). Financial Risk Management and Firm Performance: Evidence from European Cross-Border Mergers and Acquisitions, *Master of Thesis*, Marmara University.
18. Nugroho, L., Utami, W., Sanusi, Z.M. and Setiyawati, H. (2018). Corporate Culture and Financial Risk Management in Islamic Social Enterprises (Indonesia Evidence), *Int. Journal of Commerce and Finance*, 4(2), http://hdl.handle.net/11467/2757.
19. Oktay, S., Temel H. (2007), A Field Study of Credit Risk Management in Commercial
20. Banks in the Framework of Basel II Criteria, *ZKU Journal of Social Sciences*, 3(6), 163-185.
21. Oláh, J., Kovács, S., Virglerova, Z., Lakner, Z., Kovacova, M. and Popp, J. (2019). Analysis and Comparison of Economic and Financial Risk Sources in SMEs of the Visegrad Group and Serbia, Sustainability, MDPI, *Open Access Journal*, 11(7), 1-19.
22. Peker, E. (2008), Scenario Supported Value at Risk Applications, *Master of Thesis*, İstanbul Technical University.
23. Santos, R.B., Lima, F.G., Gatsios, R.C. and Almeida, R.B. (2017). Risk management and value creation: new evidence for Brazilian non-financial companies, *Applied Economics*, 49(58), 5815-5827, DOI: 10.1080/00036846.2017.1343451.
24. Srinivasan, S. and Kamalakannan, T. (2018). Multi Criteria Decision Making in Financial Risk Management with a Multi-objective Genetic Algorithm, *Comput Econ*, 52, 443-457, Available at: https://doi.org/10.1007/s10614-017-9683-7.
25. Şencal, H. (2007). Value-At-Risk Applications Before and after the Global Financial Crisis of 2007, *Doctor of Philosophy in Economics*, Fatih University.
26. Terpezan-Tabara, O.A. (2008). The Importance of Value at Risk Method in the Management of Banking Risk. In *4th International Conference of ASECU: "Development Cooperation and Competitiveness"*, Bucharest, Romania (pp. 22-24).

27. Tibiletti L. (2001) Incremental Value-at-Risk: traps and misinter-pretations. In: Kohlmann M., Tang S. (eds), *Mathematical Finance*. Trends in Mathematics. Birkhäuser, Basel. https://doi.org/10.1007/978-3-0348-8291-0_3
28. Uylangco, K. (2015). An evaluation of the effectiveness of Value-at-Risk (VaR) models for Australian banks under Basel III. *Australian Journal of Management*, 41(4), 699-718.
29. Wang, Z. (2002). The Properties of Incremental VaR in Monte Carlo Simulations. *Journal of Risk Finance*, 3(3), 14-23. Available at: https://doi.org/10.1108/eb043491.
30. Yang, Q., Wang, Y. and Ren, Y. (2019). Research on financial risk management model of internet supply chain based on data science. *Cognitive Systems Research*, 56, 50-55. Available at: https://doi.org/10.1016/j.cogsys.2019.02.001.
31. Zhang, X. and Jiang, H. (2019). Application of Copula function in financial risk analysis. *Computers & Electrical Engineering*, 77, 376-388. Available at: https://doi.org/10.1016/j.compeleceng.2019.06.011.

Proc. 18th International Conference on Statistical Sciences
Lahore, Pakistan – February 18-20, 2021, Vol. 35, pp. 427-430

BOUNDED RATIONALITY
FAST AND FRUGAL HEURISTICS FOR MAKING DECISIONS UNDER RISK

Laura Martignon
Ludwigsburg University of Education, Germany
Email: martignon@ph-ludwigsburg.de

ABSTRACT

Bounded rationality is a term coined by H. Simon, describing the theory that an individual's decision-making ability is limited by a set of constraints such as lack of information, cognitive biases or lack of time. From this perspective, decision-makers tend to choose “satisficing” solutions rather than optimal ones. In particular in difficult times of high risk pandemics, as we experience today, bounded rationality seems the ideal way for dealing with uncertainty.

1. INTRODUCTION

The tenor of this text is conceptual and aims at clarifying the use of certain labels and methodologies in the interdisciplinary field of Deciaion Making under Uncertainty. The concept was initially theorized by Herbert Simon and used in sociology, psychology, microeconomics and political philosophy. It is a departure from one of the primary assumptions of microeconomics and/or rational choice theory, which is that the actor is fully rational with respect to the situation under consideration. Gerd Gigerenzer has consistently pointed out, however, that for Simon, bounded rationality is no less good than full rationality. The mind adapts to the environment and the environment adapts to the human mind, according to the “two blades of a scissor” metaphor postulated by Simon.

