



IntechOpen

# Recent Advances in Biostatistics

*Edited by B. Santhosh Kumar*





---

# Recent Advances in Biostatistics

*Edited by B. Santhosh Kumar*

Published in London, United Kingdom

---

Recent Advances in Biostatistics

<http://dx.doi.org/10.5772/intechopen.1000344>

Edited by B. Santhosh Kumar

#### Contributors

Ajith Wickramasinghe, Ankita Pal, Anusha Jayasiri, Aseem Mishra, Atanu Bhattacharjee, Barbara J. Lutz, Dechang Chen, Dieter Rasch, Dilip Kumar Ghosh, Hazhar Talaat Abubaker Blbas, Huan Wang, Joanne N. Halls, L. Rob Verdooren, Li Sheng, Matthew A. Psioda, Olubunmi Alabi, Pankaj Chaturvedi, Sara B. Jones, Satyajit Pradhan, Tosin Bukola

© The Editor(s) and the Author(s) 2024

The rights of the editor(s) and the author(s) have been asserted in accordance with the Copyright, Designs and Patents Act 1988. All rights to the book as a whole are reserved by INTECHOPEN LIMITED. The book as a whole (compilation) cannot be reproduced, distributed or used for commercial or non-commercial purposes without INTECHOPEN LIMITED's written permission. Enquiries concerning the use of the book should be directed to INTECHOPEN LIMITED rights and permissions department ([permissions@intechopen.com](mailto:permissions@intechopen.com)).

Violations are liable to prosecution under the governing Copyright Law.



Individual chapters of this publication are distributed under the terms of the Creative Commons Attribution 3.0 Unported License which permits commercial use, distribution and reproduction of the individual chapters, provided the original author(s) and source publication are appropriately acknowledged. If so indicated, certain images may not be included under the Creative Commons license. In such cases users will need to obtain permission from the license holder to reproduce the material. More details and guidelines concerning content reuse and adaptation can be found at <http://www.intechopen.com/copyright-policy.html>.

#### Notice

Statements and opinions expressed in the chapters are those of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2024 by IntechOpen

IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 167-169 Great Portland Street, London, W1W 5PF, United Kingdom

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from [orders@intechopen.com](mailto:orders@intechopen.com)

Recent Advances in Biostatistics

Edited by B. Santhosh Kumar

p. cm.

Print ISBN 978-1-83769-805-9

Online ISBN 978-1-83769-804-2

eBook (PDF) ISBN 978-1-83769-806-6

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

7,000+

Open access books available

186,000+

International authors and editors

200M+

Downloads

156

Countries delivered to

Our authors are among the  
**Top 1%**

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)





# Meet the editor



Dr. B. Santhosh Kumar is a Professor and Head at Guru Nanak Institute of Technology, Hyderabad, India. He has 19 years of teaching and research experience. He obtained a BEng in Computer Science and Engineering from Bharathiar University, India, in 2004 and an MEng with a specialization in computer science and engineering from Anna University, India, in 2011. He also obtained a Ph.D. from Anna University in 2018. Dr.

Kumar is currently pursuing a post-doctoral fellowship at the University of Louisiana, USA, and a research fellowship at INTL International University, Malaysia. His research interests include data science, machine learning, blockchain technology, and data mining. He has 122 publications to his credit, including journal articles and conference publications. He has filed thirteen patents and one copyright. He has authored fourteen books and five book chapters. He has delivered numerous guest lectures, webinars, and keynote speeches and received eighteen awards from various professional bodies. He is a reviewer of more than 1200 papers for international journals and is a session chair for various international conferences. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), appointed as an ACM Distinguished Speaker, a fellow member of the Institute of Engineers, and a life member of the Indian Society for Technical Education. Dr. Kumar was recognized as a 2019 Researcher of the Year by the *World Book of Researchers*.





# Contents

<b>Preface</b>	<b>XI</b>
<b>Chapter 1</b> Introduction to Descriptive Statistics <i>by Olubunmi Alabi and Tosin Bukola</i>	<b>1</b>
<b>Chapter 2</b> Descriptive Statistics <i>by Hazhar Talaat Abubaker Blbas</i>	<b>13</b>
<b>Chapter 3</b> Spatial Statistics: A GIS Methodology to Investigate Point Patterns in Stroke Patient Healthcare <i>by Joanne N. Halls, Barbara J. Lutz, Sara B. Jones and Matthew A. Psioda</i>	<b>23</b>
<b>Chapter 4</b> Statistical Model for the Quality of Panoramic Images of 2D Artifacts <i>by Ajith Wickramasinghe and Anusha Jayasiri</i>	<b>45</b>
<b>Chapter 5</b> Useful Block Designs in Biostatistics <i>by L. Rob Verdooren and Dieter Rasch</i>	<b>65</b>
<b>Chapter 6</b> Perspective Chapter: Using Effect Sizes to Study the Survival Difference between Two Groups <i>by Huan Wang, Li Sheng and Dechang Chen</i>	<b>99</b>
<b>Chapter 7</b> Perspective Chapter: Linear Regression and Logistic Regression Models <i>by Dilip Kumar Ghosh</i>	<b>113</b>
<b>Chapter 8</b> QoLMiss: Package for Repeatedly Measured Quality of Life of Cancer Patients Data <i>by Ankita Pal, Satyajit Pradhan, Aseem Mishra, Pankaj Chaturvedi and Atanu Bhattacharjee</i>	<b>135</b>



# Preface

In the dynamic landscape of biomedical research, the field of biostatistics plays a pivotal role in unravelling complex patterns, extracting meaningful insights, and guiding evidence-based decision-making. As we stand at the intersection of data science, health care, and statistical methodologies, the compilation of *Recent Advances in Biostatistics* emerges as a testament to the relentless pursuit of excellence in statistical approaches within the realm of life sciences.

This anthology is a collaborative effort that brings together the expertise and perspectives of accomplished researchers, statisticians, and practitioners who have dedicated their efforts to pushing the boundaries of biostatistics. It provides a comprehensive overview of the latest methodologies, innovations, and applications that have transformed the landscape of statistical analysis in the biomedical arena.

The chapters within this volume explore a diverse array of topics, ranging from cutting-edge statistical techniques for analyzing high-dimensional biological data to advanced methods in clinical trial design and analysis. The contributors delve into the challenges posed by modern healthcare datasets, such as electronic health records and genomics, and propose novel solutions that enhance our ability to draw robust inferences and make informed decisions in the face of uncertainty.

Furthermore, this collection reflects the interdisciplinary nature of contemporary biostatistics, highlighting collaborations with epidemiologists, clinicians, bioinformaticians, and other stakeholders. The synergistic integration of statistical methods with domain-specific knowledge is emphasized throughout, reinforcing the notion that successful advances in biostatistics require a holistic and collaborative approach.

As editors, we are grateful to the esteemed authors who have shared their expertise and insights, contributing to the richness and diversity of this compilation. We hope that this volume serves as a valuable resource for researchers, students, and practitioners alike, fostering a deeper understanding of the latest developments in biostatistics and inspiring future innovations.

*Recent Advances in Biostatistics* serves as a snapshot of the ever-evolving landscape of statistical methodologies in the life sciences. We invite readers to embark on a journey through the pages of this collection, exploring the forefront of biostatistical research and its profound impact on advancing our understanding of health and disease.

**Dr. B. Santhosh Kumar**

Professor, Head of Department,  
Department of Computer Science and Engineering,  
Guru Nanak Institute of Technology,  
Hyderabad, Telangana, India



## Chapter 1

# Introduction to Descriptive Statistics

*Olubunmi Alabi and Tosin Bukola*

### Abstract

This chapter offers a comprehensive exploration of descriptive statistics, tracing its historical development from Condorcet's "average" concept to Galton and Pearson's contributions. Emphasizing its pivotal role in academia, descriptive statistics serve as a fundamental tool for summarizing and analyzing data across disciplines. The chapter underscores how descriptive statistics drive research inspiration and guide analysis, and provide a foundation for advanced statistical techniques. It delves into their historical context, highlighting their organizational and presentational significance. Furthermore, the chapter accentuates the advantages of descriptive statistics in academia, including their ability to succinctly represent complex data, aid decision-making, and enhance research communication. It highlights the potency of visualization in discerning data patterns and explores emerging trends like large dataset analysis, Bayesian statistics, and nonparametric methods. Sources of variance intrinsic to descriptive statistics, such as sampling fluctuations, measurement errors, and outliers, are discussed, stressing the importance of considering these factors in data interpretation.

**Keywords:** academic research, data analysis, data visualization, decision-making, research methodology, data summarization

## 1. Introduction

The French mathematician and philosopher Condorcet established the idea of the "average" as a means to summarize data, which is when descriptive statistics got their start. Yet, the widespread use of descriptive statistics in academic study did not start until the 19th century. Francis Galton, who was concerned in the examination of human features and attributes, was one of the major forerunners of descriptive statistics. Galton created various statistical methods that are still frequently applied in academic research today, such as the correlation and regression analysis concepts. The American statistician and mathematician in the early 20th century Karl Pearson created the "normal distribution," which is a bell-shaped curve that characterizes the distribution of many natural occurrences. Moreover, Pearson created a number of correlational measures and popularized the chi-square test, which evaluates the significance of variations between observed and predicted frequencies. With the advent of new methods like multivariate analysis and factor analysis in the middle of the 20th century, the development of electronic computers sparked a revolution in

statistical analysis. Descriptive statistics is the analysis and summarization of data to gain insights into its characteristics and distribution [1].

Descriptive statistics help researchers generate study ideas and guide further analysis by allowing them to explore data patterns and trends [2]. Descriptive statistics were used more often in academic research because they helped researchers better comprehend their datasets and served as a basis for more sophisticated statistical techniques. Similarly, Descriptive statistics are used to summarize and analyze data in a variety of academic areas, including psychology, sociology, economics, education, and epidemiology [3]. Descriptive statistics continue to be a crucial research tool in academia today, giving researchers a method to compile and analyze data from many fields. It is now simpler than ever to analyze and understand data, enabling researchers to make better informed judgments based on their results. This is due to the development of new statistical techniques and computer tools. Descriptive statistics can benefit researchers in hypothesis creation and exploratory analysis by identifying trends, patterns, and correlations between variables in huge datasets [4]. Descriptive statistics are important in data-driven decision-making processes because they allow stakeholders to make educated decisions based on reliable data [5].

## **2. Background**

The history of descriptive statistics may be traced back to the 17th century, when early pioneers like John Graunt and William Petty laid the groundwork for statistical analysis [6]. Descriptive statistics is a fundamental concept in academia that is widely used across many disciplines, including social sciences, economics, medicine, engineering, and business. Descriptive statistics provides a comprehensive background for understanding data by organizing, summarizing, and presenting information effectively [7]. In academia, descriptive statistics is used to summarize and analyze data, providing insights into the patterns, trends, and characteristics of a dataset. Similarly, in academic research, descriptive statistics are often used as a preliminary analysis technique to gain a better understanding of the dataset before applying more complex statistical methods. Descriptive statistics lay the groundwork for inferential statistics by assisting researchers in drawing inferences about a population based on observed sample data [8]. Descriptive statistics aid in the identification and analysis of outliers, which can give useful insights into unusual observations or data collecting problems [9].

Descriptive statistics enable researchers to synthesize both quantitative and qualitative data, allowing for a thorough examination of factors [10]. Descriptive statistics can provide valuable information about the central tendency, variability, and distribution of the data, allowing researchers to make informed decisions about the appropriate statistical techniques to use. Descriptive statistics are an essential component of survey research technique, allowing researchers to efficiently summarize and display survey results [11]. Descriptive statistics may be used to summarize data as well as spot outliers, or observations that dramatically depart from the trend of the data as a whole. Finding outliers can help researchers spot any issues or abnormalities in the data so they can make the necessary modifications or repairs. In academic research, descriptive statistics are frequently employed to address research issues and evaluate hypotheses. Descriptive statistics, for instance, can be used to compare the average scores of two groups to see if there is a significant difference between them. In order to create new hypotheses or validate preexisting ideas, descriptive statistics may also be used to find patterns and correlations in the data.

There are several sources of variation that can affect the descriptive statistics of a data set, some of which include: Sampling Variation, descriptive statistics are often calculated using a sample of data rather than the entire population. Therefore, the descriptive statistics can vary depending on the particular sample that is selected. This is known as sampling variation. Measurement Variation, different measurement methods can produce different results, leading to variation in descriptive statistics. For example, if a scale is used to measure the weight of objects, slight differences in how the scale is used can produce slightly different measurements.

Data entry errors are mistakes made during the data entry process which can lead to variation in descriptive statistics. Even small errors, such as transposing two digits, can significantly impact the results. Outliers, Outliers are extreme values that fall outside of the expected range of values. These values can skew the descriptive statistics, making them appear more or less extreme than they actually are. Natural Variation, Natural variation refers to the inherent variability in the data itself. For example, if a data set contains measurements of the heights of trees, there will naturally be variation in the heights of the trees. It is important to understand these sources of variation when interpreting and using descriptive statistics in academia. Properly accounting for these sources of variation can help ensure that the descriptive statistics accurately reflect the underlying data.

Some emerging patterns in descriptive statistics in academia include: Big data analysis, with the increasing availability of large data sets, researchers are using descriptive statistics to identify patterns and trends in the data. The use of big data analysis techniques, such as machine learning and data mining, is becoming more common in academic research. Visualization techniques, advances in data visualization techniques are enabling researchers to more easily identify patterns in data sets. For example, heat maps and scatter plots can be used to visualize the relationship between different variables. Bayesian statistics is an emerging area of research in academia, which involves using probability theory to make inferences about data. Bayesian statistics can provide more accurate estimates of descriptive statistics, particularly when dealing with complex data sets.

Non-parametric statistics are becoming increasingly popular in academia, particularly when dealing with data sets that do not meet the assumptions of traditional parametric statistical tests. Non-parametric tests do not require the data to be normally distributed, and can be more robust to outliers. Open science practices, such as pre-registration and data sharing, are becoming more common in academia. This is enabling researchers to more easily replicate and verify the results of descriptive statistical analyses, which can improve the quality and reliability of research findings. Overall, the emerging patterns in descriptive statistics in academia reflect the increasing availability of data, the need for more accurate and robust statistical techniques, and a growing emphasis on transparency and openness in research practices.

### **3. Benefits of descriptive statistics**

The advantages of descriptive statistics extend beyond research and academia, with applications in commercial decision-making, public policy, and strategic planning [12]. The benefits of descriptive statistics include providing a clear and concise summary of data, aiding in decision-making processes, and facilitating effective communication of findings [13]. Descriptive statistics provide numerous benefits to

academia, some of which include: **Summarization of Data:** descriptive statistics allow researchers to quickly and efficiently summarize large data sets, providing a snapshot of the key characteristics of the data. This can help researchers identify patterns and trends in the data, and can also help to simplify complex data sets. **Better decision making:** descriptive statistics can help researchers make data-driven decisions. For example, if a researcher is comparing the effectiveness of two different treatments, descriptive statistics can be used to identify which treatment is more effective based on the data. **Visualization of data:** descriptive statistics can be used to create visualizations of data, which can make it easier to communicate research findings to others.

Histograms, bar charts, and scatterplots are examples of data visualization techniques that may be used to graphically depict data in order to detect trends, outliers, and correlations [14]. Visualizations can also help to identify patterns and trends in the data that might not be immediately apparent from raw data. **Hypothesis Testing:** descriptive statistics are often used in hypothesis testing, which allows researchers to determine whether a particular hypothesis about a data set is supported by the data. This can help to validate research findings and increase confidence in the conclusions drawn from the data. **Improved data quality:** Descriptive statistics can help to identify errors or inconsistencies in the data, which can help researchers improve the quality of the data. This can lead to more accurate research findings and a better understanding of the underlying phenomena. Overall, the benefits of descriptive statistics in academia are many and varied. They help researchers summarize large data sets, make data-driven decisions, visualize data, validate research findings, and improve the quality of the data. By using descriptive statistics, researchers can gain valuable insights into complex data sets and make more informed decisions based on the data.

#### **4. Practical applications of descriptive statistics**

Descriptive statistics has practical applications in disciplines such as business, social sciences, healthcare, finance, and market research [15]. Descriptive statistics have a wide range of practical applications in academia, some of which include: **Data Summarization:** Descriptive statistics can be used to summarize large data sets, making it easier for researchers to understand the key characteristics of the data. This is particularly useful when dealing with complex data sets that contain many variables. **Hypothesis Testing:** Descriptive statistics can be used to test hypotheses about a data set. For example, researchers can use descriptive statistics to test whether the mean value of a particular variable is significantly different from a hypothesized value. **Data visualization:** descriptive statistics can be used to create visualizations of data, which can make it easier to identify patterns and trends in the data. For example, a histogram or boxplot can be used to visualize the distribution of a variable. **Comparing Groups:** Descriptive statistics can be used to compare different groups within a data set. For example, researchers may compare the mean values of a particular variable between different demographic groups, such as age or gender. **Predictive modeling:** Descriptive statistics can be used to build predictive models, which can be used to forecast future trends or outcomes. For example, a researcher might use descriptive statistics to identify the key variables that predict student performance in a particular course. The practical applications of descriptive statistics in academia are wide-ranging and varied. They can be used in many different fields, including psychology, economics, sociology, and biology, among others, to provide insights into complex data sets and help researchers make data-driven decisions (**Figure 1**).





**Figure 1.**  
*Types of descriptive statistics. Ref: <https://www.analyticssteps.com/blogs/types-descriptive-analysis-examples-steps>.*

Descriptive statistics is a useful tool for researchers in a variety of sectors since it allows them express the major characteristics of a dataset, such as its frequency, central tendency, variability, and distribution.

#### 4.1 Central tendency measurements

Central tendency metrics, such as mean, median, and mode, are essential descriptive statistics that offer information about the average or typical value in a collection [16]. One of the primary purposes of descriptive statistics is to summarize data in a succinct and useful manner. Measures of central tendency, such as the median, are resistant to outliers and offer a more representative assessment of the average value in a skewed distribution [17]. The mean, median, and mode are measures of central tendency that are used to characterize the usual or center value of a dataset. The mean of a dataset is the arithmetic average, but the median is the midway number when the data is ordered in order of magnitude. The mode is the most often occurring value in the collection. Central tendency measurements are one of the most important aspects of descriptive statistics, as they provide a summary of the “typical” value of a data set.

The three most commonly used measures of central tendency are: Mean: the mean is calculated by adding up all the values in a data set and dividing by the total number of values. The mean is sensitive to outliers, as even one extreme value can greatly affect the mean. Median: the median is the middle value in a data set when the values are ordered from smallest to largest. If the data set has an odd number of values, the median is the middle value. If the data set has an even number of values, the median is the average of the two middle values. The median is more robust to outliers than the mean. Mode: the mode is the most common value in a data set. In some cases, there may be multiple modes (i.e. bimodal or multimodal distributions). The mode is useful for identifying the most frequently occurring value in a data set. Each of these measures of central tendency provides a different perspective on the “typical” value of a data set, and which measure is most appropriate to use depends on the nature of the data and the research question being addressed. For example, if the data set contains extreme outliers, the median may be a better measure of central tendency than the mean. Conversely, if the data set is symmetrical and normally distributed, the mean may provide the best measure of central tendency.

#### 4.2 Variability indices

It is another key part of descriptive statistics is determining data variability. The spread or dispersion of data points about the central tendency readings is quantified by variability indices such as range, variance, and standard deviation [18]. Variability

measures, such as range, variance, and standard deviation, reveal information about the spread or dispersion of the data. Variability indices, such as the coefficient of variation, allow you to compare variability across various datasets with different scales or units of measurement [19]. The range is the distance between the dataset's greatest and lowest values, and the variance and standard deviation are measures of how much the data values depart from the mean. Variability indices are measures used in descriptive statistics to provide information about how much the data varies or how spread out it is. Variability indices, such as the interquartile range, give insights into data distribution while being less impacted by extreme values than the standard deviation [20]. Some commonly used variability indices include:

**Range:** The range is the difference between the largest and smallest values in a data set. It provides a simple measure of the spread of the data, but is sensitive to outliers. **Interquartile Range (IQR):** The IQR is the range of the middle 50% of the data. It is calculated by subtracting the 25th percentile (lower quartile) from the 75th percentile (upper quartile). The IQR is more robust to outliers than the range. **Variance:** The variance is a measure of how spread out the data is around the mean. It is calculated by taking the average of the squared differences between each data point and the mean. The variance is sensitive to outliers. **Standard Deviation:** The standard deviation is the square root of the variance. It provides a measure of how much the data varies from the mean, and is more commonly used than the variance because it has the same units as the original data.

**Coefficient of Variation (CV):** The CV is a measure of relative variability, expressed as a percentage. It is calculated by dividing the standard deviation by the mean and multiplying by 100. The CV is useful for comparing variability across different data sets that have different units or scales. These variability indices provide important information about the spread and variability of the data, which can help researchers better understand the characteristics of the data and draw meaningful conclusions from it.

### **4.3 Data visualization**

Data may be visually represented using graphical approaches in addition to numerical metrics. Graphs and charts, such as histograms, box plots, and scatterplots, allow researchers investigate data patterns and correlations. Box plots and violin plots are efficient data visualization approaches for showing data distribution and spotting potential outliers [21]. They may also be used to detect outliers, or data points that deviate dramatically from the rest of the data. Data visualization is an important aspect of descriptive statistics, as it allows researchers to communicate complex data in a visual and easily understandable format. Some common types of data visualization used in descriptive statistics include: **Histograms:** Histograms are used to display the distribution of a continuous variable. The data is divided into intervals (or "bins"), and the number of observations falling into each bin is displayed on the vertical axis. Histograms provide a visual representation of the shape of the distribution, and can help to identify outliers or skewness. **Box plots:** Box plots provide a graphical representation of the distribution of a continuous variable. The application of graphical approaches, such as scatterplots and heat maps, improves comprehension of correlations and patterns in large datasets [22].

The box represents the middle 50% of the data, with the median displayed as a horizontal line inside the box. The whiskers extend to the minimum and maximum

values in the data set, and any outliers are displayed as points outside the whiskers. Box plots are useful for comparing distributions across different groups or for identifying outliers. Scatter plots: Scatter plots are used to display the relationship between two continuous variables. Each data point is represented as a point on the graph, with one variable displayed on the horizontal axis and the other variable displayed on the vertical axis. Scatter plots can help to identify patterns or relationships in the data, such as a positive or negative correlation. Bar charts: Bar charts are used to display the distribution of a categorical variable.

The categories are displayed on the horizontal axis, and the frequency or percentage of observations falling into each category is displayed on the vertical axis. Bar charts can help to compare the frequency of different categories or to display the results of a survey or questionnaire. Heat maps: Heat maps are used to display the relationship between two categorical variables. The categories are displayed on both the horizontal and vertical axes, and the frequency or percentage of observations falling into each combination of categories is displayed using a color scale. Heat maps can help to identify patterns or relationships in the data, such as a higher frequency of observations in certain combinations of categories. These types of data visualizations can help researchers to communicate complex data in a clear and understandable format, and can also provide insights into the characteristics of the data that may not be immediately apparent from the raw data.

#### **4.4 Data cleaning and preprocessing**

Data cleaning and preprocessing procedures, such as imputation methods for missing data, aid in the preservation of data integrity and the reduction of bias in descriptive analysis [23]. Before beginning any statistical analysis, be certain that the data is clean and well arranged. The process of discovering and fixing flaws or inconsistencies in data, such as missing numbers or outliers, is known as data cleaning. Data preparation is the process of putting data into an appropriate format for analysis, such as scaling or normalizing the data. Data cleaning and preprocessing are essential steps in descriptive analysis, as they help to ensure that the data is accurate, complete, and ready for analysis. Some common data cleaning and preprocessing steps include: Handling missing data: Missing data can be a common problem in datasets and can impact the accuracy of the analysis. Depending on the amount of missing data, researchers may choose to remove incomplete cases or impute missing values using techniques such as mean imputation, regression imputation, or multiple imputation. Handling outliers: Outliers are extreme values that are different from the majority of the data points and can distort the analysis. Outlier identification and removal procedures, for example, assist increase the accuracy and reliability of descriptive statistics [24].

To assure the correctness and dependability of descriptive statistics, data cleaning and preprocessing require finding and dealing with missing values, outliers, and data inconsistencies [25]. Researchers may choose to remove or transform outliers to better reflect the characteristics of the data. Data transformation: Data transformation is used to normalize the data or to make it easier to analyze. Common transformations include logarithmic, square root, or Box-Cox transformations. Handling categorical data: Categorical data, such as nominal or ordinal data, may need to be recoded into numerical data before analysis. Researchers may also need to handle missing or inconsistent categories within the data. Standardizing data: Standardizing data

involves scaling the data to have a mean of zero and a standard deviation of one. This can be useful for comparing variables with different units or scales. Data integration: Data integration involves merging or linking multiple datasets to create a single, comprehensive dataset for analysis. This may involve matching or merging datasets based on common variables or identifiers. By performing these data cleaning and preprocessing steps, researchers can ensure that the data is accurate and ready for analysis, which can lead to more reliable and meaningful insights from the data.

## **5. Descriptive statistics in academic methodology**

Descriptive statistics are important in academic technique because they enable researchers to synthesize and describe data collected for research objectives [26]. Descriptive statistics is often used in combination with other statistical techniques, such as inferential statistics, to draw conclusions and make predictions from the data. In academic research, descriptive statistics is used in a variety of ways, such as describing sample characteristics. Descriptive statistics is used to describe the characteristics of a sample, such as the mean, median, and standard deviation of a variable. This information can be used to identify patterns, trends, or differences within the sample. Identifying data outliers: Descriptive statistics can help researchers identify potential outliers or anomalies in the data, which can affect the validity of the results. For example, identifying extreme values in a dataset can help researchers to investigate whether these values are due to measurement error or a true characteristic of the population.

Communicating research findings: Descriptive statistics is used to summarize and communicate research findings in a clear and concise manner. Graphs, charts, and tables can be used to display descriptive statistics in a way that is easy to understand and interpret. Testing assumptions: Descriptive statistics can be used to test assumptions about the data, such as normality or homogeneity of variance, which are important for selecting appropriate statistical tests and interpreting the results. Overall, descriptive statistics is a critical methodology in academic research that helps researchers to describe and understand the characteristics of their data. By using descriptive statistics, researchers can draw meaningful insights and conclusions from their data, and communicate these findings to others in a clear and concise manner.