Historically, reasoning and inference have always been seen as basic components of the decision process. Logic was postulated as the main ingredient for good reasoning by the Greeks. During the Middle Ages logic and faith appeared to melt in good decisions. Descartes, considered the modern father of rationalism, explained rationality in terms of the “method” that applies logic to “clear and distinct” ideas. He thus separated rational thinking from faith. Laplace went one step further and formulated a model of rational thinking that combines logic and elements of the theory of probability. The “homo oeconomicus” was later conceived as a utility maximizer in economics and game theory.

In macroeconomics, this model is also often used as a so-called representative agent to analyze economic processes. A frequently used special case of homo economicus is the time-consistent expected utility maximizer, which is the subject of behavioral economics in particular.

Herbert Simon’s view on decision making was one of the agent constrained by limited information and limited time. While in classical economic theory, the goal of any action is to maximize the benefits that result from it he stated that if individual finds an action that satisfices his “drive”, the search for further alternative actions is stopped. In a decision situation, according to Simon, often not all alternative actions and their consequences are given, but must first be worked out and calculated with the help of information.

2. DECISION MAKING UNDER RISK

Martignon & Hoffrage 82019) have postulated that the construct "Risk Literacy" can be decomposed in 4 levels or components, as in Figure 1.

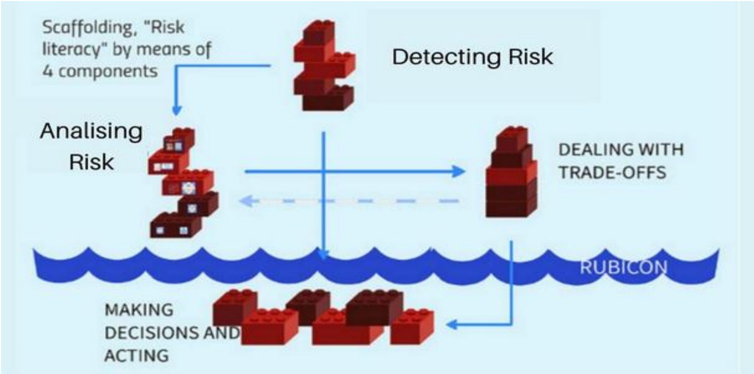


Figure 1: Scaffolding Risk Literacy by means of Four Levels

In what follows a brief description of these 4 levels is presented:

Competence level 1: **Recognizing uncertainties and risks**. Risk competence must necessarily include the ability to recognize risks as such in the first place. Risky situations are those in which something is at stake and in which one can lose something that one would not like to lose. If one is not aware of this possibility in a certain situation or when making a certain decision, one will hardly protect oneself or take precautionary measures. Or, if necessary, one gets involved in something that one would not have done if one had recognized the risk. Conversely, this does not mean that someone who takes a risk is not risk competent. It is often worth taking a risk - especially if the loss is small and unlikely, but in return a high and probable profit beckons. Here it is then necessary to weigh up - but you can only weigh up what you are aware of or at least intuitively know about. Risks which, in such a mental bar scale, do not lie in that bowl which stands for the contrasting side of an option for action also do not carry weight - and this can have unpleasant consequences.

Competency Level 2: **Analyzing and Modeling risky situations**. An analysis is a breakdown of a whole into parts. To do this, we need "to distinguish and separate", and this in turn is an activity of our mind. A risky situation can definitely be analyzed. What is the danger in it - what exactly can happen? And how often does this or that event occur within a certain reference class? In most cases, one can not only count different events, situations, objects, persons or other separable instances, but one can also describe them in detail. What properties, characteristics or symptoms do they have? On the basis of such characteristics, one can then classify and sort them. And this, in turn, allows us to inquire about correlations: which characteristics point to which possible outcome, and how meaningful are these characteristics? This necessary leads to the necessity of dealing with probabilities and conditional probabilities. How big is a risk in one condition and how big is it in another? Modeling a situation means first simplifying it and stripping it of all unimportant details - where the distinction between what is important and what is unimportant is already part of the model. Certain features are pulled out when assessing risk, while others are not.

Competence level 3: **Weighing and comparing**. What was dissected in the analysis at level 2 is now presented by means of possible alternatives. Here the one option, there the other. This is the opportunities (or advantages) of the first option, that is its risks (or disadvantages). Do the benefits outweigh the disadvantages? The same can be asked for the second, third and so on. Option for action. Here something like “synopsis” and “overview” are required, otherwise one cannot judge which option is the best. What the mind has separated in stage 2, reason must therefore somehow put together again in stage 3. Incidentally, this distinction - it is that between analysis and synthesis - is well known in many scientific disciplines, for example in chemistry or philosophy. These two activities of our mind complement each other and play together - much like pairs of muscles where we can raise our arm with one and lower it with the other. As such, then, they belong together, and so we have also shown Competency Levels 2 and 3 in the middle - and side by side - in the figure. There they stand between recognizing and deciding, which in turn can be described as the beginning and the end points of the process that takes place in the middle in the interplay between analysis and synthesis.

Competence level 4: **Deciding and acting**. In Levels 1 and 2, you “look at”, represent, analyze and model the world out there - and you form a judgment at Level 3. All of this happens in the mind before converting decisions into actions.

From the etymological perspective to de-cide means "to separate”. By saying "yes" to one alternative, one says "no" to another. Many of our decisions can be corrected later, but not all - some are irrevocable and irreversible. This makes it particularly difficult, and it takes courage to “cross the Rubicon”. Not every decision has to be based on careful analysis and judgment - often we act impulsively or intuitively or, according to the tenets of Bounded Rationality, we often based on satisficing heuristics, like fast-and-frugal trees for decision.

3. FAST AND FRUGAL TREES FOR DECISION MAKING

Decisioan are often the consequence of classifications. A doctor classifies a patient as “high risk” and the decision is then to send the patient to an intensive care unit at a hospital. Competency levels 1 and 2 are fundamental in detecting risks and analyzing faetures of the situation for a classification. The boundedly rational aspect of fast-and-frugal decision trees is that instead of using a general Bayesian analysis of a regression model for extracting consequences from a set of features, a lexicographic decision procedure is used, which we exemplify with the diagram in Figure 2, which reprsents a fast and frugal tree often used by medical doctors for deciding whether a patient with a positive Covid test should be assigned to a Care Unit at a hospital or to his/her home bed.

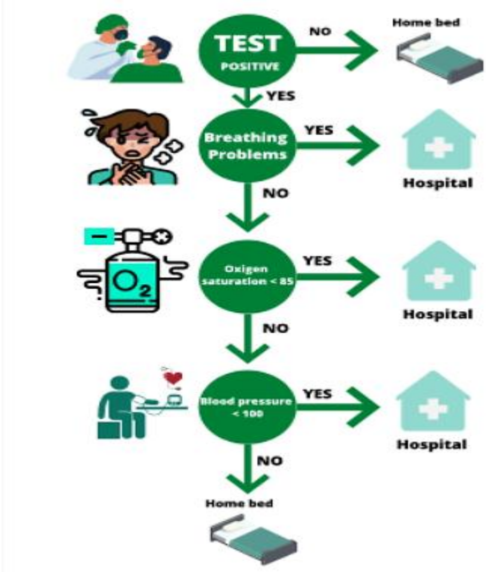


Figure 2: A Fast and Frugal Tree for Assigning Patients with Positive Tests to Hospital Care Units or a Home Bed

4. COMMENTS AND CONCLUSION

In this brief presentation the emphasis has been on "bounded-ly rational", fast and frugal decision making, to be used in cases of risk and uncertainty. It has been widely analyzed by sophisticated methods (Laskey and Martignon, 2014; Martignon and Laskey, 2019) and it has been demonstrated that such decision strategies are highly predictive and robust compared to more sophisticated standards methods.

5. ACKNOWLEDGEMENT

The author is grateful to the Organizers of the Proc. 18th International Conference on Statistical Sciences in Lahore, Pakistan especially to Prof. Saleha Habibullah and Dr. M. Iftikhar for organizing and producing the Proceedings of the Conference.

REFERENCES

1. Wikipedia. *Fast-and-frugal trees*. https://en.wikipedia.org/wiki/Fast-and-frugal_trees
2. Martignon, L. and Laskey, K. (2019). Statistical literacy for classification under risk: an educational perspective. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 13(3), 269-278.
3. Laskey, K.B. and Martignon, L. (2014). Comparing fast and frugal trees and bayesian networks for risk assessment. In Sustainability in statistics education. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).*