## **6. Pitfalls of descriptive statistics**

The possibility for misunderstanding, reliance on summary measures alone, and susceptibility to high values or outliers are all disadvantages of descriptive statistics [27]. While descriptive statistics is an essential tool in academic statistics, there are several potential pitfalls that researchers should be aware of: Limited scope: Descriptive statistics can provide a useful summary of the characteristics of a dataset, but it is limited in its ability to provide insights into the underlying causes or mechanisms that drive the data. Descriptive statistics alone cannot establish causal relationships or test hypotheses. Misleading interpretations: Descriptive statistics can be misleading if not interpreted correctly. For example, a small sample size may not accurately represent the population, and summary statistics such as the mean may not be meaningful if the data is not normally distributed.

Incomplete analysis: Descriptive statistics can only provide a limited view of the data, and researchers may need to use additional statistical techniques to fully analyze

the data. For example, hypothesis testing and regression analysis may be needed to establish relationships between variables and make predictions. Biased data: Descriptive statistics can be biased if the data is not representative of the population of interest. Sampling bias, measurement bias, or non-response bias can all impact the validity of descriptive statistics. Over-reliance on summary statistics: Descriptive statistics can be over-reliant on summary statistics such as the mean or median, which may not provide a complete picture of the data. Visualizations and other descriptive statistics, such as measures of variability, can provide additional insight into the data. To avoid these pitfalls, researchers should carefully consider the scope and limitations of descriptive statistics and use additional statistical techniques as needed. They should also ensure that their data is representative of the population of interest and interpret their descriptive statistics in a thoughtful and nuanced manner.

## **7. Conclusion**

Researchers can test the normalcy assumptions of their data by using relevant descriptive statistics techniques such as measures of skewness and kurtosis [28]. Descriptive statistics has become a fundamental methodology in academic research that is used to summarize and describe the characteristics of a dataset, such as the central tendency, variability, and distribution of the data. It is used in a wide range of disciplines, including social sciences, natural sciences, engineering, and business. Descriptive statistics can be used to describe sample characteristics, identify data outliers, communicate research findings, and test assumptions. The kind of data, research topic, and particular aims of the study all influence the right choice and implementation of descriptive statistical approaches [29].

However, there are several potential pitfalls of descriptive statistics, including limited scope, misleading interpretations, incomplete analysis, biased data, and over-reliance on summary statistics. The use of descriptive statistics in data presentation can improve the interpretability of study findings, making complicated material more accessible to a larger audience [30]. To use descriptive statistics effectively in academic research, researchers should carefully consider the limitations and scope of the methodology, use additional statistical techniques as needed, ensure that their data is representative of the population of interest, and interpret their descriptive statistics in a thoughtful and nuanced manner.

## **Conflict of interest**

The authors declare no conflict of interest.

## **Author details**

Olubunmi Alabi<sup>1\*</sup> and Tosin Bukola<sup>2</sup>


1 African University of Science and Technology, Abuja, Nigeria

2 University of Greenwich, London, United Kingdom

\*Address all correspondence to: oalabi@aust.edu.ng

## **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Agresti A, Franklin C. *Statistics: The Art and Science of Learning from Data*. Upper Saddle River, NJ: Pearson; 2009
- [2] Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 4th ed. Shelton (CT): PMPH-USA; 2014
- [3] Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: Routledge; 2013
- [4] Osborne J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment*. 2019;**10**(7):1-9
- [5] Field A, Hole G. *How To Design and Report Experiments* Sage. The Tyranny of Evaluation Human Factors in Computing Systems CHI Fringe; 2003
- [6] Anders H. *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley; 1998. p. xvii+795. ISBN 0-471-17912-4
- [7] Rebecca M. Warner's *Applied Statistics: From Bivariate Through Multivariate Techniques*. Second Edition. Thousand Oaks, California: SAGE Publications; 2012
- [8] Sullivan LM, Artino AR Jr. Analyzing and interpreting continuous data using ordinal regression. *Journal of Graduate Medical Education*. 2013;**5**(4):542-543
- [9] Hoaglin DC, Mosteller F, Tukey JW. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons; 2011
- [10] Maxwell SE, Delaney HD, Kelley K. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge; 2017
- [11] De Leeuw ED, Hox JJ. *International Handbook of Survey Methodology*. Routledge; 2008
- [12] Chatfield C. *The Analysis of Time Series: An Introduction*. CRC Press; 2016
- [13] Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. Pearson; 2013
- [14] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2016
- [15] Field A, Miles J, Field Z. *Discovering Statistics Using R*. Sage; 2012
- [16] Howell DC. *Statistical Methods for Psychology*. Cengage Learning; 2013
- [17] Wilcox RR. *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. CRC Press; 2017
- [18] Hair JF, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis*. Pearson; 2019
- [19] Beasley TM, Schumacker RE. Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *Journal of Experimental Education*. 2013;**81**(3):310-312
- [20] Dodge Y. *The Concise Encyclopedia of Statistics*. Springer Science & Business Media; 2008
- [21] Krzywinski M, Altman N. Points of significance: Visualizing samples with box plots. *Nature Methods*. 2014;**11**(2):119-120
- [22] Cleveland WS. *Visualizing data*. Hobart Press; 1993

- [23] Little RJ, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons; 2019
- [24] Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*. 2008;**52**(3):1694-1711
- [25] Shmueli G, Bruce PC, Yahav I, Patel NR, Lichtendahl KC Jr, Desarbo WS. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons; 2017
- [26] Aguinis H, Gottfredson RK. Statistical power analysis in HRM research. *Organizational Research Methods*. 2013;**16**(2):289-324
- [27] Stevens JP. *Applied Multivariate Statistics for the Social Sciences*. Routledge; 2012
- [28] Byrne BM. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. Routledge; 2016
- [29] Everitt BS, Hothorn T. *An Introduction to Applied Multivariate Analysis with R*. Springer; 2011
- [30] Kosslyn SM. *Graph Design for the Eye and Mind*. Oxford University Press; 2006



## Chapter 2

# Descriptive Statistics

*Hazhar Talaat Abubaker Blbas*

### Abstract

Descriptive statistics is a branch of statistics that deals with summarizing and describing the main features of a dataset. This chapter will cover the value of statistics, how data analysis occurs in scientific study, the distinction between a sample and the population, the different types of variables, sampling techniques, measures of central tendency, and measures of dispersion. The summary of descriptive statistics gives a succinct overview of various metrics and visual representations, enabling researchers and analysts to learn more about the features of the dataset and draw accurate conclusions.

**Keywords:** mean, median, mode, standard deviation, coefficient of variation, probability sampling, non-probability sampling

### 1. Introduction

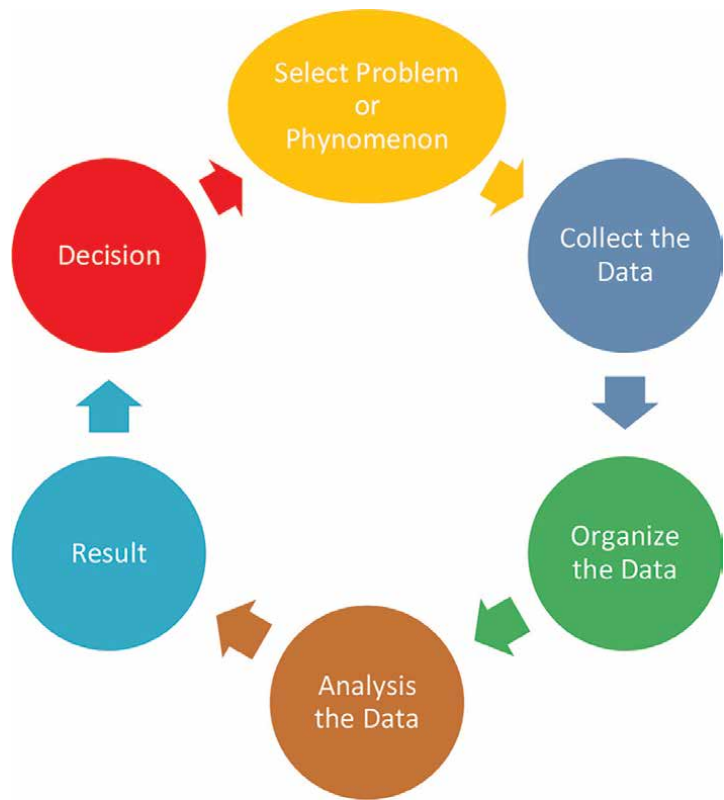
Descriptive statistics involve summarizing and describing data using numerical measures and graphical representations. It provides a concise and meaningful way to understand and communicate the main characteristics of a dataset. This introduction explores the basics of descriptive statistics, including measures of central tendency, measures of dispersion, and graphical representations. By examining these statistical tools, we can gain insights into the patterns, variability, and distribution of data, allowing us to make informed interpretations and draw meaningful conclusions.

### 2. What is the process of analyzing data in statistics?

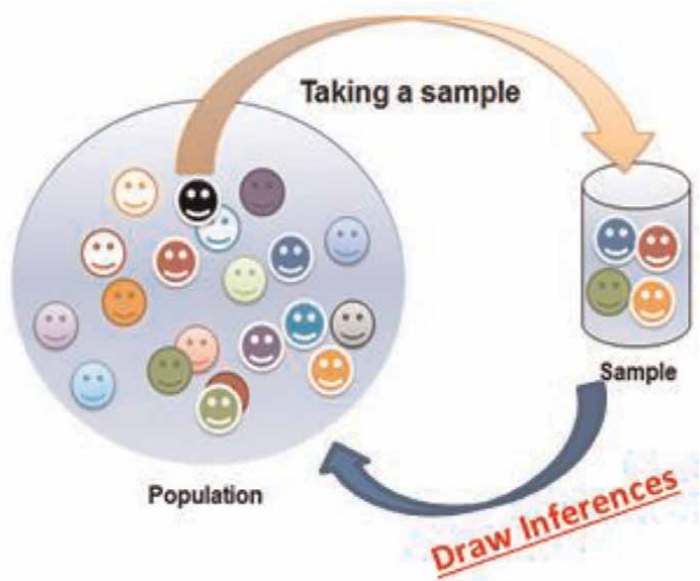
Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions as shown in **Figure 1**.

### 3. Sample and population

A population is the collection of all outcomes, responses, measurements, or counts that are of interest since sample is a subset of a population as shown in **Figure 2** [1–3].



**Figure 1.**  
*Process of data analysis in scientific research. Source: Author has been created as a new work.*

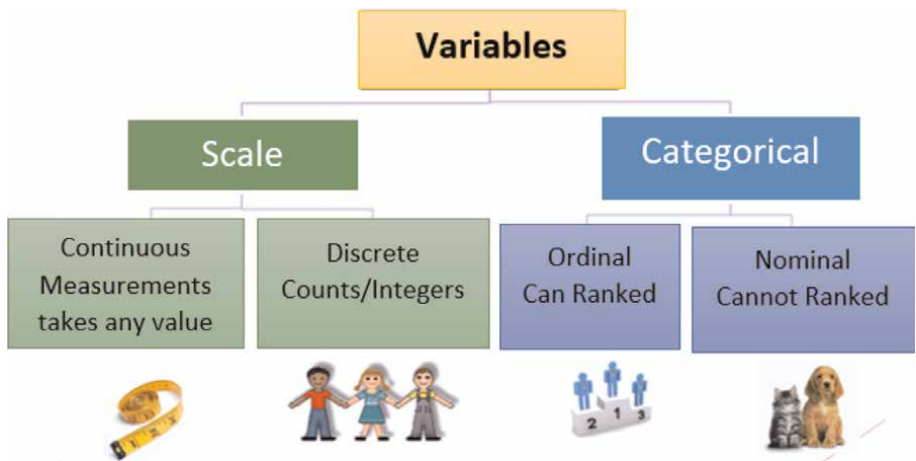


**Figure 2.**  
*Difference between sample and population. Source: Author has been created as a new work.*

#### 4. Type of variables

Variable is a characteristic that can assume different values and alphabetic. There are two common types of variables, such as quantitative variables and qualitative variables, as shown in **Figure 3** [4, 5].

1. Quantitative variables (numerical variables) are variables that represent measurable quantities or amounts. They can be further classified into two types:
  - i. Discrete variables: Discrete variables are numerical variables that can only take on specific, separate values. These values are typically whole numbers or counts and cannot be subdivided further. Examples of discrete variables include the number of children in a family, the number of customers in a store, or the number of items sold.
  - ii. Continuous variables: Continuous variables are numerical variables that can take on any value within a certain range. They can be measured with a high degree of precision and can have infinite possible values between any two points. Examples of continuous variables include height, weight, temperature, and income.
2. Qualitative variables (categorical variable): This type of variable represents data that can be divided into distinct categories or groups. Examples include gender, ethnicity, marital status, and level of education.
  - i. Nominal variables: Nominal variables are categorical variables that represent data with no ranking. Examples of nominal variables include gender (male/female), ethnicity (Asian, African, European, etc.), marital status (single, married, divorced), and eye color (blue, brown, green).



**Figure 3.**  
*Type of variables. Source: Author has been created as a new work.*

- ii. Ordinal variables: Ordinal variables represent data that has a natural order or ranking, but the differences between the categories may not be consistent or measurable. The categories can be ranked or ordered based on some criterion, but the magnitude of the difference between categories is not known. Examples of ordinal variables include satisfaction levels (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied), educational attainment (high school diploma, bachelor's degree, master's degree), and survey responses using Likert scales ("strongly agree," "agree," "neutral," "disagree," "strongly disagree").

## **5. Sampling plan**

Once the target population has been identified, next the sampling plan must be devised. Goal: Randomly select a small percent of the population that will in turn represent the ideas of the population as a whole. There are two general types of sampling techniques [1, 2, 5]:

### **5.1 Probability (random) sampling**

All members of the population must be specified prior to drawing the sample and each member of the population has equal probability of being chosen or included in the sample. There are four common types of Probability (Random) Sampling:

#### *5.1.1 Simple random sampling*

Simple random sampling is a statistical sampling technique in which each member of a population has an equal probability of being selected to be part of the sample. The selection process is conducted randomly, without any bias or preference toward certain individuals or elements in the population.

A researcher wants to conduct a survey to understand the opinions of students at a university regarding a new policy. The university has a total population of 1500 students. For example, a researcher wants to select 100 out of 1500 students as a sample. Put a unique identifier to each of the students such as a student ID number. Then, randomly select the 100 students as a sample like a lottery game.

#### *5.1.2 Systematic sampling*

Systematic sampling is a statistical sampling technique that involves selecting every  $k^{\text{th}}$  element from a population, where  $k$  is a predetermined interval. It is similar to simple random sampling but incorporates a systematic approach to the selection process.

Depending on the previous example of simple random sampling, the researcher wants to select 100 students using systematic sampling. We will calculate the sampling interval, which divides the population size by the desired sample size to determine the sampling interval. In this case, the sampling interval would be  $1000 / 100 = 10$ . Next, select a random starting point within the first  $k$  elements (in this case, the first 10 students). Next, starting from the random starting point, select every 10th student thereafter. So, you would select the 10th, 20th, 30th, and so on, until you reach the desired sample size.

### *5.1.3 Stratified sampling*

Stratified sampling is a statistical sampling technique that involves dividing the population into two or more than two homogeneous groups. Then, randomly select the desired case in each group using simple random sampling.

Depending on the previous example in simple random sampling, the researcher wants to select 100 students using stratified sampling.

First, students can be stratified based on their academic disciplines into four strata: statistics, accounting, business, and economics department.

1. Determine the sample size: Decide on the desired sample size for each stratum. Let us say you want to sample 25 students from each stratum (department), resulting in a total sample size of 100 students.
2. Divide the population into four strata: Categorize the students into the respective strata based on their academic disciplines. Each student should belong to only one stratum.
3. Determine the allocation: Calculate the proportionate allocation for each stratum by dividing the desired sample size for that stratum by the total sample size. In this case, since each stratum has the same desired sample size (25 students), the allocation would be  $1/4$  (25%) for each stratum.
4. Sample within each stratum: Perform simple random sampling within each stratum separately. Randomly select 25% (25 students) from the statistics stratum, 25% from the accounting stratum, 25% from business stratum, and 25% from the economics stratum.
5. Collect data: Once the samples are selected, collect the relevant data or information from the students in each stratum.

### *5.1.4 Cluster sampling*

Cluster sampling: Cluster sampling involves dividing the population into clusters or groups, often based on geographical proximity, and randomly selecting entire more than one clusters as the sampling units. This technique is useful when it is impractical or costly to sample individuals individually, and it can provide cost and time efficiencies.

## **5.2 Nonprobability sampling**

Every element in the population does not have an equal probability of being chosen. The process of inclusion in the sample is based on the judgment of the person selecting the sample. There are four common types of nonprobability sampling.

### *5.2.1 Judgment sampling*

Purposive sampling: Purposive sampling, also known as judgmental or selective sampling, involves handpicking individuals based on specific criteria or the researcher's judgment. This technique is often used in qualitative research or when a

specific subgroup of the population is of particular interest. Purposive sampling allows the researcher to target individuals who possess the desired characteristics or have relevant experiences.

### 5.2.2 Convenience sampling

Convenience sampling: Convenience sampling involves selecting individuals who are easily accessible or readily available to the researcher. This method is convenient and often used in situations where time, cost, or accessibility is a constraint. However, convenience sampling can introduce bias, as the sample may not be representative of the entire population.

### 5.2.3 Quota sampling

Quota sampling: Quota sampling involves setting specific quotas or targets for certain characteristics or subgroups within the population. The researcher selects individuals to fulfill the predetermined quotas until they are satisfied with the sample composition. Quota sampling allows for control over sample proportions but does not involve random selection.

### 5.2.4 Snowball sampling

Snowball sampling: Snowball sampling is a technique where initial participants are selected, and then they help identify and recruit additional participants from their social networks. This method is useful when studying hard-to-reach or hidden populations. Snowball sampling relies on referrals and networks to expand the sample size.

## 6. Measures of central tendency

It is a statistical measure that represents information about the central or middle value of a dataset. The three common measures of central tendency are the mean, median, and mode [4–6].

1. Mean (average), is calculated by summing up all the values in a dataset and dividing by the number of values. It represents the balancing point of the dataset and is sensitive to outliers. Depending on 894 people from Kurdistan Region of Iraq, the average age of people for the survey about depression and anxiety during the outbreak of COVID-19 is 33 years [1].

$$\bar{X} = \frac{\sum X_i}{n} \quad (1)$$

Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 95. The mean is calculated as  $(85 + 90 + 92 + 88 + 95) / 5 = 90$ .

2. Median: The median is the middle value in a dataset when it is arranged in ascending or descending order. If there is an even number of values, the median

is the average of the two middle values. The median is less influenced by outliers compared to the mean.

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. When arranged in ascending order, the middle value is 90. Therefore, the median is 90.

3. Mode: The mode represents the most frequently occurring value(s) in a dataset. It is the value that appears with the highest frequency. A dataset can have no mode (when all values occur equally) or multiple modes (when multiple values have the same highest frequency).

Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 90. The mode is 90 because it appears twice, which is more frequently than any other value.

## 7. Measures of dispersion (variation)

Measures of dispersion (Variation), provide information about the spread or dispersion of data points around the central tendency. The first three main measures of dispersion including range, standard deviation, and variance, are used when we have the same unit of datasets but we can use coefficient of variation once we have different units of datasets [4–8].

1. Range (R): It is the difference between the maximum and minimum values in a dataset.

$$R = \text{Highest value} - \text{Lowest value} \quad (2)$$

Example: Consider the following dataset of exam scores: 85, 90, 92, 88, 95. The range is calculated as  $95 - 85 = 10$ .

2. Variance ( $S^2$ ): It measures the average squared deviation of each data point from the mean. It provides a more precise measure of dispersion by considering the differences between individual data points and the mean. However, it is in squared units and is sensitive to outliers.

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1} \quad (3)$$

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. The variance is calculated as follows:

- Calculate the mean:  $(85 + 90 + 92 + 88 + 95) / 5 = 90$ .
- Calculate the squared deviation for each data point from the mean:  $(85-90)^2$ ,  $(90-90)^2$ ,  $(92-90)^2$ ,  $(88-90)^2$ ,  $(95-90)^2$ .
- Calculate the average of these squared deviations:  $(25 + 0 + 4 + 4 + 25) / 5 = 12.8$ . Therefore, the variance is 12.8.

1. Standard Deviation (S): It is the square root of the variance. It is the most commonly used measure of dispersion as it is in the original units of the data, making it more interpretable. It provides a measure of how much the data deviates from the mean.

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} \quad (4)$$

Example: Using the same dataset of exam scores: 85, 90, 92, 88, 95. The standard deviation is the square root of the variance calculated in the previous example, which is approximately 3.58.

2. The coefficient of variation (CV) is a relative measure of dispersion that expresses the standard deviation as a percentage of the mean. It is used to compare the variability of datasets with different means or scales. The formula for calculating the coefficient of variation is:

$$CV = \frac{S}{\bar{X}} * 100 \quad (5)$$

Here's an example to illustrate the calculation of the coefficient of variation. Consider two datasets representing the monthly sales of two stores:

Store A: Mean = \$10,000, Standard Deviation = \$2000.

Store B: Mean = \$15,000, Standard Deviation = \$3000

- CV for the store A =  $(2000 / 10,000) * 100 = 20\%$
- CV for the store B =  $(3000 / 15,000) * 100 = 20\%$

In this example, both stores have the same coefficient of variation of 20%. It indicates that the relative variability or dispersion of sales is the same for both stores, even though Store B has a higher mean and standard deviation compared to Store A.

A lower coefficient of variation indicates less variability relative to the mean, while a higher coefficient of variation suggests greater relative variability.

## Additional information

ORCID account: <https://orcid.org/my-orcid?orcid=0000-0002-5760-3019>

Google Scholar Citation: <https://scholar.google.com/citations?user=zl0eeJoAAAAJ&hl=en&authuser=1>



## **Author details**


Hazhar Talaat Abubaker Blbas

Department of Statistics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan Region, Iraq

\*Address all correspondence to: [hazhar.abubaker@su.edu.krd](mailto:hazhar.abubaker@su.edu.krd)

## **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## **References**

- [1] Aroian K, Uddin N, Blbas H. Longitudinal study of stress, social support, and depression in married Arab immigrant women. *Health care for women international*. Feb 1 2017;**38**(2): 100-117
- [2] Rosner B. *Fundamentals of biostatistics*. Cengage Learning. 2015
- [3] Bluman A. *Elementary Statistics: A Step by Step Approach* 9e. McGraw Hill; 2014
- [4] Blbas H. Statistical analysis for the most influential reasons for divorce between men and women in Erbil-Iraq. *International Journal*. Malmö, Sweden. 2019
- [5] Triola MF, Iossi L. *Essentials of Statistics*. Boston, MA, USA: Pearson Addison Wesley; 2008
- [6] Hanif M, Ahmed M, Ahmed AM. *Biostatistics for health students with manual on software applications*. Islamic Society of Statistical Sciences. 2006
- [7] Rowe P. *Essential Statistics for the Pharmaceutical Sciences*. John Wiley & Sons; 2015
- [8] Blbas HT, Aziz KF, Nejad SH, Barzinjy AA. Phenomenon of depression and anxiety related to precautions for prevention among population during the outbreak of COVID-19 in Kurdistan region of Iraq: Based on questionnaire survey. *Journal of Public Health*. 2020; **10**:1-5

# Spatial Statistics: A GIS Methodology to Investigate Point Patterns in Stroke Patient Healthcare

*Joanne N. Halls, Barbara J. Lutz, Sara B. Jones  
and Matthew A. Psioda*

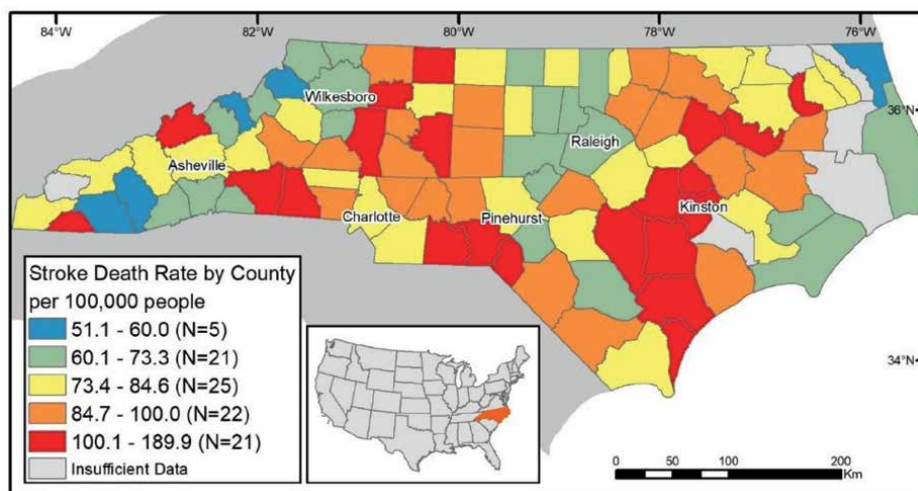
## Abstract

Stroke is the leading cause of major disability and the fifth leading cause of death in the United States. Stroke incidence across the U.S. is not uniform where the southeastern states, known as the “Stroke Belt”, have historically higher rates. Importantly, while the national average death rate due to stroke has been declining, the death rate in the Stroke Belt (from 2013 to 2015) increased 4.2% overall and 5.8% within the Hispanic population. Healthcare interventions have been designed to improve acute stroke care, but they are less prevalent in addressing post-acute care needs of stroke survivors. Therefore, this chapter will describe the results of a recent study that investigated patterns in post-stroke care using a sequence of geospatial statistics. Through this investigation, the reader will learn the sequence of Geographic Information System (GIS) techniques appropriate to use when studying complex spatial patterns.

**Keywords:** geospatial statistics, point patterns, drive time, GIS, healthcare data, stroke, North Carolina USA

## 1. Introduction

Stroke is the leading cause of major disability and the fifth leading cause of death in the U.S. [1]. Stroke incidence across the U.S. is not uniform. The southern states of Arkansas, Louisiana, Mississippi, Alabama, Tennessee, Georgia, South Carolina, and North Carolina are known as the “Stroke Belt” where there are historically higher rates [2]. Importantly, while the national average death rate due to stroke has been declining, from 2013 to 2015 the death rate in the Stroke Belt has increased 4.2% overall and 5.8% within the Hispanic population [3]. In North Carolina the average death rate from stroke is 84.6 deaths per 100,000 (age 35 and up, all races/ethnicities, both genders, 2014–2016) while the national average is 73.3 per 100,000. However, this average death rate does not tell the full story of stroke in North Carolina because the rate of death due to strokes varies substantially where 5 rural counties experience the lowest rates between 51 and 60 per 100,000 people compared with 21 counties



**Figure 1.**

*Stroke death rate, per 100,000 people aged 35+, all races/ethnicities, all genders, 2014–2016. Data source: Centers for Disease Control, Interactive Atlas of Heart Disease and Stroke [4].*

with 100 to 190 per 100,000 people (**Figure 1**). Specifically, 68 out of the 100 North Carolina counties have stroke death rates above the national average [5]. This high rate of stroke and high variability across the state has led to researching the differences between these locations to identify if there are reasons for such variability in stroke death rates.

Interventions, such as timely administration of intravenous tissue plasminogen activator (IV tPA) or mechanical thrombectomy, have improved stroke patient outcomes in acute care [6]. However, evidence-based interventions to optimize post-acute stroke recovery and address recurrent stroke after discharge have not been widely implemented. Several Transitional Care (TC) models have been designed to reduce care fragmentation and improve post-discharge outcomes [7]. The Comprehensive Post-Acute Stroke Services (COMPASS) study evaluated an evidence-based TC model, which included early telephone follow-up and an in-person clinic visit, compared with usual care in a cluster-randomized pragmatic trial in 40 (out of 110) hospitals in North Carolina [8, 9]. Average attendance at follow-up clinics in the intervention group was 35% and ranged from 6 to 70% across the 19 hospitals that implemented the intervention [8]. This variability in attendance at follow-up care led to this geographic study that investigated where attendance was highest and lowest and the relationship with drive time from patients' homes to the follow-up clinics.

## 2. Data and methods

We spatially compared where patients lived, drivetime to their assigned follow-up clinic, the Area Deprivation Index from the Health Innovation Program [10] and the designation of urban versus rural from the United States Department of Agriculture. It was hypothesized that (1) the further a patient lived from the clinic the less likely they would attend the follow-up visit; (2) higher area deprivation would correlate to

lower attendance at the clinic; and (3) urban clinics would have a higher attendance rate than rural clinics. Results from this geospatial statistical analysis could yield insights into the variability of stroke death rates across North Carolina and lead to improvements in patient access to follow-up care.

## **2.1 Geocoding address data**

Studies have shown that implementing a Geographic Information System (GIS) can provide spatial data analytics for identifying areas with less access to care and therefore potential health care disparities [11]. Additionally, by spatially referencing a variety of data layers, functions such as map overlay to identify areas that intersect between spatial layers can be utilized which can lead to statistical comparison among the varied spatial layers. For example, overlaying drinking water sources (public system versus well water), agricultural use of pesticides, and prevalence of Parkinson's Disease (PD), has identified a link between well water and PD in California [12]. To begin the GIS process, non-spatial data need to be converted into a spatial data set. Several data sets, consisting of hospital, clinic, and patient data, were geo-referenced using the World Geocoding Service within ArcGIS 10.7.1. These data included 19 hospitals that participated in the intervention study, affiliated hospital clinics where patients were assigned for follow-up care, and patient data collected between July 2016 and March 2018.

Geocoding was performed on the physical addresses of hospitals and clinics as well as residential addresses of study participants. Geocoding is an iterative process that compares an address with a reference base map to estimate a spatial location for the address. To perform the geocoding, the first step is to import the address data into a geodatabase table and then parse the addresses into several components (e.g. address number, direction prefix, street name, direction suffix, city name, and zip code) using the Address Locator tool which is an important step to compare these components with the reference base map [13]. Next, search criteria are defined and then the data are batch processed to compare them with the base map. This results in a match score for each address record. The search criteria define how spatially accurate the resulting locations will be. For example, one can specify to match a record if the city and zip code match or, for greater precision, if the address number, direction and street name also match. These criteria can greatly impact the resulting spatial accuracy of the output point data and therefore care must be taken when applying the search criteria. In this example, we specified the most stringent criteria in order to have the highest spatial accuracy of the output point data.

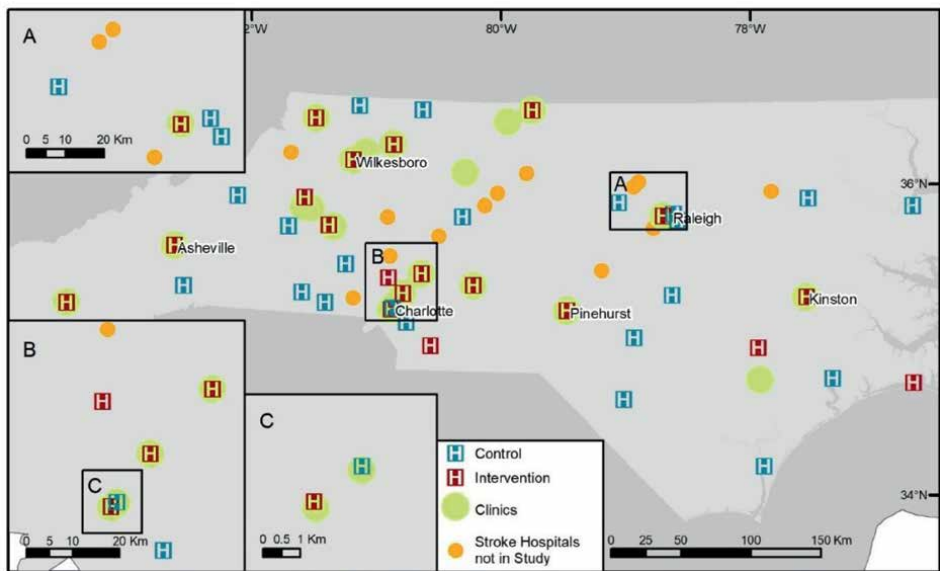
All records that can be matched using the criteria and the reference base map are given a match score that reflects the quality of the output data. It is important to check all output points that have a match score less than 100% because these points could have multiple matches to the reference map or other issues, such as a location at a large complex or a typographical error in the format of the address data. An example of a typographical error is the slight misspelling of a street name. When this occurs, the user must check all results with a score that is less than 100%, correct any errors with the address, and then re-run the geocoding to obtain a final result. If the address data have a substantial number of inconsistencies, it is best to reformat all of the data and then perform the geocoding process rather than iteratively fix each error and this will improve the spatial accuracy of the resulting point data [14].

In this study, hospitals, clinics and patient records were geocoded, iteratively checked, and final point data were derived. Of the 2689 patient records only 5 (0.2%)

were not able to be geocoded because the patients were listed as homeless or had an invalid address. Many addresses were initially unable to be geocoded due to errors in the address data. This is very common with address data because of the many opportunities to enter incorrect data. Therefore, as discussed above, it is critical to employ extensive quality control and assurance procedures during data collection and to correct address data prior to initiating the geocoding process. Because some patients resided outside the study area, i.e., beyond 50 miles of the North Carolina border, these records were removed from further analysis, which resulted in 2615 (or 97.2%) of the geocoded sample residing within the study area. Due to patient confidentiality, we cannot illustrate the point results from the geocoding of the patient data; hospital and clinic locations are shown in **Figure 2**, which demonstrates the conversion from address data to point locations.

## 2.2 Computing shortest path and drive time

Drive time is commonly used as a metric of geographic accessibility of health care services [16, 17]. Due to design of the study intervention, the patients were assigned to the follow-up clinic that was affiliated with the hospital where they received their acute stroke care, which may not have been in close proximity to where they lived due to wide catchment areas of acute stroke hospitals. Since where patients live was not considered, in some cases the patients were not assigned to the closest clinic. Patient addresses, therefore, were linked with their assigned clinic. This is an important aspect to this study because in most geographic analyses the closest, or shortest distance, is used as a measure of nearness, but in this case, the closest clinic may not have been the clinic the patient was assigned to. This is an excellent example ensuring the



**Figure 2.** Hospitals and clinics in the North Carolina, USA, study area. Hospitals were designated as control and intervention. Patients seen at the Intervention ( $N = 19$ ) hospitals were given custom care plans and assigned to clinics for follow-up care. Inset A is the Raleigh area and insets B and C are the Charlotte area. Also shown (in orange) are Joint Commission Certified primary or comprehensive stroke hospitals that did not participate in the study [15].

structure of the data and the study context guide appropriate use of spatial statistics and the interpretation of the resulting nearness or other spatial statistics results.

Open StreetMap (<https://www.openstreetmap.org/#map=4/38.01/-95.84>) and ArcGIS Network Analyst (<https://www.esri.com/en-us/arcgis/products/arcgis-network-analyst/overview>) were used to compute the shortest drive time from each patient's residence to their assigned follow-up clinic. To perform the drive time calculation recall that we considered all patients within North Carolina as well as patients that were within 50 miles of the border. Therefore, it was necessary to download the Open StreetMap data for North Carolina, South Carolina, Tennessee, and Virginia, import these into ArcGIS, and then build a comprehensive network dataset for all these state road networks. Once the network was built, the patient and clinic data were used as "origin" and "destination" data in performing the shortest drive time calculation. Additionally, the shortest drive time to the nearest clinic was also calculated to identify the number of patients who were not assigned to the closest clinic and to identify if the spatial patterns differ when comparing closest versus assigned clinic. Using the average drive time, zones around each follow-up clinic were computed using Network Analysis to derive service areas, which can then be used to identify underserved areas.

Lastly, Analysis of Variance (or ANOVA) was used to test whether drive time and visitation rates differed between urban and rural areas and cross tabulation and Pearson's Chi Square were used to compare the rate of attendance at the clinic and drive time to assigned versus closest clinic.

### **2.3 Spatial statistical analyses**

There are a variety of GIS methods to identify spatial clusters and relationships among several data layers. Importantly, the methodology should follow a series of steps to establish the neighborhood distances between observations that are valid for each unique study area [18–25]. Therefore, we computed both global and local spatial statistics to systematically assess the degree of spatial clustering. First, the Moran's I spatial autocorrelation statistic was tested using the patient location and shortest drive time to determine if they were spatially clustered. Next, we performed an Average Nearest Neighbor (ANN) calculation, which measures the degree of clustering using distances between neighboring locations. The ANN statistic is useful for confirming the spatial autocorrelation results from Moran's I and gives a measure (distance) of the amount of clustering. Importantly, when calculating the ANN statistic, you must include the size of the study area (in this case it was 127,605,669,275 m<sup>2</sup>) otherwise the statistic will use the size of the bounding box of the dataset and will likely overestimate the size of the study area which can dramatically alter the results. Once global clustering has been measured using the Moran's I and ANN, we then computed a local spatial cluster analysis (local Moran's I) to identify where the clusters are located. This technique identified several types of clusters: (1) the location of clusters with low values (shorter drive time), (2) the location of clusters with high values (longer drive time), and (3) cluster outliers where clusters of low values are surrounded by high values and vice versa where clusters of high values are surrounded by low values [26].

Regression analysis is used to look for relationships among independent, or explanatory, variables and a dependent variable. With spatial data we can use Ordinary Least Squares (OLS) Regression to identify an overall pattern in the data and Geographically Weighted Regression (GWR) to identify regression equations at

the local level, or each spatial unit such as a county, Census Tract, or other enumeration unit. In this study we used demographic data, the Area Deprivation Index, and access to community resources as independent variables and attendance rate at the follow-up clinic as the dependent variable. In North Carolina there are 100 counties, 1410 Census Tracts, and 6155 Block Groups with an average size of 22.6 sq. km. In the regression analysis, we used Census Block Groups which enables the highest granularity of resolution.

OLS is a multiple regression technique that identifies the strength of the relationships (both positively and negatively) between the independent variables with the dependent variable (rate of attendance at the follow-up clinic visit). First, all independent variables are tested and if any are colinear then they are iteratively removed from the analysis until a result can be achieved that has no multicollinearity problems. From this shorter list of independent variables, the GWR technique is used to identify local weights (importance) for each independent variable and to derive unique regression equations for each location. The GWR technique uses a local/neighborhood approach to computing multiple regression. Unlike OLS regression which looks at the entire dataset as a whole, the GWR method uses a neighborhood around each location (e.g. Block Group) to compute a regression equation. Therefore, all independent variables were tested and those that did not violate the rules of regression were included in the development of GWR models. The benefit of GWR is the ability to identify the importance of the independent variables across the study area. For example, in some areas drive time may be more important compared to other areas where demographics (e.g. age or race/ethnicity) may be more related to the attendance at follow-up clinics. The GWR technique enables the identification of spatially significant differences across the study area, rather than traditional multiple regression that looks at the entire dataset as a whole. GWR has been used in many disciplines including health studies to investigate the spatial patterns of diseases [27–29]. It is best to run many GWR trials to identify the highest performance based on the combination of independent variables. One of the most important decisions in the use of GWR is the bandwidth, or the calculation of the number of neighbors around each observation. This decision is important because it will determine the number of observations used in each unique local regression equation. The bandwidth can be a fixed distance or, so that all observations use the same neighborhood size, it can vary across the study area depending on the geographic distribution of observations. The corrected Akaike Information Criterion (AICc) identifies the optimal distance and the Cross Validation (CV) identifies the optimal number of neighbors. One can also use the distance identified from the ANN results. Because the bandwidth distance is so important in the derivation of local regression equations, we recommend testing all three approaches and using the one that yields the best results.

In this study, 1523 Block Groups with patients were included since the dependent variable was attendance at the follow-up clinic. Unlike other studies, such as studies based on Census data where there is complete geographic coverage, in this study there were gaps in coverage and the distance between neighboring polygons varied because of the large study area. Therefore, the size of the neighborhood was defined using the cross-validation approach where the optimal number of neighbors was defined locally. A well-specified model will have randomly distributed over and under predictions (residuals) of the dependent variable (attendance at follow-up clinic). If there is clustering in the over and under predictions (residuals), then it is likely that the model is missing at least one key variable.

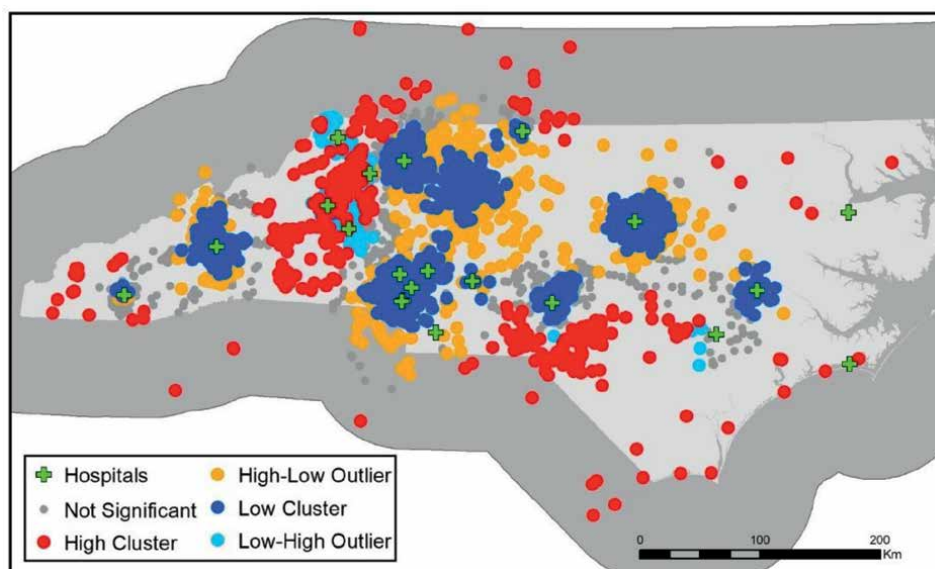


Next, Grouping Analysis, which is similar to Principal Component Analysis, is a method that is used to look for an overall spatial pattern, especially when there are many variables, with the goal of identifying statistically significant clusters in space. The Grouping Analysis technique uses the K-means statistic to identify the independent variables within each group that are as similar as possible while also identifying groups that are as different as possible. The most important variables identified through the regression analysis were used in the Grouping Analysis to identify the statistically significant groups within the study area.

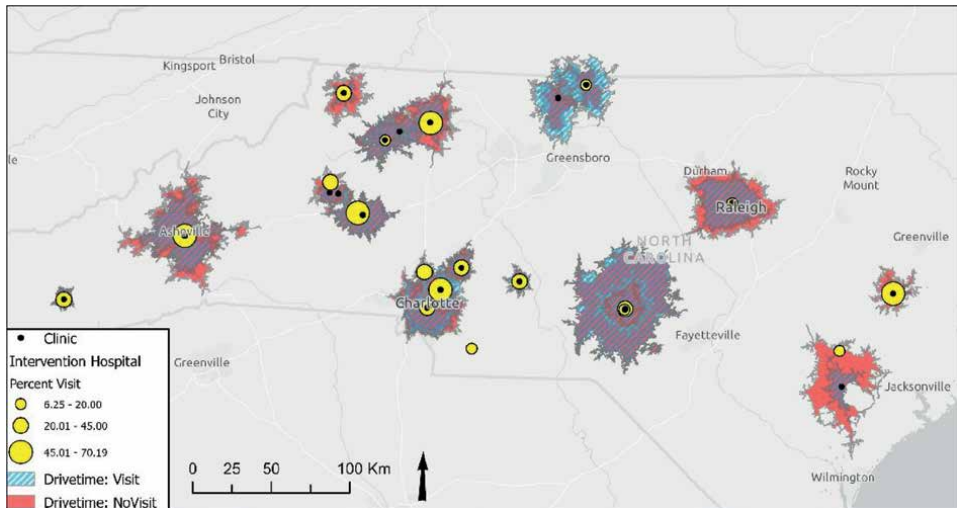
### 3. Results

#### 3.1 Drive time analysis

The average drive time from patients' residences to assigned clinics was 24 ( $\pm$  25 minutes Standard Dev) with a minimum of 0.75 to a maximum of 360 minutes. A majority, 94%, had a drive time less than 1 hr. and 74% were less than 30 minutes. The global Moran's I spatial autocorrelation statistic, which compared the patient location and drive time, identified that the data were significantly clustered with a z-score of 4.289 (P-value = 0.0000). Given the Moran I results, the next spatial statistic test, the Average Nearest Neighbor (ANN), also had highly significant clustering with a z-score of  $-35.742$  (P-value = 0.000). Given these results, we then computed a spatial cluster analysis, the local Moran's I, to identify where the clusters were located (**Figure 3**). These results identify where there are high clusters (longer drive time), low clusters (shorter drive time), and the outliers of high/low clusters (higher drive time surrounded by lower drive time) which were more prevalent than low/high clusters.



**Figure 3.**  
 Drive-time spatial cluster analysis (Local Moran's I) showing locations of high clusters (longer drive time), low clusters (shorter drive time), and outliers where low drive time clusters are surrounded by high drive time and high drive time clusters are surrounded by low drive time.



**Figure 4.** Average drive time for patients who attended the follow-up clinic (blue hatch) versus those who did not attend the clinic (red) as well as the rate of attendance (yellow graduated circles). Most locations had a higher attendance rate with shorter drive times ( $P = 0.005$ ).

The average rate of attendance at the follow-up clinics was 35% (range was 6 to 70%) and the average drive time for those who attended the clinic was significantly less at 19 minutes versus those who did not attend at 23 minutes ( $P = 0.005$ ). Given this significant relationship, we can compare the rate of attendance at the follow-up clinic with drive time by deriving drive time zones around each clinic using the average drive time for each clinic (**Figure 4**). We can see that some clinics have a much longer average drive time for those who did not attend the follow-up clinic shown in red on the map. Additionally, the larger average drive time (large zones on the map) tend to also have the lowest visit rate (smaller yellow marker size). Conversely, the smaller drive time zones had larger attendance rates. These relationships are not ubiquitous, but they are significant. What this type of mapping reveals is the importance of investigating the spatial patterns in data rather than solely relying on global statistics.

The drive time portion of the study concluded that there are locations of significant clusters of shorter and longer drive times and that shorter drive time was significantly related to higher rates of attendance.

### 3.2 Regression analysis

There were 18 independent variables tested to determine which are associated with attendance rate at the follow-up clinic (the dependent variable). **Table 1** contains the list of independent variables, sorted by decreasing importance, as indicated by the overall significance, the direction of the association (positive or negative), the Variance Inflation Factor (VIF) which indicates multicollinearity, and the list of covariates, or variables that are colinear and are potentially providing the same information. Since the goal of regression analysis is to include variables that explain a unique aspect of the dependent variable it is wise to remove redundant variables. One way of deciding which covariates to select is to use the variable with the strongest

Variable	Significance (%)	Negative (%)	Positive (%)	VIF	Covariates *
Average drive-time	100	100	0	1.25	
Rate of caregivers	100	0	100	1.88	
Not hispanic	93.59	0	100	19.19	White, Urban Area, Urban Cluster, Rural, Black
ADI	85.59	1.56	98.44	1.26	
Percent rural	77.77	0	100	296.73	Urban Area, Urban Cluster, Not Hispanic, White, Black
Average age	75.15	94.39	5.61	1.11	
Percent white	55.7	22.49	77.51	355.46	Not Hispanic, Black, Urban Area, Urban Cluster, Rural
Percent black	48.87	5.05	94.95	101.36	White, Not Hispanic, Urban Area, Urban Cluster, Rural
Percent urban area	36.62	30.4	69.6	194.63	Rural, Urban Cluster, Not Hispanic, White, Black
Density of community resources	35.31	22.3	77.7	1.91	
Percent unknown race	34.95	92.34	7.66	3.59	
Percent within urban cluster	28.53	52.31	47.69	196.62	Rural, Urban Area, Not Hispanic, White, Black
Percent Hispanic	26.09	13.67	86.33	1.84	
Percent multi-race	24.09	0	100	2.84	
Percent Asian	11.66	80.28	19.72	1.54	
Percent native American	6.2	43.58	56.42	4.43	
Percent other race	4.7	39.13	60.87	6.31	
Percent Pacific Islander	0.14	8.93	91.07	1.29	

*Negative and Positive indicate the relationship between the independent and dependent variable (attendance at the follow-up clinic). Variance Inflation Factor (VIF) indicates multicollinearity where the higher the value the greater collinearity. Covariates are listed in decreasing importance/strength.*

**Table 1.**

*List of Ordinary Least Squares regression results where the independent variables are listed in order of decreasing significance.*

positive or negative relationship as well as remove, one at a time, the variables with VIF greater than 7.5 and then re-run the OLS to check the results. In this iterative way the most important explanatory variables are identified.

The strongest independent variables were average drive time, number of caregivers, not-Hispanic, ADI, rural, and average age. As expected, drive time was very

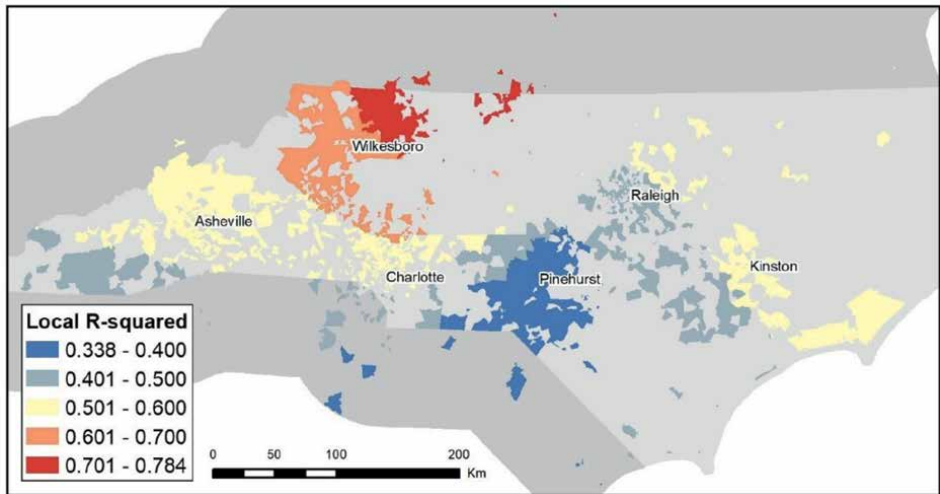
strongly negatively related to attendance and having caregivers was very positively related to attendance (both 100%). The strongest race/ethnicity variable was the percentage of the patients who were not Hispanic which was positively related to clinic attendance and many of the race/ethnicity variables covaried. Interestingly, ADI was positively related to attendance which was unexpected because higher ADI means the area is more deprived. The next strongest variable was percent rural, and it was also positively related to the attendance which indicates that patients who lived in rural areas were more likely to attend the follow-up clinic. As expected, average age was negatively related to attendance which means older patients would be less likely to attend the follow-up. The remaining independent variables were substantially less related to attendance (less than 56%).

OLS regression gives us the overall relationship between independent and dependent variables, but when you have a large study area with potentially different geographic influences, the GWR can provide insight into the varying importance of independent variables. Therefore, GWR was tested with many iterations of the independent variables to identify the combination that yields the most significant results but also low multicollinearity.

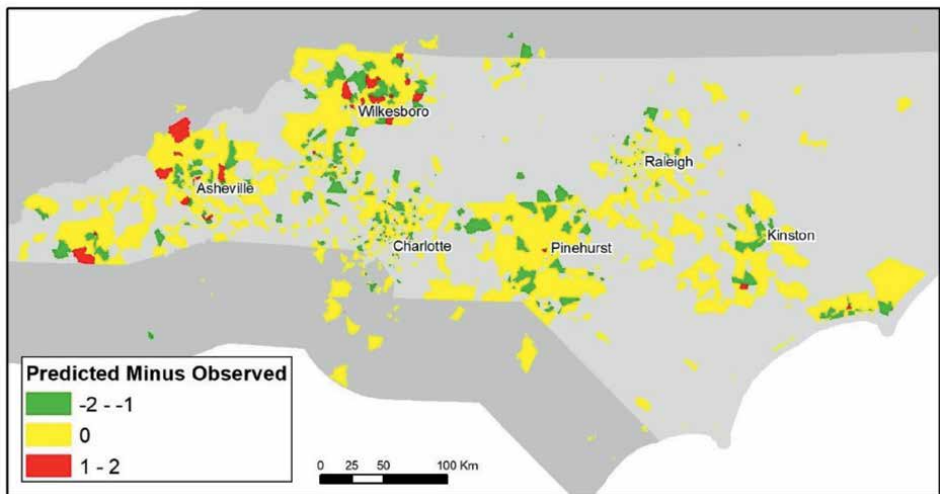
There are a series of steps one should take to interpret GWR results. First, in this study, the cross-validation method for determining bandwidth size resulted in an average of 55 neighbors which is a relatively small neighborhood around each observation/Block Group considering there were 1523 Block Groups. This is an excellent result because it informs us that local analysis is providing information about the significance of independent variables. Conversely, if the CV method resulted in a much larger number of neighbors, perhaps 500, then that would suggest a much larger geographic area should be used to derive the regression equations for each Block Group. Additionally, GWR performs best when you have a large number of observations because the analysis will have enough nearby observations to create a unique regression equation.

Next, an investigation of the residuals informs us as to whether the local regression equations are using observations that are close to the mean, or close to the regression line. In this study, only 154 block groups (10.1%) had standardized residuals greater than or less than 1.5 standard deviations, so most of the Block Groups (89.9%) had regression equations that predicted the dependent variable (attendance at the clinic) close to the mean. The Moran's I test for spatial autocorrelation confirmed that the GWR residuals were randomly distributed ( $z$ -score = 0.0955875), or not clustered, which means the local GWR did not have obvious missing variables.

The Condition Number is used to test for local multicollinearity, where values above 30 may have unreliable results. None of the Block Groups had Condition Numbers above 30. Given the Moran's I and Condition Number results, the local  $R^2$  values indicate where the GWR equations fit the dependent variable (**Figure 5**). The areas highlighted in blue (0.338 to 0.5) indicate places where the local regression equations are not explaining as much of the variance in comparison to the areas in light and dark red (0.601 to 0.784) where there were very high regression results indicating equations that contain sufficient variables to predict attendance. To test whether the local  $R^2$  results were randomly distributed, a Moran's I Spatial Autocorrelation test had a  $z$ -score of 139.79 ( $p$ -value = 0.01) which is very high (greater than 2.58 is significant) indicating that the pattern was significantly clustered and therefore the GWR  $R^2$  results are reliable.



**Figure 5.**  
 Local  $R^2$  results from geographically weighted regression analysis.

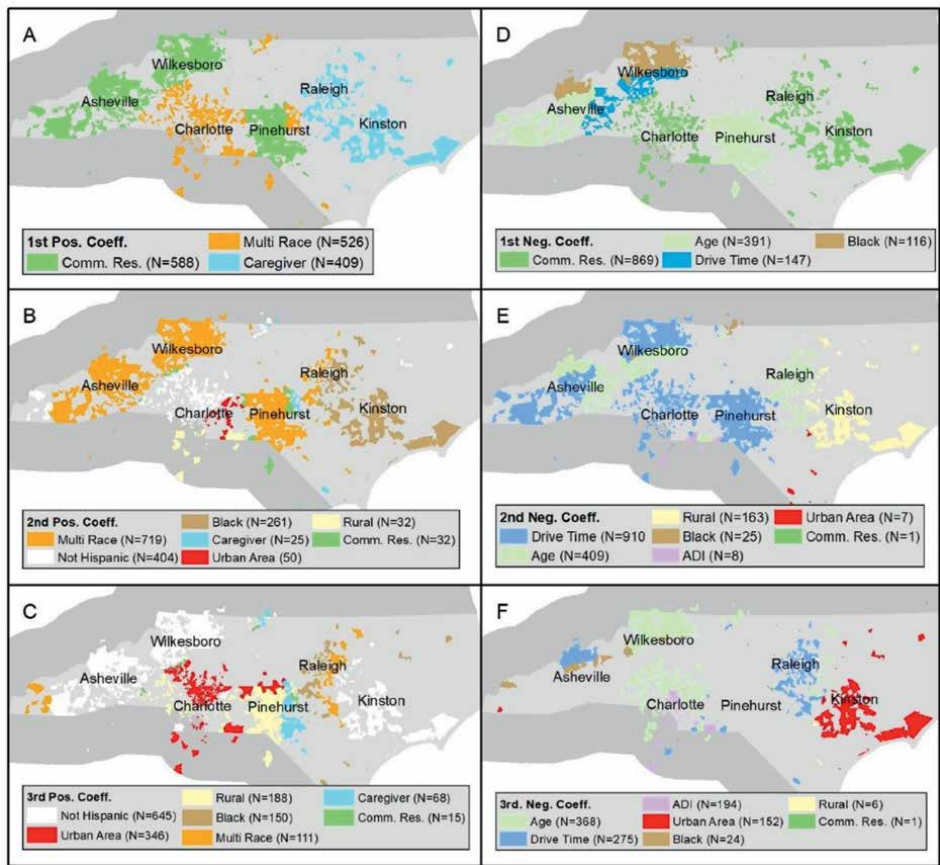


**Figure 6.**  
 Results from Geographically Weighted Regression (GWR) where observed attendance at the follow-up clinic was subtracted from predicted attendance. Areas in yellow (93%) had the same predicted attendance as actual (observed), areas in green had fewer predicted (4%) than actual and areas in red (3%) had more predicted than actual.

Given the good condition number, Moran's I and  $R^2$  results, the next check is to see where the predicted attendance differs from the observed attendance (**Figure 6**). Only 60 Block Groups (3.94%) under predicted the attendance by 1 or 2 people and only 43 Block Groups (2.82%) over predicted by 1 or 2 people. To test whether the difference between predicted and observed was clustered, the Moran's I Spatial Autocorrelation test had a z-score of 1.082583 indicating that the pattern is not significantly different from random ( $P = 0.278993$ ). Given that more than 93% of the Block Groups predicted attendance within 1 person of the actual attendance,

we concluded that the variables included in the GWR analysis were able to predict attendance at the follow-up clinic.

Given the high local  $R^2$ , excellent predicted versus observed results, and random standardized residuals, the next step was to investigate where each independent variable was important across the study area by the strength of each coefficient. For each Block Group, we identified the top three variables with the greatest contribution (largest coefficients) and most negative coefficients (Figure 7). The density of Community Resources had both a strong positive relationship with attendance in the south-central (Pinehurst area) and western regions (Figure 7A) and a negative relationship in urban (Charlotte and Raleigh) and eastern areas (Figure 7D). The number of people with multiple races had the largest and second largest positive variables in many areas. In the Raleigh and eastern areas, the second most positive influence was having a caregiver (Figure 7B). Recall that the OLS results indicated that having a caregiver was one of the most important variables and the GWR analysis corroborates this relationship and illustrates where this is important.



**Figure 7.** Geographic weighted regression results where the largest positive (A, B, and C) and negative (D, E, and F) independent variables for each Census Block Group. A is the largest positive coefficient, B is the 2nd largest and C is the 3rd largest coefficient. Independent variables also had a negative relationship with attendance at the follow-up clinic where D had the largest negative coefficient, E is the 2nd largest negative coefficient and F is the 3rd largest negative coefficient.

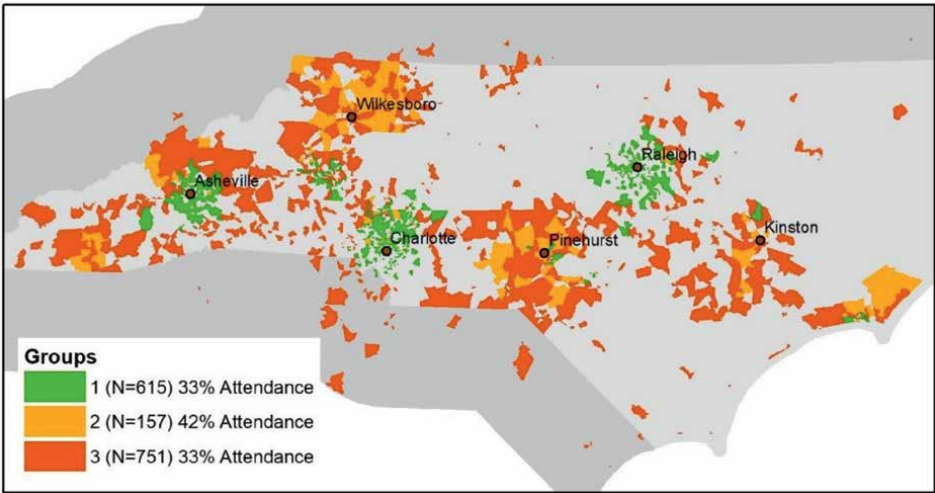


The multi-race characteristic was consistently a positive variable in the Charlotte and Pinehurst south-central areas (**Figure 7A–C**). In the far eastern part of the study area, being black was the 3rd highest positive coefficient while in the western area being black had a negative relationship with attendance at the follow-up visit. Interestingly, given the importance in the OLS results, the “not Hispanic” independent variable was not a 1st positive variable, but was very important in the 2nd and 3rd most positive variables and most dominant in the western part of the study area. This illustrates the importance of performing GWR to find where the independent variables are most important.

Similar to the OLS results, drive time was a consistent and important negative variable where the longer drive times had lower rates of attendance at the follow-up clinic visit, but in the 2nd and 3rd coefficient. Similarly, average age was a strong negative variable in the western area, where being older correlated with being less likely to attend the clinic visit. Lastly, the only area where being rural negatively related to attendance was in the eastern area. ADI, which was a strong positive variable in the OLS regression analysis, was a negative variable in the GWR analysis. This is the expected relationship and was only important in 202 Block Groups (13%) in the Charlotte region.

### 3.3 Grouping analysis

Given the identification of the importance, both positively and negatively, of the independent variables used in the regression analysis, the next step was to use the independent variables to create spatial clusters, or groups, where each group has shared similarities among the independent variables. Grouping Analysis identified 3 statistically significant groups (**Figure 8**). Like an unsupervised classification in remote sensing, the groups were identified, but we need to identify what these groups represent. Using the list of variables and their values within each group, the following characteristics defined each group:



**Figure 8.** Grouping analysis identified 3 groups across the study area, where N is the number of Block Groups. Group 1 had short drive time, high density of community resources and low ADI. Group 2 had average drive time, low density of community resources, very high number of caregivers, and above average ADI. Group 3 had very long drive time, low density of community resources, and above average ADI.

1. Group 1 (N = 615 Block Groups): urban, short drive time, very high community resources, and below average ADI. This group had an average attendance rate of 33% and is considered less vulnerable given these results.
2. Group 2 (N = 157): rural, average drive time, very low community resources, very high number of caregivers, not Hispanic, and above average ADI. This group had an average attendance rate of 42% and is considered moderate vulnerability because of these results.
3. Group 3 (N = 751): is rural much like group 1, but has very long drive time, low community resources, average number of caregivers, and above average ADI. This group had an average attendance rate of 33% and is considered highly vulnerable because of the results in this group.

Therefore, even though these are statistically significant groups, the rate of attendance is not dependent on these characteristics alone and care should be taken to not generalize too much and instead focus on the GWR results that provide spatially explicit variables of importance. The importance of the groups presented here is for comparing with the GWR and other results to identify areas of vulnerability, such as Group 3, and then design targeted strategies to improve patient care, especially in the most vulnerable groups.

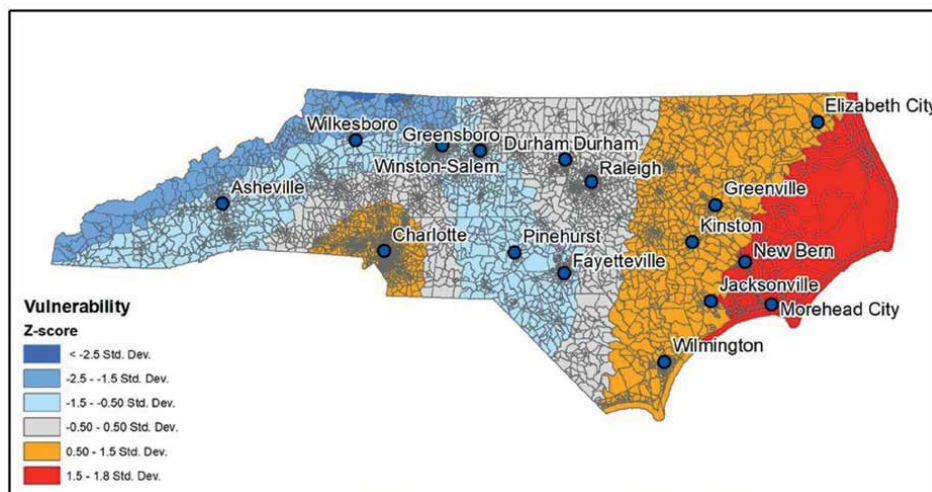
#### **4. Discussion**

This chapter has outlined a workflow for identifying patterns in point data (cluster analysis) and identifying spatial relationships between independent and dependent variables through regression analysis. The example study has identified the relationship between attendance at stroke care follow-up visits and drive time, ethnicity, and other social characteristics. These characteristics vary in space, geographically, where they are more positively or negatively related to clinic attendance and these results confirm that race and ethnicity are important factors [30–32]. Geographic variability in social characteristics has been related to health outcomes in previous studies. For example, one study found that accessibility to pharmaceutical products is directly related to varying social characteristics [33] and another study found disparate resources between urban and rural nursing home facilities [34].

Drive time varied significantly across the study area and was directly related to attendance at the follow-up clinic. These results are comparable to other studies that have shown the direct relationship between health care use and accessibility [16, 32, 35].

Cluster and Grouping Analysis identified statistically significant locations of social characteristics as related to attendance at the follow-up visit. Other studies have used similar cluster analysis techniques [18, 21–23, 36–39] and research is being conducted to develop new methods of cluster analysis [22]. As with all studies that use GWR, it is important to review the results to identify places where the regression equations have lower results (e.g.  $R^2$ , residuals, and over and under predictions) because these indicate places where the local regression equations are not explaining the variance in the input data. In this study, 65% (982 out of 1523) of the Block Groups had  $R^2$  greater than 0.5, which is very high for GWR analysis, but this also means 35% of the variance was not explained so future work could investigate these areas further to try and identify missing variables that may improve the explanation for attendance at the follow-up clinic.





**Figure 9.**  
 Statewide index of vulnerability derived using GWR weights. Areas in blue have very low z-scores, negative standard deviations and low to very low vulnerability. Conversely, areas in orange and red have positive standard deviations, high z-scores, and high to very high vulnerability.

Using GWR, we identified 10 important variables in the prediction of attendance at follow-up clinic visits. The unique equations for each Census Block Group have coefficients/weights for each variable and these weights varied across the study area. Other studies have used GWR to identify spatially varying patterns in health [27–29]. Building on the results from this study, the coefficients for the 10 independent variables were applied across the study area using spatial interpolation and then combined to create an overall Index of Vulnerability (**Figure 9**). This approach highlights areas with higher z-scores, higher vulnerability, in the eastern part of North Carolina and in the Charlotte area, with respect to attending a clinic follow-up visit. Future work could address these areas and the variables identified in the local GWR analysis to focus health care interventions.

## 5. Conclusions

The strategy and use of spatial statistics outlined here provides a framework for others to use as they investigate spatial patterns in data. There are many other approaches, such as geostatistical analyses that interpolate surfaces (e.g. kriging), thus the approach described here is in no way comprehensive, but is one strategy that logically progresses through a series of data analytic strategies to identify statistical patterns in vector-based geographic data. Given the relatively low rate of attendance at follow up clinics and geographic variability across the study area, this study identified several factors that are related to attendance (e.g. drive time, presence of a caregiver, presence of community resources). However, as with most geographic studies, future research is needed to further investigate additional factors that may relate to patient attendance because some parts of the study area had relatively low explanatory power. These results complement the overall results described in previous research [8, 40, 41] as well as other studies that have also concluded that health care access is directly related to proximity [16, 17, 20]. Given that this study also identified

drive time as one of the most important factors for clinic attendance, we recommend that hospitals take into consideration where patients live when they assign follow-up care. This could be accomplished by creating a regional system of integrated stroke follow-up clinics that would allow patients to receive stroke-specific follow-up care at a clinic that is closest to where they live regardless of hospital affiliation. This is especially important in urban areas where there may be several clinics to choose from. Several European countries have successfully implemented regional integrated healthcare delivery networks. With this approach, the nurse care coordinator could automatically compute drive-time, discuss the routing results with the patient, and provide them a list of clinics including the estimated drive time, enabling the patients a choice of follow-up clinics. Providing follow-up care by telehealth is another option especially as accessibility to reliable broadband technology continues to improve [42]. These strategies may improve the rate of attendance at follow-up clinic visits and would be more patient-centric than the current hospital-centric approach to stroke care in the U.S.

## **Acknowledgements**

Geospatial analysis was conducted at the University of North Carolina Spatial Analysis Lab. Patient data for this project was obtained through a project funded through a Patient-Centered Outcomes Research Institute Award (PCS-1403-14532). The contents of this chapter are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee. University of North Carolina Wilmington students Zachary Hahn and Alexandria Reimold assisted with the geocoding portion of the project.

## **Conflict of interest**

The authors declare no conflict of interest.

## **Disclaimer**

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute PCORI, its Board of Governors or Methodology Committee.

## Author details

Joanne N. Halls<sup>1\*</sup>, Barbara J. Lutz<sup>2</sup>, Sara B. Jones<sup>3</sup> and Matthew A. Psioda<sup>4</sup>

1 Department of Earth and Ocean Sciences, University of North Carolina  
Wilmington, Wilmington, NC, USA

2 School of Nursing, University of North Carolina Wilmington, Wilmington, NC,  
USA


3 Department of Epidemiology, Gillings School of Global Public Health, University of  
North Carolina Chapel Hill, Chapel Hill, NC, USA

4 Department of Biostatistics, Gillings School of Global Public Health, University of  
North Carolina Chapel Hill, Chapel Hill, NC, USA

\*Address all correspondence to: [hallsj@uncw.edu](mailto:hallsj@uncw.edu)

## IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics-2021 update: A report from the American Heart Association. *Circulation*. 2021;**143**(8):e254-e743. DOI: 10.1161/CIR.0000000000000950
- [2] Karp DN, Wolff CS, Wiebe DJ, Branäs CC, Carr BG, Mullen MT. Reassessing the stroke belt: Using small area spatial statistics to identify clusters of high stroke mortality in the United States. *Stroke*. 2016;**47**:1939-1942. DOI: 10.1161/STROKEAHA.116.012997
- [3] Hall EW, Vaughan AS, Ritchey MD, Schieb L, Casper M. Stagnating national declines in stroke mortality mask widespread county-level increases, 2010-2016. *Stroke*. 2019;**50**:3355-3359. DOI: 10.1161/STROKEAHA.119.026695
- [4] Centers for Disease Control. Interactive Atlas of Heart Disease and Stroke. 2019. Available from: <https://www.cdc.gov/dhhdsp/maps/atlas/index.htm>
- [5] Yang Q, Tong X, Schieb L, Vaughan A, Gillespie C, Wiltz JL, et al. Vital Signs: Recent Trends in Stroke Death Rates — United States, 2000-2015. *MMWR Morbidity and Mortality Weekly Report*. 2017;**2017**:933-939. Available from: <https://www.cdc.gov/mmwr/volumes/66/wr/mm6635e1.htm>
- [6] Powers WJ et al. Guidelines for the Early Management of Patients with Acute Ischemic Stroke: 2019 update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals from the American Heart Association/American Stroke Association. *Stroke*. 2019;**50**(12):e344-e418
- [7] Hirschman KB, Shaid E, McCauley K, Pauly MV, Naylor MD. Continuity of care: The transitional care model. *Online Journal of Issues in Nursing*. 2015;**20**:1. DOI: 10.3912/OJIN.Vol20No03Man01
- [8] Duncan PW, Bushnell CD, Jones SB, Psioda MA, Gesell SB, D'Agostino RB, et al. Randomized pragmatic trial of stroke transitional care: The COMPASS study. *Circulation: Cardiovascular Quality and Outcomes*. 2020;**13**(6):e006285. DOI: 10.1161/circoutcomes.119.006285
- [9] Johnson AM, Jones SB, Duncan PW, Bushnell CD, Coleman SW, Mettam LH, et al. Hospital recruitment for a pragmatic cluster-randomized clinical trial: Lessons learned from the COMPASS study. *Trials*. 2018;**19**:74. DOI: 10.1186/s13063-017-2434-1. Available from: <https://rdcu.be/dbbMU>
- [10] HIPxChange. Area Deprivation Index Datasets. 2020
- [11] Ferguson WJ, Kemp K, Kost G. Using a geographic information system to enhance patient access to point-of-care diagnostics in a limited-resource setting. *International Journal of Health Geographics*. 2016;**10**:15. DOI: 10.1186/s12942-016-0037-9
- [12] Gatto NM, Cockburn M, Bronstein J, Manthripragada AD, Ritz B. Well-water consumption and Parkinson's disease in rural California. *Environmental Health Perspectives*. 2009;**117**(12):1912-1918. DOI: 10.1289/ehp.0900852
- [13] Zandbergen PA. Geocoding quality and implications for spatial analysis. *Geography Compass*. 2009;**3**:647-680. DOI: 10.1111/j.1749-8198.2008.00205.x

- [14] Matci DM, Avdan U. Address standardization using the natural language process for improving geocoding results. *Computers, Environment and Urban Systems*. 2018;**70**:1-8. DOI: 10.1016/j.compenvurbsys.2018.01.009
- [15] The Joint Commission. Comprehensive Stroke Center. 2019 [October 10, 2019]. Available from: <https://www.jointcommission.org/en/accreditation-and-certification/certification/certifications-by-setting/hospital-certifications/stroke-certification/advanced-stroke/comprehensive-stroke-center/>
- [16] Brual J, Gravely-Witte S, Suskin N, Stewart DE, Macpherson A, Grace SL. Drive time to cardiac rehabilitation: at what point does it affect utilization? *International Journal of Health Geographics*. 2010;**9**:27. DOI: 10.1186%2F1476-072X-9-27
- [17] Hare TS, Barcus HR. Geographical accessibility and Kentucky's heart-related hospital services. *Applied Geography*. 2007;**27**:181-205. DOI: 10.1016/j.apgeog.2007.07.004
- [18] Barro AS, Kracalik IT, Malania L, Tsertsivadze N, Manvelyan J, Imnadze P, et al. Identifying hotspots of human anthrax transmission using three local clustering techniques. *Applied Geography*. 2015;**60**:29-36. DOI: 10.1016/j.apgeog.2015.02.014
- [19] Chen J, Roth RE, Naito AT, Lengerich EJ, MacEachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: An analysis of U.S. cervical cancer mortality. *International Journal of Health Geographics*. 2008;**7**(1):57. DOI: 10.1186/1476-072X-7-57
- [20] Coppi R, D'Urso P, Giordani P. A fuzzy clustering model for multivariate spatial time series. *Journal of Classification*. 2010;**27**(1):54-88. DOI: 10.1007/s00357-010-9043-y
- [21] Fritz CE, Schuurman N, Robertson C, Lear S. A scoping review of spatial cluster analysis techniques for point-event data. *Geospatial Health*. 2013;**7**(2):183-198. DOI: 10.4081/gh.2013.79
- [22] Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z, et al. Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics*. 2016;**2016**:15. DOI: 10.1186/s12942-016-0056-6
- [23] Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Statistics in Medicine*. 2008;**27**(25):5111-5142. DOI: 10.1002/sim.3342
- [24] Yamada I, Rogerson PA, Lee G. GeoSurveillance: A GIS-based system for the detection and monitoring of spatial clusters. *Journal of Geographical Systems*. 2009;**11**(2):155-173. DOI: 10.1007/s10109-009-0080-1
- [25] Iftimi A, Montes F, Mateu J, Ayyad C. Measuring spatial inhomogeneity at different spatial scales using hybrids of Gibbs point process models. *Stochastic Environmental Research and Risk Assessment*. 2017;**31**(6):1455-1469. DOI: 10.1007/s00477-016-1264-0
- [26] Roberson S, Dutton M, Macdonald M, Odoi A. Does place of residence or time of year affect the risk of stroke hospitalization and death? A descriptive spatial and temporal epidemiologic study. *PLoS One*. 2016;**11**(1):13. DOI: 10.1371/journal.pone.0145224

- [27] Kauh B, Schweikart J, Krafft T, Keste A, Moskwyn M. Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics*. 2016;**15**(1):38. DOI: 10.1186/s12942-016-0068-2
- [28] Cabrera-Barona P, Murphy T, Kienberger S, Blaschke T. A multi-criteria spatial deprivation index to support health inequality analyses. *International Journal of Health Geographics*. 2015;**14**:11. DOI: 10.1186/s12942-015-0004-x
- [29] Comber AJ, Brunsdon C, Radburn R. A spatial analysis of variations in health access: Linking geography, socio-economic status and access perceptions. *International Journal of Health Geographics*. 2011;**10**:44. DOI: 10.1186/1476-072X-10-44
- [30] Plantinga L, Howard VJ, Judd S, Muntner P, Tanner R, Rizk D, et al. Association of duration of residence in the southeastern United States with chronic kidney disease may differ by race: The REasons for geographic and racial differences in stroke (REGARDS) cohort study. *International Journal of Health Geographics*. 2013;**12**(1):17. DOI: 10.1186/1476-072X-12-17
- [31] Moore JX, Donnelly JP, Griffin R, Safford MM, Howard G, Baddley J, et al. Community characteristics and regional variations in sepsis. *International Journal of Epidemiology*. 2017;**46**(5):1607-1617. DOI: 10.1093/ije/dyx099
- [32] Wennerholm C, Grip B, Johansson A, Nilsson H, Honkasalo M-L, Faresjö T. Cardiovascular disease occurrence in two close but different social environments. *International Journal of Health Geographics*. 2011;**10**:5. DOI: 10.1186/1476-072X-10-5
- [33] Amstislavski P, Matthews A, Sheffield S, Maroko AR, Weedon J. Medication deserts: Survey of neighborhood disparities in availability of prescription medications. *International Journal of Health Geographics*. 2012;**11**(1):48. DOI: 10.1186/1476-072X-11-48
- [34] Lin S-W, Yen C-F, Chiu T-Y, Chi W-C, Liou T-H. New indices for home nursing care resource disparities in rural and urban areas, based on geocoding and geographic distance barriers: A cross-sectional study. *International Journal of Health Geographics*. 2015;**14**(1):28. DOI: 10.1186/s12942-015-0021-9
- [35] Freyssenge J, Renard F, Schott AM, Derex L, Nighoghossian N, Tazarourte K, et al. Measurement of the potential geographic accessibility from call to definitive care for patient with acute stroke. *International Journal of Health Geographics*. 2018;**17**:1. DOI: 10.1186/s12942-018-0121-4
- [36] van Rheenen S et al. An analysis of spatial clustering of stroke types, In-hospital mortality, and reported risk factors in Alberta, Canada, using geographic information systems. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*. 2015;**42**(5):299-309
- [37] Solano R et al. Retrospective space-time cluster analysis of whooping cough re-emergence in Barcelona, Spain, 2000-2011. *Geospatial Health*. 2014;**8**(2):455-461
- [38] Hagenlocher M et al. Assessing socioeconomic vulnerability to dengue fever in Cali, Colombia: Statistical vs expert-based modeling. *International Journal of Health Geographics*. 2013;**2013**:12

[39] Queiroz JW et al. Geographic information systems and applied spatial statistics are efficient tools to study Hansen's disease (leprosy) and to determine areas of greater risk of disease. *American Journal of Tropical Medicine and Hygiene*. 2010;**82**(2):306-314

[40] Gesell SB, Bushnell CD, Jones SB, Coleman SW, Levy SM, Xenakis JG, et al. Implementation of a billable transitional care model for stroke patients: The COMPASS study. *BMC Health Services Research*. 2019;**19**(1):1-14. DOI: 10.1186/s12913-019-4771-0. Available from: <https://rdcu.be/dbbPC>

[41] Lutz BJ, Reimold AE, Coleman SW, Guzik AK, Russell LP, Radman MD, et al. Implementation of a transitional care model for stroke: Perspectives from frontline clinicians, administrators, and COMPASS-TC implementation staff. *The Gerontologist*. 2020;**60**(6):1071-1084. DOI: [doi.org/10.1093/geront/gnaa029](https://doi.org/10.1093/geront/gnaa029)

[42] Adeoye O, Nyström KV, Yavagal DR, Luciano J, Nogueira RG, Zorowitz RD, et al. Recommendations for the establishment of stroke systems of care: A 2019 update. *Stroke*. 2019;**50**:e187-e210. DOI: 10.1161/STR.0000000000000173





# Statistical Model for the Quality of Panoramic Images of 2D Artifacts

*Ajith Wickramasinghe and Anusha Jayasiri*

### Abstract

The field of digital imaging emphasizes the quality of 2D artifact images, often facing challenges when capturing large images due to their wide field of view. A successful technique for addressing this is panoramic image creation, which involves merging overlapping segments from a larger image. Research in this domain focuses on understanding the visual quality aspects of panoramic images. This study aims to achieve two main objectives: firstly, to identify the key visual quality attributes associated with panoramic images, and secondly, to propose predictor variables for a statistical model that assesses the quality of 2D artifact panoramic images. To accomplish this, the researchers conducted a case study centered on generating panoramic images of mural paintings found in Sri Lankan temples. Through their investigation, they pinpointed color balance and noise & distortion as the most significant factors influencing the overall quality of these images. The researchers employed three methods to create the panoramas: an innovative technique, alongside two established methods—Photoshop and Hugin. Experts in Visual Arts evaluated the resulting images using a four-point Likert scale. Color balance and noise & distortion were used as predictor variables, while overall quality was the response variable. The gathered data underwent analysis using ordinal logistic regression within the Minitab statistical package. The outcomes underscored the pivotal roles of color balance and noise & distortion in determining the quality of panoramic images. Moreover, the findings showcased the model's high accuracy in fitting the data, reinforcing its effectiveness in assessing panoramic image quality.

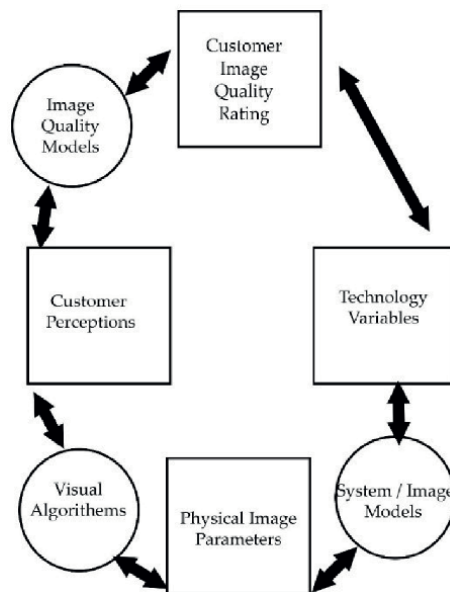
**Keywords:** quality attributes, color balance, noise and distortion, 2D artifacts, panoramic images

### 1. Introduction

It was identified that there are less amount of research works in Sri Lanka in the area of digitization of 2D artifacts for conservation. It was able to find out a report written by Schmid [1]. It talks about the main points such as the mission background & the objectives, identification of issues and draft strategy for conservation. According to the report, it had been mentioned that Central Cultural Fund (CCF) and Department of Archeology (DOA) requested international expert in conservation for the evaluation of the condition of Sigiriya Paintings in Sri Lanka. The report

presents about the discussion of main issues of Sigiriya paintings such as lack of documentation, need for monitoring, creation of a permanent record through 3D laser-scanning, Protection against rainwater, preliminary scientific investigation and documentation, emergency stabilization, construction of new visitors' platforms and reduction of number of visitors to the painting pocket. It is proposed some actions and precautions to rectify above issues. Digitization of all existing written and visual documentation and creation of simple computer repository of the existing documentation were proposed as some of the actions for above issues in the area of digital technology [1]. Quality aspect of the digitization process is an important research area. There were different notions related to image quality in the past. In response to the lack of a unified view on image quality, considerable effort was dedicated to imposing structure on the existing concepts. This led to the development of image quality theory, which was first introduced in 1988 [2]. Originally conceived as a four-way approach, this theory has evolved into what is now known as the "image quality circle." This robust framework effectively organizes the diverse range of ideas that contribute to the understanding of image quality. As a result, it serves as a valuable process model, streamlining and directing research, product development, marketing, and technology-related endeavors. The image quality circle can be diagrammatically shown as shown in the **Figure 1**.

There are four major components in this circle and those components are linked via one another by three links called, System/Image Models, Visual Algorithms and Image Quality Models. Technology Variables encompass the fundamental elements that can be controlled by imaging designers to alter the existing quality of an image. These variables encompass factors such as paper parameters, toner size, dots per inch (resolution) and other relevant aspects. Physical Image Parameters are quantitative and it can be any measurable aspect of an image. Further, Physical image parameters are called objective measures of image quality. One approach is applying the process



**Figure 1.**  
*Image quality circle.*

of image quality metrics to measure the quality of the image with reference to the full or partial reference of the original image. Structural similarity index matrix (SSIM) and peak signal-to-noise ratio (PSNR) are two examples for quality matrix. There is no limit for this physical image parameters except that they need to be physical and measurable. Customer perceptions are derived with a set of perceptual attributes of image quality, mostly visual, that form the basis of the quality preference or judgment by the customer. Some of the examples are darkness, sharpness, and graininess. Quality Ratings to the images given by customers refers to the comprehensive evaluation of image quality provided by them based on their judgment. This rating is represented on an interval scale, allowing for the expression of overall image quality as either a numerical value or a qualitative descriptor, such as “excellent”, “good” or “bad”. The authors conducted a research in which they generated panoramic images using three different methods: a novel method and two other existing methods. Experts in the field of Visual Arts assessed the visual quality of the created images. Subsequently, statistical models were developed using the ordinal logistic regression technique in the Minitab statistical package. The predictor variables and response variables from the collected data set were utilized in the creation of these models for the three methods. The results clearly indicate that the statistical model derived from the novel method outperformed the other two methods in terms of accuracy. These findings demonstrate that two crucial attributes significantly influence the quality of panoramic images, as supported by the highly accurate model. As a result, the proposed statistical model can be employed in any application that involves the generation of digital images with 2D artifacts.

## **2. Research problem**

In the background exploration, it has been identified the importance of conservation of two-dimensional artifacts in the Asian countries [3]. There are two notions in this field of research as preservation and conservation of valuable artifacts. Preservation involves safeguarding cultural property by implementing measures that effectively minimize physical and chemical decline of the damage [4]. Through such proactive actions, the loss of informational content can be effectively averted [5]. The UNESCO [6] definition says.

“In the domain of cultural property, the aim of conservation is to maintain the physical and cultural characteristics of the object to ensure that its value is not diminished and that it will outlive our limited time span”. The term “digital preservation” encompasses a range of methods aimed at ensuring the longevity of digital materials well into the future, as stated by the Council on Library and Information Resources [7]. This concept focuses on the sustainable management and accessibility of digital resources over time.

According to this definition, digital preservation is the management and maintenance of digital objects such as manuscripts, maps, rare books and other significant cultural materials. These digital objects can be accessed and used by future requirements. Further, it is required to study the theory and philosophy of conservation and explore the basis and framework of conservation, restoration, preservation theory and practice in the globalized world [8]. Therefore, conservation represents a more expansive domain, constituting a dedicated profession aimed at safeguarding cultural property for the benefit of future generations. The scope of conservation activities encompasses several important aspects such as examination, documentation,

treatment, and preventive care, all of which are bolstered by extensive research and education efforts [5]. This comprehensive approach ensures the enduring preservation and appreciation of valuable cultural assets.

It is a known fact that there are several types of valuable artworks in Sri Lanka which are needed to be conserved for the archeological aspect of next generation. It has been observed that various techniques are used for conservation of those artifacts in most of the places like Colombo museum, some historical temples in Sri Lanka such as Bellanwila Rajamaha viharaya, Kelani Rajamaha Viharaya and Sapugaskanda Rajamaha Viharaya. At the analysis of the methods applied for the conservation of valuable artifacts in those places, it can be understood that maintaining the quality of artifacts is a critical factor. In this context, identification of techniques to study the quality will be an important research area. Accordingly, Authors have identified that the problem of this research is “Statistical Model for the Quality of Panoramic Images of 2D Artifacts”.

### **3. Literature review**

#### **3.1 Literature review on the creation of panoramic images of 2D artifacts**

Author was able to critically review several researches done in different countries by analyzing the techniques used by them such as panoramic image creation. At the analysis, it was found that panoramic image creation is an application of image stitching technology [9–12]. Authors have identified that image stitching as the main technique of 2D image digitization [13, 14]. Sruthi et al. explain the idea of panorama image creation. There are two main techniques for image stitching as direct method and feature based method. Further, it talks about the approach and a specific method, called scale invariant feature transform (SIFT) for detecting local features in an image. After finding local features, overlapping areas are identified. Using dynamic programming method, a minimal cost path is selected to stitch images. Here, seven steps were used for the process of developing a panoramic image. By cutting at the overlapping places, images are merged together to form the final panorama image [15]. This research uses a scientific method for creation of a panorama image. Kokate et al. present the idea of image mosaicing based on feature extraction. It explains the idea of panoramic image production. Two approaches, direct and feature based techniques are discussed in this paper. Further, the difference between those two approaches are discussed. Components of image stitching are discussed as calibration, registration and blending. Steps of feature-based image stitching are discussed in detail. This paper talks about two concepts called local feature descriptor and feature detector. Accordingly, two techniques for describing local feature descriptor such as SIFT and speeded up robust features (SURF) are discussed in this paper by analyzing the individual pixels of the images. Harris corner detector is described as a feature detector with the technical details. RANSAC algorithm is used as the Homography detection algorithm in that research [12]. This research supports to understand the idea of local feature descriptor and the techniques for feature detection. It talks about two approaches with the comparison of the suitability rather than just applying a particular technique. Ultimately, better approach is selected for image stitching. This comparison helps Authors to analyze the suitability of a particular technique rather than using one technique directly in the stitching process within the research. Wu et al. discuss about the applications of SIFT in different fields, such as machine vision,

image retrieval and image stitching. This paper systematically analyzes SIFT and its variants [16]. Parallax is a displacement or difference in the apparent position of an object along two different views. Lens distortion is the appearance of straight lines as curved lines inward or outward to the center of the object. Scene motion is the visible lines of moving objects in a photograph. Exposure is the amount of light per unit area reaching to a frame of a photographic film. Ebtsam et al. present the idea of panoramic image creation. It talks about a different aspect, called field of view and how it affects for panoramic image creation. As there is a difference of the field of view between the human visual system and a typical camera, a requirement arises to get several pictures from a camera and stitch them to form a composite image with a much larger field of view. This paper aims to provide a comprehensive survey of feature-based image stitching. It covers the primary components involved in image stitching and presents a framework for a complete image stitching system based on feature-based approaches. By exploring these key aspects, this study seeks to offer valuable insights into the field of image stitching and its applications.

According to this research, there are many feature descriptors such as SIFT, SURF, Histogram of oriented gradients (HOG), Gradient Location and Orientation Histogram (GLOH), Principal Component Analysis SIFT (PCA-SIFT), Pyramidal HOG (PHOG), and Pyramidal Histogram of Visual Words (PHOW). Some of them are described in detail. Finally, the current challenges of image stitching process have also been discussed in this paper [17]. This paper uses a methodical approach. Actually, it elaborates the concept of feature-based techniques with more details. Levin and Weiss introduce the idea of having a quality panorama image by the evaluation of the techniques used for image stitching. Then, it explains how to measure the quality of image stitching. The focus of this study lies in two main areas: first, evaluating the similarity of the stitched image to each of the input images, and second, assessing the visibility of the seam between the stitched images. To achieve these objectives, an approach must be adopted that ensures the stitched image is as similar as possible to the input images both geometrically and photometrically, while simultaneously ensuring the seam between the stitched images remains imperceptible. This dual aspect approach aims to enhance the overall quality and seamlessness of the final stitched image.

It had been presented several cost functions for these requirements and define the mosaic image as their optimum [18]. Mikolajczyk and Schmid proposed a method to compare the performance of descriptors for local interested regions. It was calculated for different image transformations such as rotation, scale change, view point change, image blur, JPEG compression and illumination change. Further, experiment was done for the interest region descriptors in the presence of real geometric and photometric transformations. At the experiment, GLOW (Gradient Location and Orientation Histogram) obtains the best results followed by SIFT [19]. Balntas et al. have identified and demonstrated that the existing dataset and evaluation protocols regarding the benchmark have led to the inconsistency in results in the literature. So they have proposed a new public benchmark for local descriptors. They have mentioned that the new benchmark would enable the community to gain new insights since it is more significantly large than any existing dataset in the field [20]. Maponga presents that the image stitching has a lot of researches in area of medical imaging, computer vision, satellite imaging and video conferencing. It talks about two main approaches: direct method and feature-based method. Direct approach utilizes all the pixels of the image but it has disadvantages such as quite inflexible and greatly affected by exposure differences of the same object in different images to be stitched.

Furthermore, it is undesirable for real time applications as it performs slowly. Feature-based technique performs better depending on what exactly feature-based technique was implemented. Several techniques such as SIFT, SURF and PHOW were discussed. Advantages and Disadvantages were discussed [21]. Khan et al. Present that the concept of Image Mosaicing is currently a vibrant and dynamic research area within the realms of computer vision and computer graphics. It encompasses a wide array of diverse algorithms focused on detecting and describing features in images. These algorithms play a crucial role in the development and advancement of Image Mosaicing techniques, contributing to the exploration of new possibilities in visual representation and synthesis.

In this paper, it is studied image stitching technique called SIFT algorithm which is rotation, scale invariant as well as more effective in presence of noise. However, it needs high computational time. Additionally, the paper examines the SURF algorithm, which demonstrates robustness in terms of execution time and illumination invariance. Another algorithm explored is ORB, which exhibits rotation and scale invariance and boasts improved execution time. However, its performance tends to degrade in the presence of noise [22]. Patil and Gohatre present various techniques for the process of image stitching under various light conditions. Results obtained show that for the day light condition SIFT works better and for night light condition it is shown that Harris /Hessian detector performs better than SIFT detector [23]. There are different application areas of image stitching such as remote sensing which are applied for the domains of agriculture and natural disaster. Further, the application of remote sensing images becomes much widespread with the development of satellite technology [24].

Williams et al. discuss about post-processing solutions for creating quality digital images by combining captured portions of objects. Further, some alternative approaches such as robotic systems and some linear array scanners that are moved through the large images by stitching image components were also discussed. Even though they are high accurate, they are high expensive systems. This paper talks about digitization environment which affects the post-processing for the stitching procedure. Furthermore, it elaborates basic operations for image stitching and some software tools that can be used for panorama creation [9]. This paper is more important in terms of getting idea for alternative approaches for image stitching. It is really important to consider the digitization environment and the software tools which can be used for panorama creation in this research. Sarlin et al. have presented a new way to think about the feature matching problem. In most of the above applications, methods have been used by using local feature detecting and matching for stitching technique. But, in this paper, idea has been changed to use novel neural network architecture to learn the matching process from pre-existing local features [25].

### **3.2 Literature review on statistical data analysis: categorical data analysis**

Statistics can be used for the analysis of data in the nature of quantitative and qualitative. Statistical methods are used by researchers to analyze data, present the results and interpret them in the particular domain [26]. Statistical analysis is a crucial process behind how we make discoveries in the areas such as science, social science. Further, it will lay the foundation to make decisions based on the results, and make predictions. This allows us to understand a subject area much more deeply. It was researched to identify a suitable method to develop a quality model for the panoramic images. Accordingly, regression model in statistics [27]

was identified as one of the suitable models. A categorical variable is characterized by a measurement scale that comprises a defined set of categories. In this type of variable, data points are grouped into distinct, non-numeric categories rather than continuous numerical values.

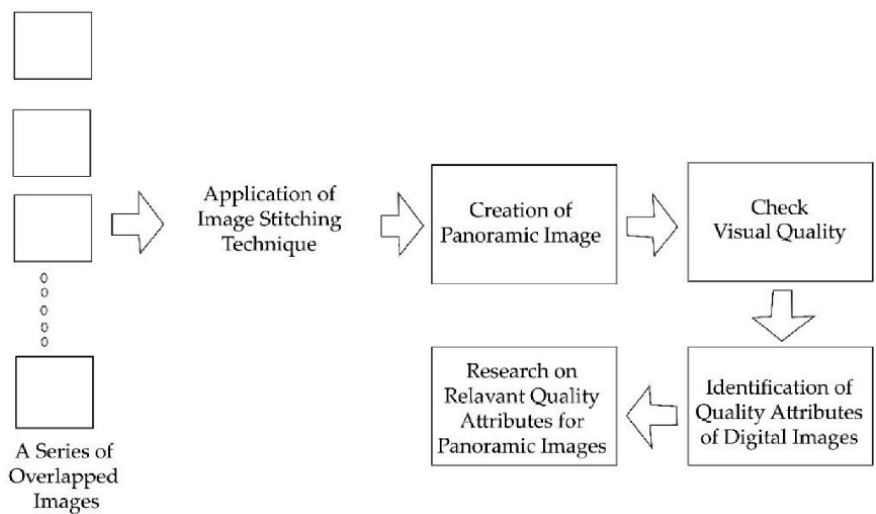
Accordingly data type that can be applied as categorical variable is defined as the categorical data [28, 29]. Ordinal variable is one of categorical data type which has a natural ordering. Some of the examples are: level of a course (high, medium and low), overall quality of an art work (excellent, good, average and poor). Categorical data can be analyzed using regression model in statistics. It was researched to identify a suitable method to develop a quality model for the panoramic images. Logistic regression was identified as a suitable method to find the significance of independent and dependent variables of the statistical model [30].

#### 4. Research methodology

This research methodology covers technical details related to five key areas. It includes a research of the essential attributes of panoramic images for visual quality, the process of capturing mural paintings and creating panoramic images from them, the implications of testing panoramic image creations using existing methods, the proposal of a new method for panoramic image development, and the subjective evaluation of the quality of the created panoramic images. These areas are thoroughly discussed within this study.

##### 4.1 Research on quality attributes of panoramic images

Figure 2 illustrates how to create panoramic images by stitching a series of overlapped image components. Image quality (IQ) models which is set up to quantify the overall image quality usually consist of a sub-set of quality attributes (QA) [31–36]. Therefore, the number of selected QAs is a very important step. Therefore,



**Figure 2.**  
*Steps of research on quality attributes for panoramic images.*

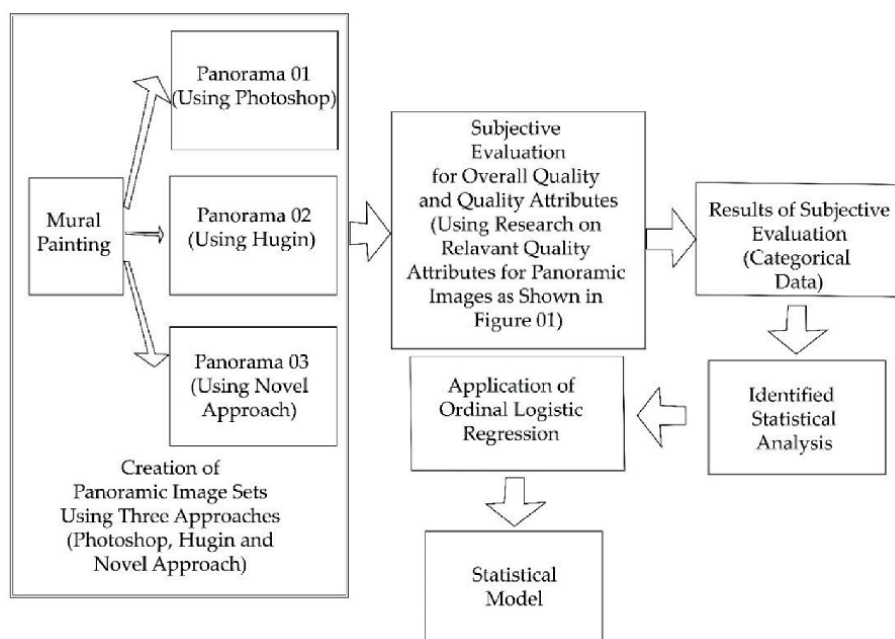
an issue exists between the accuracy of selected quality attributes (QAs) and the quantity of available QAs. Additionally, when the model exhibits high dimensionality, it leads to a more comprehensive evaluation of image quality (IQ), encompassing numerous aspects but also increasing complexity [37, 38]. On the other side, having too few QAs might result in inaccurate estimation of quality. So, selecting the most accurate set of QA will be the crucial task in this scenario. Accordingly, Authors researched on a set of image attributes that would contribute a quality panoramic image output [2, 31, 35–37]. It was well studied the underline theory of creating panoramic images, called image stitching. During the literature review, attention was drawn to the challenge of fully automated panoramic image stitching, and it discussed two categories of techniques prevalent in the research: direct and feature-based approaches [39, 40]. These categories represent distinct methods employed to tackle the task of seamlessly stitching panoramic images. According to Szeliski, it has been explained how the image stitching techniques evolved from the past up to the modern level of techniques and what are the millstones in the techniques developments which were resulted to enhance the quality of the panorama creation process [10]. It is worth emphasizing that the primary challenges associated with panorama creation include the presence of seams, the blurring effect caused by parallax, lens distortion, scene motion, and exposure variations among the panoramas. This clarification underscores the key factors that contribute to the overall quality and potential issues encountered during the process of creating panoramas [10, 41]. Based on the investigation, it was observed that noise, distortion and color balance emerged as the three primary quality attributes (QAs) with high significance in the domain of panoramic images [42]. Additionally, noise and distortions were found to be closely intertwined, forming a combined factor during the evaluation of the visual quality of digital images on display. Consequently, for the purpose of evaluation, these two factors were treated as a single entity.

#### **4.2 The process of capturing mural painting and developing panoramic images of mural paintings**

In this process, it was identified the requirement of photo shooting a set of historically valued 2D artifacts. Considering the traditional value of the artifacts, three temples having large-scale murals were selected for the research as a real-world case study. Further, it was ensured to cover different nature of artifacts to avoid the homogeneity. Subsequently, it was required to acquire a series of overlapped image components of the same image for creation of panoramic images of selected 2D artifacts. In this task, there are three different image acquisition methods which are covered large-scale mural paintings in the context of panoramic image development. There are three distinct setups used for capturing images in panorama creation. In the first setup, the camera is mounted on a tripod, and images are acquired by rotating the camera. On the other hand, the second setup involves placing the camera on a sliding plate, and images are obtained by moving the camera along the sliding plate. The last setup differs from the previous two setups, as it involves holding the camera in a person's hands. In this setup, images are captured by either rotating around a fixed spot or by walking in a direction perpendicular to the camera's viewing direction (**Figure 3**).

In all three set-ups, a still image digital camera or a smart phone embed with digital camera can be used to capture images. In this case study, majority of 2D artifacts which are existing in these traditional temples are in flat shape. Therefore, planner





**Figure 3.**  
 Developing a statistical model of predictor variables for the quality of panoramic images of mural paintings.

panorama images will have to be developed in this process. Further, authors wanted to use a simple approach which can be managed easily in the regular monitoring process of the conservation activities. So, third set-up is applied in this research which camera is held in a person's hands and the images are captured by walking in a direction perpendicular to the camera's view direction while ensuring the overlapping image components to avoid stitching issues.

### 4.3 Implications of the testing of panoramic image creations using existing methods

The purpose of this testing is to get the idea of developing panoramic images in different areas of digital images and the behavior of the final outputs of panoramic images with respect to color balance, noise and distortion quality attributes using different software tools. Accordingly, it was designed an experiment for this evaluation. Three types of digital images were selected and two types of available software which supports panoramic image creation were used for observing the quality and behavior of final outcomes. This testing was done for three types of image categories: existing 2D digital image set, 2D captured digital images set of an artifact and 2D captured digital image set of murals. Even though, several software tools were identified in the field, majority of them are expensive tools and are not available in the market for normal or academic purpose. Accordingly, two software tools: Photoshop (available in the market) and Hugin (an open software downloaded from internet) were used for testing the stitching technique under the above three testing cases.

By looking at the testing results, it was identified that there were some drawbacks and quality issues of the created panoramas in the areas of color balance, noise & distortion and overall quality for the selected three cases.

#### **4.4 Proposing a new method for the development of panoramic images**

Image stitching algorithm is used for the development of panoramic images. It was researched on a flexible and efficient mechanism to implement computer vision algorithms. Then, it was able to identify, OpenCV [43] that supports many algorithms related to computer vision and machine learning. According to the literature review, SIFT algorithm has been justified as a suitable algorithm for robust, reliable, efficient and quality output at the stitching operations for creating panoramas in area of 2D artifact even with noise and distortion. Based on the findings, it was identified that stitcher class in OpenCV software library is having stitching pipeline very similar to the algorithm proposed by Matthew and David [39]. Matthew and David has proposed their algorithm including the detection of SIFT features of images for panorama creation efficiently. Accordingly, stitcher class was selected for the implementation of the proposed algorithm. As authors have planned to use Python programming language, OpenCV-Python library was used in this implementation as Python bindings designed to solve computer vision problems.

#### **4.5 Subjective evaluation for the quality of created panoramic images**

To create panoramic images, the authors captured a series of digital image portions, meticulously covering selected large murals in the temples. Each captured portion was stored separately during the image acquisition process. The authors opted to create an odd number of panoramic image sets for each method, specifically generating five sets as part of this research.

Accordingly,  $5 \times 3$  panoramic images were obtained in this process. In the case of panoramic images, no original digital images are available for the reference [44]. So, objective evaluation IQ matrix cannot be used. Furthermore, esthetic aspect is very important in the context of artifact quality evaluation. As it is known, objective evaluation IQ matrix does not incorporate this factor at the evaluation. Accordingly, subjective evaluation is selected for this research for the evaluation of panoramic images using the quality attributes: color balance, noise & distortion with the overall quality of the panorama.

In this experiment, participants selected for evaluation were specifically chosen from the field of visual arts to ensure a comprehensive understanding of individual perceptions regarding quality attributes. A total of 15 experts from the faculty of visual arts were selected to participate in the evaluation process. The panel of experts were individually presented with sets of panoramic images using three distinct software tools. They were then asked to assess the quality attributes of each image, including color balance, noise & distortion, as well as overall quality. For color balance and overall quality, the experts used the following rating scale: Excellent (E), Good (G), Average (A), and Poor (P). The rating scale for noise & distortion consisted of the following options: No noise (N), Average (A), High (H), and Too Much (T). After the evaluation was completed, a qualitative dataset was obtained based on the assessments made using the three different methods.

The authors conducted research to determine a suitable regression model for this specific type of data and found that the ordinal logistic regression method could be effectively applied to assess the significance of independent and dependent variables in the statistical model [27]. Subsequently, they performed regression analysis using ordinal logistic regression and introduced a set of predictor variables for a statistical

model aimed at evaluating the quality of panoramic images of mural paintings through a novel, more accurate approach [28, 29, 45–47].

## 5. Results and discussion

### 5.1 Statistical data analysis: categorical data analysis for the data set obtained using new method

The primary objective of this research is to identify a comprehensive set of predictor variables to be incorporated into a statistical model for evaluating the quality of panoramic images depicting mural paintings. To accomplish this, an ordinal logistic regression model was utilized, taking into account the presence of four distinct ordered categories for the data, namely Excellent, Good, Average, and Poor. The evaluation process involved the utilization of the Minitab statistical package. Following statistical data analysis shows the testing of the level of significance of selected quality attributes to the overall quality for a statistical model for the quality of panoramic images of mural paintings through the results of the panoramic images created using the new method. The link function is Logit. The results of this categorical data analysis comprises several tables which are explained below.

**Table 1** presents the statistics of frequencies of responses for the overall quality of the generated panoramic images using the proposed method. The data is categorized into four ordered categories: Excellent, good, average, and poor, along with their corresponding total frequencies (count).

**Table 2** presents the key statistics derived from the regression analysis conducted in this study. This analysis offers valuable insights into the relevance of color balance and noise & distortion categories concerning the overall quality of panoramic images. Notably, the P-values for the “Average” and “Good” categories of the color balance predictor variable are both 0.000. These values being less than 0.05 indicate statistically significant associations between the “Good” and “Average” color balance categories and the overall quality. Conversely, the P-value for the “Poor” category of the color balance predictor variable is 0.994, indicating that there is no statistically significant association between the “Poor” category and the overall quality.

Moving on to the noise & distortion predictor variable, the P-value for the “Average” category is 0.013. Given that this P-value is smaller than 0.05, it indicates a statistically significant association with the overall quality. Conversely, the P-value for the “High” category of the noise & distortion predictor variable is 0.996, surpassing the significance threshold of 0.05. Consequently, it can be concluded that there is no

Variable	Value	Count
Overall Quality	Excellent	16
	Good	46
	Average	12
	Poor	1
	Total	75

**Table 1.**  
*Response information.*

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Const(1)	1.55373	0.705252	2.20	0.028			
Const(2)	7.78226	1.46466	5.31	0.000			
Const(3)	44.7692	6276.88	0.01	0.994			
Color balance							
Good	-2.9638	0.807985	-3.67	0.000	0.05	0.01	0.25
Average	-6.6842	1.41653	-4.72	0.000	0.00	0.00	0.02
Poor	-43.377	6276.88	-0.01	0.994	0.00	0.00	
Noise and distortion							
Average	-1.8036	0.722281	-2.50	0.013	0.16	0.04	0.68
High	-19.593	4454.98	-0.00	0.996	0.00	0.00	

*Log-likelihood = -35.545.*

**Table 2.**  
*Logistic regression table.*

DF	G	P-Value
5	71.937	0.000

**Table 3.**  
*Test of all slopes equal to zero.*

statistically significant association between the “High” category of noise & distortion and the overall quality.

Furthermore, the results highlight significant relationships between certain predictor variables and the overall quality of panoramic images. Notably, both the “Good” and “Average” categories of color balance show statistically significant associations with the overall quality. Additionally, the “Average” category of noise & distortion also demonstrates a statistically significant relationship with the overall quality. These findings shed light on the factors that impact the overall quality perception of panoramic images.

In addition to the previous findings, the results presented in **Table 3** provide further support for the presence of a predictor variable (either color balance or noise & distortion) that significantly influences the overall quality of panoramic images. This is evidenced by the G value of 71.937, which is notably large, and the P-value being less than 0.05. These findings suggest compelling evidence to conclude that at least one of the estimated coefficients of the predictor variable is significantly different from zero. Consequently, it can be inferred that either the color balance or noise & distortion (or both) play a crucial role in impacting the overall quality of the panoramic images.

**Table 4** displays the results of the Goodness-of-Fit Tests. The computed Chi-square statistics for both the Pearson and Deviance methods are 8.82540 and 9.99793, respectively. Furthermore, the associated P-values for these methods are 0.976 and 0.953, respectively. It is noteworthy that these P-values exceed the significance

Method	Chi-Square	DF	P
Pearson	8.82540	19	0.976
Deviance	9.99793	19	0.953

**Table 4.**  
*Goodness-of-fit tests.*

Pairs	Number	Percent	Summery measures	Value
Concordant	1316	84.7	Somers' D	0.81
Discordant	50	3.2	Goodman-Kruskal Gamma	0.93
Ties	188	12.1	Kendall's Tau-a	0.46
Total	1554	100.0		

**Table 5.**  
*Measure of association.*

threshold of 0.05. Consequently, it can be inferred that there is insufficient evidence to suggest that the model inadequately fits the data.

In **Table 5**, the measure of association between the response variable and the predicted probabilities is presented. The results indicate that 84.7% of the pairs show concordance, 3.2% are discordant, and 12.1% are tied pairs. These figures provide valuable insights into the relationship between the response variable and the predicted probabilities.

These findings indicate a high degree of agreement within the predicted probabilities, suggesting that the model possesses a strong predictive ability. Additionally, the statistics of Somers' D, Goodman-Kruskal Gamma (close to 1.0), and Kendall's Tau-a further support the superior predictive capacity of the model. This implies a robust association between the response variable and the predicted probabilities.

The combined findings from **Tables 3–5** offer further compelling evidence in favor of the proposed predictor variables, namely color balance and noise & distortion, for the statistical model designed to evaluate the quality of panoramic images of mural paintings, as demonstrated in **Table 2**. These results reinforce the effectiveness of the selected variables in capturing and assessing the overall quality of such panoramic images.

These results demonstrate a higher level of accuracy and reaffirm the suitability of the chosen predictor variables for the model.

## 5.2 Ordinal logistic regression analysis for the data set obtained using an existing method, Photoshop

This **Table 6** provides insights into the significance of the color balance and noise and distortion categories in relation to the overall quality of panoramic images. The statistical analysis yielded the following P-values for the respective categories of the color balance predictor variable: 0.254 for "Good," 0.897 for "Average," and 0.184 for "Poor." Moreover, the P-values for the "High" and "Too Much" categories of the noise & distortion variable are 0.182 and 0.028, respectively. These results indicate that, apart from the "Too Much" category of noise & distortion, which shows a P-value

Predictor	Coef	SE Coef	Z	P	Odds ratio	95% CI	
						Lower	Upper
Const(1)	-2.10231	1.42105	-1.48	0.139			
Const(2)	2.10230	1.42105	1.48	0.139			
Color balance							
Good	1.72536	1.51106	1.14	0.254	5.61	0.29	108.53
Average	-0.188025	1.45311	-0.13	0.897	0.83	0.05	14.30
Poor	-2.08513	1.56892	-1.33	0.184	0.12	0.01	2.69
Noise & distortion							
High	-0.975611	0.731260	-1.33	0.182	0.38	0.09	1.58
Too Much	-1.44111	0.656656	-2.19	0.028	0.24	0.07	0.86

*Log-Likelihood = -52.615.*

**Table 6.**  
*Photoshop: Logistic regression table.*

Predictor	Coef	SE Coef	Z	P	Odds ratio	95% CI	
						Lower	Upper
Const(1)	-25.5821	26863.0	-0.00	0.999			
Const(2)	-21.3436	26863.0	-0.00	0.999			
Color balance							
Good	23.0328	26863.0	0.00	0.999	1.00696E+10	0.00	*
Average	20.7671	26863.0	0.00	0.999	1.04486E+09	0.00	*
Poor	18.9467	26863.0	0.00	0.999	1.69215E+08	0.00	*
Noise & distortion							
Average	1.92115	2.09863	0.92	0.360	6.83	0.11	417.54
High	-0.0533333	2.10786	-0.03	0.980	0.95	0.02	59.03
Too much	-1.81660	2.17684	-0.83	0.404	0.16	0.00	11.59

*Log-Likelihood = -33.404.*

**Table 7.**  
*Hugin: Logistic regression table.*

lower than 0.05, there are no statistically significant associations between any of the color balance or noise & distortion categories and the overall quality.

### 5.3 Ordinal logistic regression analysis for the data set obtained using an existing method, Hugin

**Table 7** displays the P-values for the “Poor”, “Average,” and “Good,” categories of the color balance predictor variables. They are equal and the value is 0.999. Additionally, the P-values for the “Average,” “High,” and “Too Much” categories of

the noise & distortion variable are 0.360, 0.980, and 0.404, respectively. Since all the P-values in this table exceed the significance threshold of 0.05, it can be concluded that there are no statistically significant associations between any of the color balance or noise & distortion categories and the overall quality.

#### **5.4 Comparison of the ordinal regression analysis of already existing two methods with the proposed new method**

According to the research done for the identification of critical attributes for the visual quality of the panoramic images, it was proofed that color balance and noise & distortion are two significant attributes in this context. It implies that there should be a possibility to generate a statistical model which describes color balance and noise & distortion, are crucial attributes affecting the quality of panoramic images.

Comparing three logistic regression tables related to the new method and the other two methods, the logistic regression table derived from the novel method provides evidence supporting the enhanced accuracy of the proposed predictor variables, color balance, and noise & distortion, for the statistical model used to evaluate the quality of panoramic images of mural paintings, as presented in **Table 2**. This reaffirms the effectiveness and validity of the selected variables in accurately assessing the overall quality of such panoramic images.

Therefore, this result shows the superiority of the new method compared to other two existing methods.

## **6. Conclusion**

The authors are able to provide a solid justification for accomplishing their primary objectives of this study: to identify the visual quality attributes of panoramic images and to propose predictor variables for a statistical model aimed at assessing the quality of panoramic images of mural paintings.

The research encompassed a diverse selection of mural paintings, varying in terms of painting types and content complexity. These murals were situated on both the ceilings and walls of temples, exhibiting different orientations, including both vertical and horizontal shapes. The murals depicted a range of subjects, with some showcasing multiple objects and others featuring individual objects. As a result, a significant implication of this research is that the conclusions drawn can be applied to mural paintings of any type.

Furthermore, the study highlights an important consideration regarding the image capturing process. It emphasizes the need for careful attention to the level of overlap between consecutive image components and the maintenance of parallelism between the camera and the object. This ensures minimal noise & distortion while maximizing color balance. In terms of future enhancements, the research could be extended to encompass 3D artifacts to investigate potential variations in quality attributes and predictor variables within a 3D context.

Examining the regression analysis table using the novel method reveals key statistical findings within the logistic regression analysis. Specifically, it highlights the significance levels of color balance and noise & distortion quality attributes in relation to the overall quality of panoramic images. Notably, the P-values within this table indicate that both the “Good” and “Average” categories of the color balance predictor variable are recorded as 0.000. Given that these P-values are below the threshold of

0.05, the analysis indicates that statistically significant associations exist between the “Good” and “Average” categories of color balance and the overall quality of the panoramas.

In summary, the research effectively accomplishes its two main objectives: identifying the visual quality attributes of panoramic images and proposing predictor variables for a statistical model to evaluate the quality of panoramic images of mural paintings.

## **Author details**


Ajith Wickramasinghe\* and Anusha Jayasiri

Faculty of Dance and Drama, Department of Information Technology, University of the Visual and Performing Arts, Colombo, Sri Lanka

\*Address all correspondence to: [ajith.w@vpa.ac.lk](mailto:ajith.w@vpa.ac.lk)

## **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 



## References

- [1] Schmid W. Final Report - Sri Lanka Mission to the UNESCO World Heritage Site of Sigiriya. 2016
- [2] Engeldrum PG. A theory of image quality: The image quality circle. *Journal of Imaging Science and Technology*. 2004;**48**(5):446-456
- [3] Costin CL. Legal and policy issues in the protection of cultural heritage in South Asia and the pacific, cultural heritage in Asia and the pacific: Conservation and policy. In: *Proceedings of a Symposium*; 8-13 September 1991; Honolulu, Hawaii. Burbank, California: Wetland Graphics; 1993
- [4] Preservation and Access and the Research Libraries Group. Preserving Digital Information, Report of the Task Force on Archiving of Digital Information. [Internet]. 1996. Available from: <https://archive.org/details/PreservingDigitalInformationTaskForceReport1996> [Accessed: 18 September 2018]
- [5] About conservation. [Internet]. 2017. Available from: [http://www.conservation-us.org/about-conservation/definitions#.WZA\\_jFHhXIU](http://www.conservation-us.org/about-conservation/definitions#.WZA_jFHhXIU) [Accessed: 9 December 2019]
- [6] UNESCO. [Internet]. 2022. Available from: <https://uis.unesco.org/en/glossary>
- [7] Waters DJ. Digital Preservation? CLIR Issues. [Internet]. 1998. Available from: <http://www.clir.org/pubs/issues/issues.html> [Accessed: 5 September 2018]
- [8] Australia ICOMOS. Australia International Council on Monuments and Sites, Principles, Theory & Philosophy of Conservation. Australia ICOMOS; 2023 [Internet]. Available from: <https://australia.icomos.org/resources/australia-icomos-heritage-toolkit/principles-theory-philosophy-of-conservation/>. [Accessed: 2 February 2023]
- [9] Williams D, Williamson, Burns PD. Image stitching: Exploring practices, software and performance. In: *Proceedings in IS&T's Archiving Conference*. ResearchGate; 2013
- [10] Szeliski R. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*. 2006;**2**(1):1-104
- [11] Shikha A. A review on image stitching and its different methods. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2015;**5**(5). Available from: [www.ijarcsse.com](http://www.ijarcsse.com)
- [12] Kokate M, Wankhede V, Rohit S. Survey: Image Mosaicing based on feature extraction. *International Journal of Computer Applications*. 2017;**165**(1)
- [13] Hetal MP, Pinal JP, Sandip GP. Comprehensive study and review of image Mosaicing methods. *International Journal of Engineering Research & Technology (IJERT)*. 2012;**1**(9)
- [14] Dalwai A, Ansari M, Khan M. Operative use of image stitching algorithm based on feature extraction. *International Journal of Advanced Research in Computer Science*. 2015;**6**(2). Available from: [www.ijarcs.info](http://www.ijarcs.info)
- [15] Sruthi P, Dinesh S. Panoramic image creation. *IOSR Journal of Electronic and Communication Engineering (IOSR-JECE)*. 2017;12-24. e-ISSN: 2278-2834, p-ISSN: 2278-8735

- [16] Wu J, Cui Z, Sheng Victor S, Zhao P, Su D, Gong S. A comparative study of SIFT and its variants. *Measurement Science Review*. 2013;**13**(3)
- [17] Ebtsam A, Mohammed E, Hazem E. Image stitching based on feature extraction techniques: A survey. *International Journal of Computer Applications*. 2014;**99**(6)
- [18] Anat Levin AZ, Weiss Y. Seamless Image Stitching in the Gradient Domain. The Hebrew University of Jerusalem; 2000
- [19] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005;**27**(10):1615-1630
- [20] Balntas V, Lenc K, Vedaldi A, Mikolajczyk K. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. pp. 5173-5182
- [21] Maponga R. Image Stitching Techniques. ESE Senior Capstone Project; 2017 Tech Notes. 2017
- [22] Khan D, Maqsood A, Khan K. Operative use of image stitching algorithm based on feature extraction. *International Journal of Advanced Research in Computer Science*. 2015;**6**(2). Available from: [www.ijarcs.info](http://www.ijarcs.info)
- [23] Patil V, Gohatre U. Performance comparison of image stitching methods under different illumination conditions. *International Journal of Engineering Technology Science and Research (IJETSR)*. 2017;**4**(9). Available from: [www.ijetsr.com](http://www.ijetsr.com)
- [24] Jianxia W, Yawei W. Modified SURF applied in remote sensing image stitching. *International Journal of Signal Processing, Image Processing and Pattern Recognition*. 2015;**8**(8):1-10
- [25] Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A. Superglue: Learning feature matching with graph neural networks. In: *Proceedings of the 2 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. pp. 4938-4947
- [26] Hanneman RA, Kposowa AJ, Riddle M. *Research Methods for the Social Sciences: Basic Statistics for Social Research*. John Wiley & Sons, Ltd; 2013
- [27] Selvamuthu D, Das D. *Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control*. Springer Nature Singapore Pte Ltd; 2018. DOI: 10.1007/978-981-13-1736-1. ISBN: 978-981-13-1735-4. ISBN: 978-981-13-1736-1 (eBook)
- [28] Agresti A. *Analysis of Ordinal Categorical Data Second Edition*, Wiley Series in Probability and Statistics. John Wiley and Sons; 2010
- [29] Simonoff JS. *Analyzing Categorical Data*. Springer; 2003
- [30] Ramosacaj M, Hasani V, Dumi A. Application of logistic regression in the study of students' performance level (case study of Vlora University). *Journal of Educational and Social Research (Rome, Italy: MCSER Publishing)*. 2015;**5**(3)
- [31] Granados A, Pelayo VM, Arillo JR. Automatizing chromatic quality assessment for cultural heritage image digitization. *El Profesional de la Información*. 2019;**28**(3):e280305. DOI: 10.3145/epi.2019.may.05
- [32] Engeldrum PG. A new approach to image quality. In: *IS and T's 42nd Annual Conf. Proc*; IS and T. VA: Springfield; 1989. p. 461

- [33] Engeldrum PG. Measuring Key Customer Print Quality Attributes, TAPPI Symposium Process and Product Quality Division. 1989. p. 101
- [34] Bartleson C. The combined influence of sharpness and graininess on the quality of color prints. *The Journal of Photographic Science*. 1982;**30**:33-38
- [35] Engeldrum PG. Measuring customer perception of print quality. *Tappi Journal*. 1990;**73**:161
- [36] Engeldrum PG. Image Quality Modeling: Where Are We? IS and T's 1999 PICS Conference. 1999
- [37] Burns PD. Image Quality Concepts, Handbook of Digital Imaging. John Wiley and Sons, Ltd; 2015. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/9781118798706.hdi004>
- [38] Chandler DM. Seven challenges in image quality assessment: Past, present, and future research. *International Scholarly Research Notices*. 2013;1-53. Article ID 905685. DOI: 10.1155/2013/905685
- [39] Matthew B, David G. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*. 2007;**74**(1):59-73
- [40] Brown M, Lowe DG. Recognizing panoramas. In: *International Conference on Computer Vision*. Vol. 3. 2003. p. 1218
- [41] Szeliski R. *Computer Vision: Algorithms and Applications*. Springer Science and Business Media; 2010
- [42] Pedersen M, Bonnier N, Hardenberg JY, Albregtsen F. Attributes of a New Image Quality Model for Color Prints. *Society for Imaging Science and Technology*; 2009
- [43] OpenCV team. About. OpenCV. [Internet]. Available from: <https://opencv.org/about/> [Accessed: 22 February 2021]
- [44] Pedersen M, Hardeberg J. Survey of full-reference image quality metrics. In: *Høgskolen i Gjøviks rapportserie 5. The Norwegian Color Research Laboratory (Gjøvik University College)*; 2009. ISSN: 1890-520X
- [45] Upton GJG. *Categorical Data Analysis by Example*. Wiley and Sons, Inc, Publication; 2016
- [46] Andersen EB. *The Statistical Analysis of Categorical Data*. Springer; 1990
- [47] Downey AB. *Think Stats*. Sebastopol, CA: O'Reilly Media, Inc.; 2015. p. 95472



# Useful Block Designs in Biostatistics

*L. Rob Verdooren and Dieter Rasch*

## Abstract

Randomized Complete Block Designs (RCBD), Balanced Incomplete Block Designs (BIBD) and the so-called Generalized Lattice Designs as Alpha Designs are useful designs in Biostatistics. A complete table of BIBDs with the smallest number  $b$  of blocks with at most  $v = 25$  treatments and block sizes  $k$  for  $2 < k \leq \frac{v}{2}$  is presented. Such a table did not exist until now. The analysis of the different block designs (Randomized Complete Block design, BIBDs and Alpha Design) is here not done with the commercial statistical packages SAS or SPSS. These packages can now only be hired for a year and are quite expensive. We used the package of **R** for the analysis, which is free of charge and it is now used in most Universities.

**Keywords:** RCBD, BIBD, alpha designs, smallest BIBDs for  $v = 25$  treatments and block sizes  $k$  for  $2 < k \leq \frac{v}{2}$ , analysis with R

## 1. Introduction

Experiments in Biostatistics to compare treatments need homogeneous conditions. R.A. Fisher, a statistician at Rothamsted Experimental Station in Hertfordshire in England, published in 1926 an article “The arrangement of field experiments” [1]. Within 10 and a half pages Fisher gives all principles of experimental designs: replication, randomization and blocking. In agriculture with variety trials, the experimental field was often laid down next to a ditch. Plots parallel to the ditch have the same growing conditions, but plots farther away from the ditch have other growing conditions than plots next to the ditch. When one wants to investigate  $v$  varieties Fisher proposed starting with the  $v$  plots next to the ditch. These  $v$  plots form then a block of plots with the same growing conditions. The varieties are placed in this first block in a randomized order. The plots adjacent to the first block, farther away from the ditch form a new block of  $v$  plots. In this second block, the  $v$  varieties are placed again in a random order. This design is called a randomized complete block design for  $v$  varieties with  $b = 2$  blocks. The complete blocks are also called replications, because all the  $v$  varieties are present in this block. The statistical model for the yield in such a randomized complete block design with  $v$  varieties and  $b$  blocks is that of a two-fold analysis of variance and is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, \dots, v; j = 1, \dots, b; \quad (1)$$

In (1),  $y_{ij}$  is the random model of the observed yield  $y_{ij}$ ,  $\mu$  is the general mean,  $\alpha_i$  is the varietal effect,  $\beta_j$  is the block effect. Further, we have the side conditions  $\sum_{i=1}^v \alpha_i = 0$

and  $\sum_{j=1}^b \beta_j = 0$ . Furthermore, the  $e_{ij}$  terms are independently distributed random errors with a normal distribution, each having an expectation of 0 and a variance of  $\sigma^2$ . If the blocks of a block design are randomly selected from a huge set of blocks available, we have a mixed model with random  $\beta_j$ . Due to the rarity of this scenario in Biostatistics, we will not be discussing it in this chapter.

If all pairs of varietal mean differences  $\alpha_p - \alpha_q$  are by Least Squares Method estimated by  $\bar{y}_p - \bar{y}_q$  for  $p \neq q = 1, \dots, v$  and where  $\bar{y}_p = \sum_{j=1}^b y_{pj}/b$  and  $\bar{y}_q = \sum_{j=1}^b y_{qj}/b$  it can be shown that all pairs have the same variance  $\frac{2\sigma^2}{b}$ . This is a nice property of a randomized complete block design.

In Biostatistics, Complete Block Designs are very useful. The researcher must only be looking for the same experimental conditions for his  $v$  treatments. Often  $b$  blocks with the same number  $k$  experimental units per block are used.

We present the analysis with **R** of a randomized complete Block Designs in Section 2.

But unfortunately, the Biostatistician often comes in the situation that his number of treatments  $v$  is larger than the block size  $k$ . In this case, the Balanced Incomplete Block Design (*BIBD*) for the investigation of his  $v$  treatments is a good alternative design. A *BIBD* consists of  $b$  blocks each with  $k$  experimental units but  $k < v$ . The number of times a treatment is used in a *BIBD* is  $r$ , the number of replicates. Further, each pair of treatments occurs in a *BIBD*  $\lambda$  times together in all the blocks;  $\lambda = r \cdot (k - 1) / (v - 1)$  is called the number of concurrences. In a *BIBD*, we observe the property of equal variance for all treatment effects and treatment effect differences estimated by the Least Squares Method. Specifically, the variance for a treatment effect is given by  $(\sigma^2 / (r \cdot v)) \cdot (1 + k \cdot r \cdot (v - 1) / (\lambda \cdot v))$  and the variance for a treatment effect difference is given by  $\sigma^2 (2k / \lambda v)$ .

The *BIBDs* are discussed in Section 3, their analysis with **R** in Section 4.

Often the number of blocks in a *BIBD* is very large and the design is not useful in field trials or in other Biostatistics trials. Section 5 introduces an alternative method known as Alpha Designs and their analysis with **R** is presented in Section 6.

The analysis of the different block designs (Randomized Complete Block design, *BIBDs* and Alpha Design) is here not done with the commercial statistical packages SAS or SPSS. These packages can now only be hired for a year and are quite expensive. We used the package of **R** for the analysis, which is free of charge and it is now used in most Universities.

## 2. Analysis of a randomized complete block design with R

Kuehl [2] gives Example 8.1 of a randomized complete block design on pages 257–258. Current nitrogen fertilization recommendations for wheat included applications of specified amounts at specified stages of plant growth. The recommendations were developed through the use of periodic stem tissue analysis of nitrate content of the plants. Stem tissue analysis was thought to be very effective to monitor the nitrogen status of the crop and provide a basis for predicting required nitrogen for optimum production. In certain situations, however, the stem nitrate tests were found to over-predict the required nitrate amounts. Consequently, the researcher wanted to evaluate the effect of several different fertilization timing schedules on the stem tissue nitrate amounts and wheat production to refine the recommendation procedure. The treatment design included six different nitrogen application timing and rate schedules that were thought to provide the range of conditions necessary to evaluate the process.

For the purpose of comparison, a control treatment with no nitrogen, treatment (1), was included, as well as the current standard recommendation.

The experiment was conducted in an irrigated field with a water gradient along one direction of the experimental area. Since plant responses are affected by variability in the amount of available moisture, the field plots were grouped into blocks of six plots such that each block occurred in the same part of the water gradient. Thus, any differences in plant responses caused by the water gradient could be associated with the blocks. The resulting design was a randomized complete block design with four blocks of six field plots to which the nitrogen treatments were randomly allocated.

The layout of the experimental plots in the field is shown below. The observed NO<sub>3</sub> nitrogen content (ppm  $\times 10^{-2}$ ) from a sample of wheat stems is shown for each plot with the treatment number in parentheses before it.

Block 1	(2) 40.89	(5) 37.99	(4) 37.18	(1) 34.98	(6) 34.89	(3) 42.07	Irrigation
Block 2	(1) 41.22	(3) 49.42	(4) 45.85	(6) 50.15	(5) 41.99	(2) 46.69	Gradient
Block 3	(6) 44.57	(3) 52.68	(5) 37.61	(1) 36.94	(2) 46.65	(4) 40.23	↓
Block 4	(2) 41.90	(4) 39.20	(6) 43.29	(5) 40.45	(3) 42.91	(1) 39.97	

We now bring the data in the R-package using the following R commands.

```
> Block = c(rep(1,6), rep(2,6), rep(3,6), rep(4,6))
> Treatment = c(2,5,4,1,6,3,1,3,4,6,5,2,
6,3,5,1,2,4,2,4,6,5,3,1)
> NO3 = c(40.89, 37.99, 37.18, 34.98, 34.89, 42.07,
41.22, 49.42, 45.85, 50.15, 41.99, 46.69,
44.57, 52.68, 37.61, 36.94, 46.65, 40.23,
41.90, 39.20, 43.29, 40.45, 42.91, 39.97)
> Block = as.factor(Block)
> Treatment = as.factor(Treatment)
> DATA = data.frame(Block, Treatment, NO3)
> head(DATA)
  Block Treatment  NO3
1     1         2 40.89
2     1         5 37.99
3     1         4 37.18
4     1         1 34.98
5     1         6 34.89
6     1         3 42.07
> tail(DATA)
  Block Treatment  NO3
19    4         2 41.90
20    4         4 39.20
21    4         6 43.29
22    4         5 40.45
23    4         3 42.91
24    4         1 39.97
> with(DATA, tapply(NO3, Block, mean))
      1      2      3      4
38.00000 45.88667 43.11333 41.28667
> with(DATA, tapply(NO3, Treatment, mean))
      1      2      3      4      5      6
38.2775 44.0325 46.7700 40.6150 39.5100 43.2250
> ANOVA = aov( NO3 ~ Block + Treatment)
> summary(ANOVA)
              Df Sum Sq Mean Sq F value    Pr(>F)
Block           3  197.0    65.67    9.120 0.0012 **
Treatment       5  201.3    40.26    5.592 0.00419 **
Residuals     15   108.0     7.20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model)
Call:
lm(formula = NO3 ~ Block + Treatment, data = DATA)
Residuals:
    Min       1Q   Median       3Q      Max
-4.2633 -1.3381 -0.1625  1.4590  4.8683
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.206     1.643   20.816 1.77e-12 ***
Block2        7.887      1.549    5.091 0.00133 ***
Block3        5.113      1.549    3.301 0.004854 **
Block4        3.287      1.549    2.121 0.050953 .
Treatment2    5.755      1.897    3.033 0.008389 **
Treatment3    8.492      1.897    4.476 0.000444 ***
Treatment4    2.338      1.897    1.232 0.236941
Treatment5    1.233      1.897    0.650 0.525800
Treatment6    4.948      1.897    2.607 0.019803 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.683 on 15 degrees of freedom
Multiple R-squared:  0.7867,    Adjusted R-squared:  0.6729
F-statistic: 6.915 on 8 and 15 DF,  p-value: 0.0007034
```

**Note:** The coefficient Block2 represents the estimated difference in effect between Block 2 and Block 1.

The coefficient Treatment2 represents the estimated difference in effect between Treatment 2 and Treatment 1. For a Randomized Complete Block Design, the estimate with the Least Squares Method of a treatment is equal to the mean of the treatment in the experiment.

The estimate of the standard error of the differences of the least squares mean of two treatment means for a Randomized Complete Block Design uses the same estimate for the variance  $\sigma^2$ , it is the square of the residual standard error  $s = 2.683$ , hence  $s^2 = 7.1985$  with 15 degrees of freedom. The estimate of the standard error of the differences of the least squares means of two treatment means is given by  $\sqrt{(s^2 (2/4))} = \sqrt{3.5993} = 1.897$ .

```
> # To find the estimates of the Treatment means with
> # the Least Squares Method we use the R-package
> # lsmeans
> library(lsmeans)
> lsm = lsmeans(model, "Treatment" , alpha=0.05)
> lsm
```

Treatment	lsmean	SE	df	lower.CL	upper.CL
1	38.3	1.34	15	35.4	41.1
2	44.0	1.34	15	41.2	46.9
3	46.8	1.34	15	43.9	49.6
4	40.6	1.34	15	37.8	43.5
5	39.5	1.34	15	36.7	42.4
6	43.2	1.34	15	40.4	46.1

Results are averaged over the levels of: Block  
Confidence level used: 0.95

**Note:** The estimate of the standard error of a Least Square Mean of a treatment is  $\sqrt{s^2/4} = \sqrt{7.1985/4} = \sqrt{1.7996} = 1.342$ .

```
> contrast(lsm, method = "pairwise")
contrast estimate SE df t.ratio p.value
1 - 2 -5.755 1.9 15 -3.033 0.0742
1 - 3 -8.492 1.9 15 -4.476 0.0049
1 - 4 -2.337 1.9 15 -1.232 0.8150
1 - 5 -1.232 1.9 15 -0.650 0.9849
1 - 6 -4.947 1.9 15 -2.607 0.1553
2 - 3 -2.737 1.9 15 -1.443 0.7025
2 - 4 3.417 1.9 15 1.801 0.4934
2 - 5 4.522 1.9 15 2.383 0.2226
2 - 6 0.807 1.9 15 0.426 0.9978
3 - 4 6.155 1.9 15 3.244 0.0505
3 - 5 7.260 1.9 15 3.826 0.0168
3 - 6 3.545 1.9 15 1.868 0.4559
4 - 5 1.105 1.9 15 0.582 0.9907
4 - 6 -2.610 1.9 15 -1.376 0.7402
5 - 6 -3.715 1.9 15 -1.958 0.4079
```

Results are averaged over the levels of: Block  
P value adjustment: tukey method for comparing a family of 6 estimates.

But the investigator mentioned now that treatment 4 was the standard fertilizer recommendation for wheat. The nitrate nitrogen in the stem of the wheat plant measured throughout the growing season is used to assess nitrogen requirements for optimum wheat yields. The investigator would be interested in differences between any of the individual nitrogen timing treatments and the current recommendation of each stage of growth. The Dunnett's test can be used to compare the standard recommendation to each of the other timing treatments including the no nitrogen control treatment. The no nitrogen control provides a means of evaluating the nitrogen available without fertilization in these particular plots.



In the Dunnett's test table, as described by Dunnett [3, 4], for a degree of freedom (df) of 15 and  $k = 5$  (the number of treatments to compare with the control), with a significance level ( $\alpha$ ) of 0.05 (two-sided), the critical value is 2.82. Therefore we must compare the absolute value of the difference in means between treatment  $i$  and 4 with  $2.82 \times SE(\bar{y}_i - \bar{y}_4) = 2.82 \times \sqrt{3.5993} = 5.35$ .

Only the absolute difference of treatment 3 and 4, 6.16, is larger than 5.35.

$$|\bar{y}_1 - \bar{y}_4| = 2.34 < 5.35 \quad (2)$$

$$|\bar{y}_2 - \bar{y}_4| = 3.42 < 5.35 \quad (3)$$

$$|\bar{y}_3 - \bar{y}_4| = \mathbf{6.16} > 5.35 \quad (4)$$

$$|\bar{y}_5 - \bar{y}_4| = 1.11 < 5.35 \quad (5)$$

$$|\bar{y}_6 - \bar{y}_4| = 2.61 < 5.35 \quad (6)$$

With the R- package "nCDunnett", we can find the quantile of the Dunnett's test. This package can be used for the central and non-central Dunnett's test.

The degrees of freedom is  $\nu = 15$ . We have 5 treatments which are compared with the control hence, the correlation coefficient is 0.5 for two comparisons, this is given for all 5 treatment comparisons by the vector  $\rho = c(0.5, 0.5, 0.5, 0.5, 0.5)$ .

For the test, the non-centrality parameter is  $\delta = 0$ . This is indicated by the vector  $\delta = c(0, 0, 0, 0, 0)$  for the 5 treatment comparisons. We want to use the significance level  $\alpha = 0.05$  for a two-sided test. We indicate this by the confidence coefficient  $p = 1 - \alpha = 0.95$  and, in the command, we use the indication two-sided = TRUE. The computation is done with 32 points of the Gaussian quadrature method. The R- commands are then:

```
> library(nCDunnett)
> nu = 15
> rho = c(0.5, 0.5, 0.5, 0.5, 0.5)
> delta = c(0, 0, 0, 0, 0)
> p=0.95
> qNCDun(p, nu, rho, delta, 32, two.sided = TRUE)
[1] 2.816321
```

This quantile point 2.82 is also given by the table of Dunnett [4].

With the R-package "multcomp" for multi-comparisons, we can find the significance of the Dunnett's test where control is Treatment 4.

```
> Treatment = relevel(Treatment, ref = "4")
> levels(Treatment)
[1] "4" "3" "1" "2" "5" "6"
> # Dunnett's test
> MODEL = aov( NO3 ~ Block + Treatment)
> post_test = glht(MODEL, linfct = mcp(Treatment ="Dunnett"))
> summary(post_test)
      Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = NO3 ~ Block + Treatment)
Linear Hypotheses:
      Estimate Std. Error t value Pr(>|t|)
3 - 4 == 0      6.155      1.897   3.244  0.022 *
1 - 4 == 0     -2.338      1.897  -1.232  0.623
2 - 4 == 0      3.418      1.897   1.801  0.294
5 - 4 == 0     -1.105      1.897  -0.582  0.964
6 - 4 == 0      2.610      1.897   1.376  0.529
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Only the difference of Treatment 3 and Treatment 4 has a significant p-value  $\Pr(>|t|) = 0.022 < 0.05$ .

### 3. Balanced incomplete block designs

Balanced Incomplete Block designs consist of  $b$  blocks of size  $k$  experimental units each but  $k < v$ . The number of times a treatment is used in a *BIBD* is  $r$ , the number of replicates. Further, each pair of treatments occurs in a *BIBD* together in all the blocks  $\lambda$  times;  $\lambda = r \cdot (k - 1) / (v - 1)$  is called the number of concurrences.

In a *BIBD*, we observe the property of equal variance for all treatment effects and treatment effect differences estimated by the Least Squares Method. Specifically, the variance for a treatment effect is given by  $(\sigma^2 / (r \cdot v)) \cdot (1 + k \cdot r \cdot (v - 1) / (\lambda \cdot v))$  and the variance for a treatment effect difference is given by  $\sigma^2 (2k / \lambda v)$ .

#### 3.1 Theoretical background

We give at first an introduction into the theory behind *BIBDs*.

We consider experiments that include a treatment factor as well as disturbance factor that affects the experimental result. Block designs are experimental designs to eliminate the influence of that. Restricting ourselves to one treatment factor is not a loss of generality because, in the case of multiple disturbance factors, we can incorporate all factor level combinations as treatments of a new factor. We assume that the treatment factor has  $v$  levels, called treatments and the  $i$ th treatment occurs  $r_i$  ( $i = 1, \dots, v$ ) times then the  $r_i$  are called replications. The block may have  $b$  levels, called blocks. Block  $b_j$  has  $k_j$   $j = 1, \dots, b$  elements which are called block sizes.

If each treatment occurs equally often, say  $r$  times, in a block design, then the design is called to be equireplicate and if all blocks have the same size  $k$ , the design is called proper.

Any block design can be represented by the matrix

$$\begin{pmatrix} D(r_1, \dots, r_v) & \mathcal{N} \\ \mathcal{N}^T & D(k_1, \dots, k_b) \end{pmatrix} \quad (7)$$

which includes the diagonal matrices  $D(r_1, \dots, r_v)$  and  $D(k_1, \dots, k_b)$  of replications and block sizes, respectively. Additionally, it includes the matrix  $\mathcal{N}$ , which is called incidence matrix. The elements of the incidence matrix  $\mathcal{N} = (n_{ij})$  with  $v$  rows and  $b$  columns show how often the  $i$ th treatment (representing the  $i$ th row) occurs in the  $j$ th block (representing the  $j$ th column). If all  $n_{ij}$  is either 0 or 1, the incidence matrix and the corresponding block design are called binary.

The elements of the incidence matrix  $\mathcal{N} = (n_{ij})$  lead to the two diagonal matrices because  $\mathcal{N}e_b = D(r_1, \dots, r_v)$  and  $\mathcal{N}^T e_v = D(k_1, \dots, k_b)$ ; e.g.  $e_b$  is a column vector of length  $b$  of ones.

The levels of the block factor are called blocks. The  $b$  column sums  $k_j$  of the incidence matrix are the elements of  $D(k_1, \dots, k_b)$  and are called block sizes. The  $v$  row

sums  $r_i$  of the incidence matrix are the elements of  $D(r_1, \dots, r_v)$  and are called replications. A block design is complete, if the elements of the incidence matrix are all positive ( $n_{ij} \geq 1$ ). A block design is incomplete, if the incidence matrix contains at least one zero. Blocks are called incomplete, if in the corresponding column of the incidence matrix there is at least one zero.

In block designs, the randomization has to be done as follows: the experimental units in each block are randomly assigned to the treatments, occurring in this block. This randomization is done for each block separately.

In complete block designs with  $v$  plots per block, where each of them is assigned to exactly one of the  $v$  treatments, the randomization is completed. If  $k < v$ , (incomplete block designs) the abstract blocks, obtained by the mathematical construction have to be randomly assigned to the real blocks.

For incomplete binary block designs in place of the incidence matrix, often a shorter writing is in use. Each block is represented by a bracket including the symbols (numbers) of the treatments, contained in the block.

A block design with a symmetric incidence matrix is a symmetric block design.

It can easily be seen that the sum of the replications  $r_i$  as well as the sum of all block sizes  $k_j$  equals the Number  $N$  of the experimental units of a block design. Therefore, for each block design, we have:

$$\sum_{i=1}^v r_i = \sum_{j=1}^b k_j = N, \quad (8)$$

Especially for equireplicate and proper block designs ( $r_i = r$  and  $k_j = k$ ) this gives:

$$vr = bk \quad (9)$$

A (completely) balanced incomplete block design (*BIBD*) is a proper and equireplicate incomplete block design with the additional property that each pair of treatments occurs in equally many, say in  $\lambda$ , blocks. A *BIBD* with  $v$  treatments with  $r$  replications in  $b$  blocks of size  $k < v$ , is called a  $B(v, k, \lambda)$ -design. A *BIBD* for a pair  $(v, k)$  is called elementary, if it cannot be decomposed in at least two *BIBDs* for this pair  $(v, k)$ . A *BIBD* for a pair  $(v, k)$  is a minimum *BIBD* for this pair  $(v, k)$ , if  $r$  (and by this also  $b$  and  $\lambda$ ) is minimum. For a *BIBD* besides (9) we receive the relation for the number of concurrences  $\lambda$ :

$$\lambda(v-1) = r(k-1) \quad (10)$$

The Eqs. (9) and (10) are necessary but not sufficient conditions for the existence of a *BIBD*. For instance, the quintuple  $v = 16, r = 3, b = 8, k = 6, \lambda = 1$  fulfills the necessary conditions but no *BIBD* with these parameters exists. The reason is that Fisher's inequality is violated, which as Fisher [5] showed is also a necessary condition.

$$b \geq v \quad (11)$$

Hence the design with  $v = 16, r = 3, b = 8, k = 6, \lambda = 1$  is not a *BIBD*.

A researcher likes to have the experiment in replications, which means that a set of blocks forms a replication where all the treatments are present once. Such an incomplete block design is called **resolvable**. For a resolvable *BIBD*, Bose [6] showed that a necessary condition is:

$$b \geq r + v - 1 \quad (12)$$

But even if (9), (10) and (11) are valid, a *BIBD* not necessarily exists. Cases for this are  $v = 22, k = 8, b = 33, r = 12, \lambda = 4$  and  $v = 34, r = 12, b = 34, k = 12, \lambda = 4$ . The minimum *BIBD* for  $v = 22$  and  $k = 8$  is  $v = 22, k = 8, b = 66, r = 24, \lambda = 8$ .

The so-called unreduced or trivial *BIBD* can, for any pair  $(v, k)$  with positive integers  $v$  and  $k, k < v$ , always be constructed by forming all  $k$ -combinations of the  $v$  numbers.

Hence  $r = \binom{v}{k}, r = \binom{v-1}{k-1}, \lambda = \binom{v-2}{k-2}$ . Often a *BIBD* can be found as a part of such an unreduced *BIBD* and this is a reduced *BIBD*.

One case for which such a reduction is not possible is that with  $v = 8$  and  $k = 3$ .

There is no other case for  $v \leq 25$  and  $2 < k < v - 1$  where no unreduced *BIBD* exists; see for more cases with  $v > 25$  Rasch et al. [7] and Section 3.3.

In addition to Completely Balanced Incomplete Block Designs (*BIBDs*), Partially Balanced Incomplete Block Designs (*PBIBDs*) are also known, where not only one number of concurrences but two may occur. The consequence of this is that estimators of some treatment differences have two different variances. In this chapter, we do not consider these designs.

Example 3.1 (from Rasch and Herrendörfer [8, 9]).

For  $v = 7$  and  $k = 3$  the trivial *BIBD* is:

$$\begin{array}{l} (1, 2, 3) \ (1, 3, 6) \ (1, 6, 7) \ (2, 4, 7) \ (3, 5, 6) \\ \textbf{(1, 2, 4)} \ \textbf{(1, 3, 7)} \ (2, 3, 4) \ (2, 5, 6) \ (3, 5, 7) \\ (1, 2, 5) \ (1, 4, 5) \ (2, 3, 5) \ (2, 5, 7) \ (3, 6, 7) \\ (1, 2, 6) \ (1, 4, 6) \ (2, 3, 6) \ \textbf{(2, 6, 7)} \ (4, 5, 6) \\ (1, 2, 7) \ (1, 4, 7) \ (2, 3, 7) \ (3, 4, 5) \ \textbf{(4, 5, 7)} \\ (1, 3, 4) \ \textbf{(1, 5, 6)} \ (2, 4, 5) \ \textbf{(3, 4, 6)} \ (4, 6, 7) \\ (1, 3, 5) \ (1, 5, 7) \ (2, 4, 6) \ (3, 4, 7) \ (5, 6, 7) \end{array} \quad (13)$$

There are three elementary *BIBD*, two of them with parameters  $b = 7, r = 3, \lambda = 1$  and blocks  $\{(1, 2, 4), (1, 3, 7), (1, 5, 6), (2, 4, 5), (2, 6, 7), (4, 6, 7), (3, 4, 6)\}$ —they are in the trivial *BIBD* bold and italic. The incidence matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \quad (14)$$

A further *BIBD* with parameters  $b = 7$ ,  $r = 3$ ,  $\lambda = 1$  is the septuple  $\{(1,2,6), (1,3,4), (1,5,7), (2,3,7), (2,4,5), (4,6,7), (3,5,6)\}$ —its blocks are in the trivial *BIBD* italic but not bold.

The set of the residual 21 of the 35 blocks is a further elementary  $\{(1,2,4), (1,3,7), (1,5,6), (2,3,5), (2,6,7), (4,5,7), (3,4,6)\}$  block design—its blocks are in the trivial *BIBD* neither bold nor italic. Rasch and Herrendörfer [8, 9] showed that this set is an elementary *BIBD*.

Let  $N$  be the incidence matrix of a *BIBD* for the pair  $(v, k)$ . If we replace all Zeros of  $N$  by Ones and all Ones by Zero, we obtain the incidence matrix of a new *BIBD* which is called the complementary *BIBD* of the original one. If the original *BIBD* has the parameters  $v, r, b, k, \lambda$ , then the complementary *BIBD* has the parameters

$$v_c = v; b_c = b, r_c = v - r, k_c = v - k \text{ and } \lambda_{ck} = b - 2r + \lambda. \quad (15)$$

We consider the blocks  $\{(1,2,4), (1,3,7), (1,5,6), (2,4,5), (2,6,7), (4,6,7), (3,4,6)\}$  of the first elementary *BIBD* from above. By replacing each of the blocks by the corresponding block with the treatments not in the original block, we obtain the complementary block design  $\{(3,5,6,7), (2,4,5,6), (2,3,4,7), (1,4,6,7), (1,3,4,5), (1,2,3,6), (1,2,5,7)\}$ .

It is sufficient if a table of *BIBDs* contains only the designs with  $\{v, k \leq \frac{v}{2}\}$  because we can for the case  $k > \frac{v}{2}$  easily construct all designs using the complementary *BIBDs*.

Since October 2022, the R Package `ibd` by B. N. Mandal has the function `bibd(v,b,r,k,lambda,ntrial,pbar=FALSE)`

that generates a balanced incomplete block design with given number of treatments ( $v$ ), number of blocks ( $b$ ), number of replications ( $r$ ), block size ( $k$ ) and number of concurrences ( $lambda$ ); `ntrial` is the number of trials (default is one) and `pbar` is the logical value indicating whether progress bar will be displayed or not (default is `FALSE`). The function works best for most values of treatments ( $v$ ) up to 30 and block size ( $k$ ) up to 10. However, for block size ( $k$ ) up to 3, much larger number of treatments ( $v$ ) may be used.

In the output of the function `bibd()` of Mandal his package `ibd` gives besides the design is also given `NNP`, where  $N$  is the incidence  $(v \times b)$ -matrix  $(n_{ij})$  and  $NP$  is  $N'$  (the transpose of  $N$ ). For a *BIBD*, the matrix `NNP` is a  $(v \times v)$  symmetric matrix with  $r$  on the diagonal and  $lambda$  in the rest of the matrix. Also it gives `Aeff` for the lower bound to the A-efficiency of the generated design and `Deff` for the lower bound to the D-efficiency of the generated design. For a *BIBD*, the `Aeff` and `Deff` must be 1 or 0.999999.

### 3.2 A table of all smallest $b$ for *BIBD* for $v \leq 25$

To construct a *BIBD* is often not easy and needs methods of combinatorics (finite geometries and others) which are described in Rasch and Herrendörfer [8, 9]. Therefore, we present a link to a website of Springer [https://doi.org/10.1007/978-3-662-67078-1\\_9](https://doi.org/10.1007/978-3-662-67078-1_9) belonging to Rasch and Verdooren [10] for a complete table of *BIBDs* with the smallest number  $b$  of blocks with at most  $v = 25$  treatments and block sizes  $k$  for  $2 < k \leq \frac{v}{2}$ . Such a table did not exist until now. Fisher and Yates [11] published an incomplete table of *BIBDs*.

Below, we give all 110 Balanced Incomplete Block Designs (*BIBDs*) with  $v < 26$  and  $k \leq v/2$  and smallest number  $b$  of blocks of size. The *BIBDs* for  $k > v/2$  are the complementary *BIBDs*, which can be easily be obtained by replacing the treatments given in the blocks of the original *BIBD* by the treatments not occurring in the original blocks.

Example 3.1.

Design 2 (German “Plan” = Design, “Behandlungen” = Treatments) below.  
Plan 2

$v$	$k$	$b$	$r$	$\lambda$
7	3	7	3	1

Block Behandlungen ( = treatments)

---

1	2	4	1
2	3	5	2
3	4	6	3
4	5	7	4
5	6	1	5
6	7	2	6

Design 2 has the complementary BIBD with the parameters  $v_c=7, b_c=7, k_c = 7-3 = 4, r_c = 7-3 = 4, \lambda_c = 7-2 \cdot 3 + 1 = 2$  and the blocks are now:

Block Behandlungen ( = treatments)

---

1	3	5	6	7
2	1	4	6	7
3	1	2	5	7
4	1	2	3	6
5	2	3	4	7
6	1	3	4	5
7	2	4	5	6

In Rasch and Verdooren [10], we find the following table of smallest  $b$  for BIBDs. The designs of these 110 BIBDs can be found on the website of Springer [https://doi.org/10.1007/978-3-662-67078-1\\_9](https://doi.org/10.1007/978-3-662-67078-1_9).

If the R program ibd on a PC with 64 bits processor did not give a solution after 5 minutes, then we mentioned this design as not constructible by ibd. For Design 21, ibd gives a design with a NNP matrix that does not belong to a BIBD and also  $A_{eff}$  and  $D_{eff}$  was not 0.9999999.

$v$	$k$	$b$	$r$	$\lambda$	Design	Constructible by ibd
6	3	10	5	2	1	Yes
7	3	7	3	1	2	Yes
8	3	56	21	6	3	Yes
	4	14	7	3	4	Yes
9	3	12	4	1	5	Yes
	4	18	8	3	6	Yes
10	3	30	9	2	7	Yes
	4	15	6	2	8	Yes
	5	18	9	4	9	Yes
11	3	55	15	3	10	Yes

$v$	$k$	$b$	$r$	$\lambda$	Design	Constructible by ibd
12	4	55	20	6	11	Yes
	5	11	5	2	12	Yes
	3	44	11	2	13	Yes
	4	33	11	3	14	Yes
	5	132	55	20	15	Yes
13	6	22	11	5	16	Yes
	3	26	6	1	17	Yes
	4	13	4	1	18	Yes
	5	39	15	5	19	Yes
	6	26	12	5	20	Yes
14	3	182	39	6	21	Yes
	4	91	26	6	22	Yes
	5	182	65	20	23	Yes
	6	91	39	15	24	Yes
	7	26	13	6	25	Yes
15	3	35	7	1	26	Yes
	4	105	28	6	27	Yes
	5	42	14	4	28	Yes
	6	35	14	5	29	Yes
	7	15	7	3	30	Yes
16	3	80	15	2	31	Yes
	4	20	5	1	32	Yes
	5	48	15	4	33	Yes
	6	16	6	2	34	Yes
	7	80	35	14	35	Yes
17	8	30	15	7	36	Yes
	3	136	24	3	37	No
	4	68	16	3	38	Yes
	5	68	20	5	39	No
	6	136	48	15	40	No
18	7	136	56	21	41	No
	8	34	16	7	42	Yes
	3	102	17	2	43	Yes
	4	153	34	6	44	Yes
	5	306	85	20	45	No
	6	51	17	5	46	Yes
	7	306	119	42	47	No
	8	153	68	28	48	No

<i>v</i>	<i>k</i>	<i>b</i>	<i>r</i>	$\lambda$	Design	Constructible by ibd
19	9	34	17	8	49	No
	3	57	9	1	50	Yes
	4	57	12	2	51	Yes
	5	171	45	10	52	No
	6	57	18	5	53	Yes
	7	57	21	7	54	Yes
	8	171	72	28	55	No
	9	19	9	4	56	Yes
20	3	380	57	6	57	Yes
	4	95	19	3	58	Yes
	5	76	19	4	59	Yes
	6	190	57	15	60	No
	7	380	133	42	61	No
	8	95	38	14	62	No
	9	380	171	72	63	No
	10	38	19	9	64	No
21	3	70	10	1	65	Yes
	4	105	20	3	66	Yes
	5	21	5	1	67	No
	6	42	12	3	68	No
	7	30	10	3	69	Yes
	8	105	40	14	70	Yes
	9	35	15	6	71	No
	10	42	20	9	72	No
22	3	154	21	2	73	Yes
	4	77	14	2	74	Yes
	5	462	105	20	75	No
	6	77	21	5	76	Yes
	7	44	14	4	77	Yes
	8	66	24	8	78	Yes
	9	154	63	24	79	No
	10	77	35	15	80	Yes
23	11	42	21	10	81	No
	3	253	33	3	82	Yes
	4	253	44	6	83	Yes
	5	253	55	10	84	Yes
	6	253	66	15	85	No
	7	253	77	21	86	No



$v$	$k$	$b$	$r$	$\lambda$	Design	Constructible by ibd
	8	253	88	28	87	No
	9	253	99	36	88	No
	10	253	110	45	89	No
	11	23	11	5	90	Yes
24	3	184	23	2	91	Yes
	4	138	23	3	92	No
	5	552	115	20	93	No
	6	92	23	5	94	No
	7	552	161	42	95	No
	8	69	23	7	96	No
	9	184	69	24	97	No
	10	276	115	45	98	No
	11	552	253	110	99	No
	12	46	23	11	100	No
25	3	100	12	1	101	Yes
	4	50	8	1	102	No
	5	30	6	1	103	Yes
	6	100	24	5	104	No
	7	100	28	7	105	No
	8	75	24	7	106	No
	9	25	9	3	107	Yes
	10	40	16	6	108	No
	11	300	132	55	109	No
	12	50	24	11	110	No

**Table 1.**  
 BIBDs with smallest  $b$  for  $v \leq 25$  and  $\{v, 2 < k \leq \frac{v}{2}\}$ .

**Table 1** provides a comprehensive list of 110 designs of *BIBD* that can be used by experimenters to find the desired design, given that  $v \leq 25$ . These designs can be downloaded on the website of Springer [https://doi.org/10.1007/978-3-662-67078-1\\_9](https://doi.org/10.1007/978-3-662-67078-1_9) belonging to Rasch and Verdooren [10]. Be aware that the website uses the German words “Plan” and “BUBD” for *Design* and *BIBD* respectively.

But we know that designs with more than 100 blocks will be applied less frequently. In place of using such a design, we recommend to decrease the number of plots  $k$  if the number of blocks  $b$  is then smaller. Another possibility is, to increase  $v$  by a placebo treatment and use the design with this  $v$  with smaller  $b$ .

**Example 3.2.**  
 We consider design 87 with parameters  $v = 23$ ,  $b = 253$ ,  $r = 88$ ,  $k = 8$  and  $\lambda = 28$ . If we can use  $k = 7$  then the design 77 with parameters  $v = 23$ ,  $b = 44$ ,  $r = 14$ ,  $k = 7$  and  $\lambda = 4$  needs much less effort.

**Example 3.3.**

We consider design 14 with parameters  $v = 12$ ,  $b = 33$ ,  $r = 11$ ,  $k = 4$  and  $\lambda = 3$ . If we can use  $v = 13$  (with a placebo treatment) then the design 18 with parameters  $v = 13$ ,  $b = 13$ ,  $r = 4$ ,  $k = 4$  and  $\lambda = 1$  needs much less effort.

**3.3 A conjecture about trivial BIBD**

The following conjecture was already published in several articles—see for instance, Rasch and Herrendörfer [8, 9, 12–14], Rasch et al. [7].

For  $v = 8$  and  $k = 3$ , the elementary and smallest *BIBD*  $(v, b, r, k, \lambda)$  is the trivial one with  $b = 56$ ,  $r = 21$ ,  $\lambda = 6$ . This can be shown by verifying that there is no quintuple  $(8, b, r, 3, \lambda)$  with a number of blocks  $b < 56$  fulfilling Eqs. (2), (3) and (4). Certainly, the complementary *BIBD* of this design is both elementary and trivial. The conjecture is as follows:

**Conjecture** For  $3 \leq k \leq \frac{v}{2}$  the case  $(v, k) = (8, 3)$  (Design 3 in **Table 1**) is the only one where the trivial *BIBD* is elementary.

If we look at **Table 1**, we see that the conjecture is true for  $v \leq 25$ . In Rasch et al. [7] and Teuscher and Rasch [15], many cases are given where the conjecture is true. A counter-example could not be shown.

**3.4 BIBD with too large number of blocks  $b$** 

It is true that a *BIBD* needs often a large number  $b$  of blocks. The Biostatistician who cannot use the large number  $b$  of blocks for the *BIBD* can use a type of Incomplete Block Designs as Lattice Designs. In Cochran and Cox [16], Lattice Designs are given in Chapter 10. But these Lattice Designs exist only when the number of treatments  $v$  is an exact squares. An extension of these Lattice Designs are the generalized lattice designs as *Alpha Designs* by Patterson and Williams [17].

The variance of the estimator of treatment differences of these designs is not the same for all treatment differences but often about the same order. A researcher likes to have the experiment in replications, which means that a set of blocks forms a replication where all the treatments are present once. Such an incomplete block design is called *resolvable*. If  $v$  is equal to the product of  $b$  and  $k$ , resolvable *Alpha Designs* are possible. We describe them in Section 5.

**4. Analysis of a BIBD with R**

In a *BIBD*, the treatment effects are usually estimated with the intrablock analysis and the Least Squares Method using matrix solution of the normal equations with  $b = (X'X)^{-1}X'y$  and the variance of a varietal contrast  $p'b$  is found as  $\sigma^2 p'(X'X)^{-1}p$ . However, now *BIBD* trials are analyzed with recovery of interblock analysis using a mixed model where the replications and treatments are fixed factors, the incomplete block effects in the replications and the experimental error are random effects. With the REML (Restricted Maximum Likelihood method) from

Patterson and Thompson [18], the variance components are estimated. These estimated variance components are then inserted into the variance-covariance matrix  $V$ . The practical best linear unbiased estimator for  $b$  is calculated as  $b = (X'V^{-1}X)^{-1}X'V^{-1}y$ .

In  $R$ , the intrablock analysis and the recovery of interblock analysis information can be done. We demonstrate this with the following example of Cochran and Cox [16] page 443–444; with the following example of an experiment in a resolvable Balanced Incomplete Block Design with  $v = 6$  treatments,  $k = 2$  number of units per block,  $b = 15$  number of blocks,  $r = 5$  number of replications and  $\lambda = r \cdot (k - 1) / (v - 1) = 5 \cdot (2 - 1) / (6 - 1) = 1$ .

The objective of this experiment was to compare the effects of length of cold storage on the tenderness and flavor of beef roasts. The treatments were six periods of storage (0, 1, 2, 4, 9 and 18 days); these are denoted by the treatment symbols 1, 2, 3, 4, 5, 6, respectively. Thirty roasts from the round of an animal were used. Four muscles are provided 6 roasts, while 3 muscles each provided 2 roasts. The roasts from any muscle group naturally pair up, as each roast on the left side of an animal corresponds to a roast on the right side. From previous experience, it was believed that the 2 roasts in any pair would give the same results, hence these two roasts form a block. Variation among different pairs from the same muscle was expected to be somewhat larger, and variation among muscles to be still larger.

These options prompted the use of an incomplete block design in blocks of  $k = 2$ , each block comprising the left and right roasts in a pair. When grouping the blocks into replications, it was natural to put roasts from the same muscle to the same replicate. In this case, the first 4 muscles could be allocated to separate replications, allowing for a distinct replicate to be formed for each muscle. The remaining replication consisted of the 3 smaller muscles.

The treatments in a block were randomized. The order of the blocks from the design of this resolvable BIBD was also randomized per replication. Scoring for tenderness was done by 4 judges, each marking on a scale from 0 to 10. The scores shown are their total (out of 40). A high score indicating very tender beef.

The plan of this experiment with the treatment in parentheses and the scores for tenderness are given below.

Rep I	Rep II	Rep III	Rep IV	Rep V
<i>Block 1</i> (1) 7 (2) 17	<i>Block 4</i> (1) 17 (3) 27	<i>Block 7</i> (1) 10 (4) 25	<i>Block 10</i> (1) 25 (5) 40	<i>Block 13</i> (1) 11 (6) 27
<i>Block 2</i> (3) 26 (4) 25	<i>Block 5</i> (2) 23 (5) 27	<i>Block 8</i> (2) 26 (6) 37	<i>Block 11</i> (2) 25 (4) 34	<i>Block 14</i> (2) 24 (3) 21
<i>Block 3</i> (5) 33 (6) 29	<i>Block 6</i> (4) 29 (6) 30	<i>Block 9</i> (3) 24 (5) 26	<i>Block 12</i> (3) 34 (6) 32	<i>Block 15</i> (4) 26 (5) 32

The analysis with  $R$  is as follows for the BIBD from Cochran and Cox [16], page 443–444.

```

> rep = c(1,1,1,1,1,1,2,2,2,2,2,2,
          3,3,3,3,3,3,4,4,4,4,4,4,
          5,5,5,5,5,5)
> block = c(1,1,2,2,3,3,4,4,5,5,
            6,6,7,7,8,8,9,9,10,10,
            11,11,12,12,13,13,14,14,15,15)
> treat = c(1,2,3,4,5,6,1,3,2,5,
            4,6,1,4,2,6,3,5,1,5,
            2,4,3,6,1,6,2,3,4,5)
> score=c( 7,17,26,25,33,29,17,27,23,27,
           29,30,10,25,26,37,24,26,25,40,
           25,34,34,32,11,27,24,21,26,32)
> rep = as.factor(rep)
> block = as.factor(block)
> treat = as.factor(treat)
> # ANOVA Table such as given in Cochran and Cox page 444
> Anova = aov(score ~rep + treat + (rep:block))
> summary(Anova)
              Df Sum Sq Mean Sq F value    Pr(>F)
rep              4   298.5    74.62    9.649 0.00183 **
treat            5  1059.8   211.95   27.408 1.56e-05 ***
rep:block       10   213.4    21.34    2.759 0.06246 .
Residuals       10    77.3     7.73
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Note that in this Anova we cannot use the Sum of Squares
> # for treat because it is not the last variable for Residuals.
> Anova.1 = aov(score ~ block + treat)
> summary(Anova.1)
              Df Sum Sq Mean Sq F value    Pr(>F)
block          14  1051.5    75.10    9.712 0.000491 ***
treat           5   520.2   104.03   13.453 0.000359 ***
Residuals      10    77.3     7.73
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
> # In this Anova.1 table we have the correct test for treat.
> # Intrablock analysis
> model = lm( score ~ rep + block + treat)
> summary(model)

Call:
lm(formula = score ~ rep + block + treat)

Residuals:
    Min       1Q   Median       3Q      Max
-3.083 -1.167  0.000  1.167  3.083

Coefficients: (4 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.41667    2.27058   3.266 0.008482 **
rep2           7.91667    3.21109   2.465 0.033366 *
rep3           3.33333    3.21109   1.038 0.323687
rep4          12.08333    3.21109   3.763 0.003704 **
rep5           6.66667    3.21109   2.076 0.064615 .
block2         5.08333    3.21109   1.583 0.144491
block3         8.16667    3.21109   2.543 0.029199 *
block4         0.50000    3.21109   0.156 0.879360
block5        -3.00000    3.21109  -0.934 0.372179
block6            NA         NA      NA      NA
block7        -0.08333    3.21109  -0.026 0.979806
block8         8.83333    3.21109   2.751 0.020447 *
block9            NA         NA      NA      NA
block10        4.91667    3.21109   1.531 0.156732
block11       -1.41667    3.21109  -0.441 0.668469
block12            NA         NA      NA      NA
block13       -2.41667    3.21109  -0.753 0.469036
block14       -2.33333    3.21109  -0.727 0.484104
block15            NA         NA      NA      NA
treat2         9.16667    2.27058   4.037 0.002372 **
treat3        12.33333    2.27058   5.432 0.000288 ***
treat4        13.66667    2.27058   6.019 0.000129 ***
treat5        16.16667    2.27058   7.120 3.22e-05 ***
treat6        14.66667    2.27058   6.459 7.26e-05 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.781 on 10 degrees of freedom

Multiple R-squared:  0.9531,    Adjusted R-squared:  0.864 F-statistic:
10.7 on 19 and 10 DF,  p-value: 0.0002612
```

**Note:** The coefficient `rep2` represents the estimated difference in effect between rep 2 and rep 1.

The coefficient `treat2` represents the estimated difference in effect between treat 2 – treat 1.

The estimate of the Standard error for the differences between the least squares means of two treatment means is the same in the case of a *BIBD* the same. Estimate for the variance  $\sigma^2$  is square of residual standard errors = 2.271, hence  $s^2 = 7.734$  with 10 degrees of freedom. The estimate of the Standard error for the differences between the least squares means of two treatment means is  $\sqrt{[2 \cdot k \cdot s^2 / (\lambda \cdot v)]} = \sqrt{[2 \cdot 2 \cdot 7.734 / (1 \cdot 6)]} = 2.271$ , where  $\lambda = r \cdot (k - 1) / (v - 1) = 5 \cdot (2 - 1) / (6 - 1) = 1$ .

```
> ANOVA.model = anova(model)
> ANOVA.model
Analysis of Variance Table

Response: score

          Df Sum Sq Mean Sq F value    Pr(>F)
rep           4  298.47   74.617    9.6487 0.0018344 **
block        10  753.00   75.300    9.7371 0.0006399 ***
treat         5  520.17  104.033   13.4526 0.0003591 ***
Residuals    10   77.33    7.733

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Intrablock analysis with the Least Squares estimates
> # of the treat means. We use the R-package lsmeans.

> library(lsmeans)

> lsm = lsmeans(model, "treat" , alpha=0.05)

NOTE: A nesting structure was detected in the fitted model:block %in% rep

> lsm

treat lsmean    SE df lower.CL upper.CL
1      14.6 1.55 10     11.2     18.1
2      23.8 1.55 10     20.3     27.3
3      27.0 1.55 10     23.5     30.4
4      28.3 1.55 10     24.8     31.8
5      30.8 1.55 10     27.3     34.3
6      29.3 1.55 10     25.8     32.8

Results are averaged over the levels of: block, rep

Confidence level used: 0.95
```

**Note:** The estimate of the Standard error for the least squares estimate of a treatment mean is same in the case of a *BIBD* the same. Estimate for the variance  $\sigma^2$  is square of residual standard error  $s = 2.271$ , hence  $s^2 = 7.734$  with 10 degrees of freedom. The estimate of the Standard error for the least squares estimate of treatment mean is with  $\lambda = r \cdot (k - 1) / (v - 1) = 5 \cdot (2 - 1) / (6 - 1) = 1$  given by  $\sqrt{[(s^2 / (r \cdot v)) \cdot (1 + (k \cdot r \cdot (v - 1) / (\lambda \cdot v)))]} = \sqrt{[(7.734 / (5 \cdot 6)) \cdot (1 + (2 \cdot 5 \cdot (6 - 1) / (1 \cdot 6)))]} = \sqrt{2.40613} = 1.551$ .

```
> contrast(lsm, method = "pairwise")
contrast estimate    SE df t.ratio p.value
1 - 2          -9.17 2.27 10  -4.037  0.0213
1 - 3         -12.33 2.27 10  -5.432  0.0028
1 - 4         -13.67 2.27 10  -6.019  0.0013
1 - 5         -16.17 2.27 10  -7.120  0.0003
1 - 6         -14.67 2.27 10  -6.459  0.0007
2 - 3          -3.17 2.27 10  -1.395  0.7298
2 - 4          -4.50 2.27 10  -1.982  0.4129
2 - 5          -7.00 2.27 10  -3.083  0.0903
2 - 6          -5.50 2.27 10  -2.422  0.2351
3 - 4          -1.33 2.27 10  -0.587  0.9897
3 - 5          -3.83 2.27 10  -1.688  0.5668
3 - 6          -2.33 2.27 10  -1.028  0.8982
4 - 5          -2.50 2.27 10  -1.101  0.8704
4 - 6          -1.00 2.27 10  -0.440  0.9972
5 - 6           1.50 2.27 10   0.661  0.9826
```

Results are averaged over the levels of: block, rep

P value adjustment: tukey method for comparing a family of 6 estimates

**Note:** To get the interblock estimates, we must use the mixed model with the random effect of blocks in replications. We use, therefore, the R-package lme4.

```
> # Interblock estimates for treat means with random Blocks
> # Interblock analysis with recovery of information.
> library(lme4)
> MODEL = lmer( score ~ rep + treat + (1|block))
> MODEL
Linear mixed model fit by REML ['lmerModLmerTest']
Formula: score ~ rep + treat + (1 | block)
REML criterion at convergence: 122.1044
Random effects:
Groups   Name             Std.Dev.
block    (Intercept)  2.878
Residual                    2.717
Number of obs: 30, groups:  block, 15
Fixed Effects:
(Intercept)          rep2          rep3          rep4          rep5
treat2
    11.5634      2.6667      1.8333      8.8333      0.6667
9.0956
      treat3      treat4      treat5      treat6
    12.3618    13.7235    16.7777    15.6613
```

**Note:** To get the estimates for the fixed effects rep and treat, we use the R-package `lmerTest`.

```
lmerTest.
> library(lmerTest)
> MODEL.1 = lmer( score ~ rep + treat + (1|block))
> lsm = ls_means(MODEL.1)
> lsm
```

Least Squares Means table:

	Estimate	Std. Error	df	t value	lower	upper	Pr(> t )
rep1	22.8333	1.9978	9.2	11.4291	18.3307	27.3360	9.422e-07 ***
rep2	25.5000	1.9978	9.2	12.7639	20.9973	30.0027	3.595e-07 ***
rep3	24.6667	1.9978	9.2	12.3468	20.1640	29.1693	4.809e-07 ***
rep4	31.6667	1.9978	9.2	15.8506	27.1640	36.1693	5.282e-08 ***
rep5	23.5000	1.9978	9.2	11.7628	18.9973	28.0027	7.339e-07 ***
treat1	14.3634	1.5810	20.0	9.0848	11.0650	17.6617	1.575e-08 ***
treat2	23.4590	1.5810	20.0	14.8378	20.1606	26.7574	3.032e-12 ***
treat3	26.7251	1.5810	20.0	16.9036	23.4267	30.0235	2.701e-13 ***
treat4	28.0869	1.5810	20.0	17.7649	24.7885	31.3852	1.062e-13 ***
treat5	31.1410	1.5810	20.0	19.6967	27.8426	34.4394	1.505e-14 ***
treat6	30.0247	1.5810	20.0	18.9906	26.7263	33.3231	3.011e-14 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Confidence level: 95%

Degrees of freedom method: Satterthwaite

## 5. Alpha designs

Patterson and Williams [17] introduced the concept of *Alpha designs* for variety trials in binary connected incomplete block designs with a block size of  $k$ . They start with a rectangular array with column lengths  $k$  (the size of the incomplete blocks) of the sequence of 1, ...,  $v$  varieties and shift the columns according to an array. For many combinations of varieties  $v$  and block sizes  $k$ , they give a procedure to construct Alpha designs. The name “Alpha” comes from the first letter in the Greek alphabet that was used to construct the design. John and Williams [19] made more Alpha designs based on cyclic designs; see for the definition of cyclic design chapter 3 of their book. Tables of cyclic designs are given by John et al. [20] and Lamacraft and Hall [21]. There is a computer program CycDesign [22] available to generate incomplete block designs as alpha designs and cyclic designs (see the website of VSN-international: <http://www.vsnl.co.uk/software/cycdesign/>). This package has general algorithms for generating incomplete block and row-column designs, which give better results than the alpha and cyclic construction methods. In variety testing trials one wants to use



resolvable incomplete block designs where the design can be divided into  $r$  groups (= replications) such that each group contains each of the  $v$  crosses exactly once.

The resolvable incomplete block designs, and particularly the so-called *generalized lattice* (GL) or *Alpha designs*, have become most suitable for crop variety trials; because they make it easier to find designs for a large number of varieties and different (even small) sizes of incomplete blocks, see Williams [23] and Patterson et al. [24].

The program CycDesignN [22] gives such resolvable incomplete block designs. All these above-mentioned designs are connected. In a connected incomplete block design, one can estimate all differences between the varieties. Patterson and Silvey [25] indicate about 70% saving in use of land and labor for variety trials, if certain incomplete block designs are adopted rather than complete block designs. Partially replicated designs are now very popular for large variety trials: see e.g. Cullis et al. [26] and Williams et al. [27].

One well-known other block design procedure is the OPTTEX procedure from the SAS package but numerous other packages are available including a number of open-source **R** packages.

In oil palm breeding trials, the alpha designs are very useful to make connected partial diallel or incomplete diallel crossing scheme of the female parent *dura* and the male parent *pisifera* to produce the wanted *tenera* hybrids. Because the *tenera* palms are planted at the corners of equilateral triangles with side lengths of 9 m, the plots with  $6 \times 6$  palms are quite large. The Alpha designs of Patterson and Williams are then used to find resolvable incomplete block designs with the computer program CycDesignN. This is described in Verdooren [28]. See further Verdooren et al. [29] where the analysis of oil palm breeding trials in incomplete block designs is given for estimating the General Combining Ability of the parents using mixed models.

The varietal effects are usually estimated with the intrablock analysis and the Least Squares Method using matrix solution of the normal equations with  $b = (X'X)^{-1}X'y$  and the variance of a varietal contrast  $p'b$  is found as  $\sigma^2 p'(X'X)^{-1}p$ . However, now varietal trials are analyzed with recovery of interblock analysis information using a mixed model where the replications and varieties are fixed factors, the incomplete block effects in the replications and the experimental error are random effects.

With the REML (Restricted Maximum Likelihood) method of Patterson and Thompson [15], the variance components are estimated. These estimated variance components are then inserted into the variance-covariance matrix  $V$ . The practical best linear unbiased estimator for  $b$  is calculated as  $b = (X'V^{-1}X)^{-1}X'V^{-1}y$ . In **R**, the intrablock analysis and the recovery of information with interblock analysis can be done.

## 6. Analysis of an Alpha Design with R

Kuehl [2] gives the following exercise 10.4. A variety trial was conducted in an alpha design  $\alpha(0,1,2)$  resolvable design;  $\alpha(0,1,2)$  means an alpha design with: 0 some varieties are not together in a block, 1 some varieties are one times together in a block; 2 some varieties are twice together in a block. There were  $v = 18$  varieties in  $r = 4$  replicate groups. Hence, this is a resolvable design. There were 3 blocks with 6 varieties in each replicate. Hence, the block size is  $k = 6$  and the number of blocks  $b = 4 \cdot 3 = 12$ .

Varieties 1 and 5 are control varieties. The table gives first the yield  $y$  in kg/plot and then in parentheses the variety number.

Replicate I						
Block						
1	88.2 (5)	82.5 (10)	84.3 (15)	87.0 (6)	84.5 (12)	88.9 (8)
2	82.4 (1)	82.9 (14)	83.1 (3)	84.7 (13)	83.3 (16)	89.0 (4)
3	93.1 (2)	82.7 (11)	88.9 (17)	88.6 (18)	84.1 (9)	87.5 (7)
Replicate II						
Block						
4	85.4 (4)	73.0 (11)	84.2 (7)	80.3 (14)	79.6 (10)	86.0 (6)
5	87.9 (8)	85.1 (9)	79.4 (18)	80.7 (13)	89.3 (5)	81.5 (3)
6	82.4 (1)	88.5 (2)	87.0 (12)	85.4 (17)	85.9 (15)	79.1 (16)
Replicate III						
Block						
7	83.6 (6)	79.4 (17)	81.3 (4)	80.5 (9)	80.9 (8)	79.3 (1)
8	80.4 (7)	88.2 (5)	82.3 (14)	88.0 (12)	90.0 (2)	83.6 (3)
9	81.4 (18)	84.8 (15)	81.0 (10)	81.2 (13)	79.1 (11)	83.8 (16)
Replicate IV						
Block						
10	80.5 (16)	77.1 (11)	84.4 (17)	90.4 (6)	82.9 (14)	83.0 (12)
11	87.9 (8)	78.9 (18)	81.4 (1)	83.5 (2)	82.2 (15)	79.0 (3)
12	84.2 (7)	83.0 (10)	87.6 (9)	81.7 (13)	91.3 (5)	87.4 (4)

The analysis with **R** is as follows for the Alpha Design from Kuehl [2], Exercise 10.4.

```
> replicate = c(rep(1,18), rep(2,18), rep(3,18), rep(4,18))
> block = c(rep(1,6), rep(2,6), rep(3,6), rep(4,6), rep(5,6),
  rep(6,6), rep(7,6), rep(8,6), rep(9,6), rep(10,6),
  rep(11,6), rep(12,6))
> variety = c(5,10,15,6,2,8,1,14,3,13,16,4,2,11,17,18,9,7,
  4,11,7,14,10,6,8,9,18,13,5,3,1,2,12,17,15,16,
  6,17,4,9,8,1,7,5,14,12,2,3,18,15,10,13,11,16,
  16,11,17,6,14,12,8,18,1,2,15,3,7,10,9,13,5,4)
> yield = c(88.2,82.5,84.3,87.0,84.5,88.9,
  82.4,82.9,83.1,84.7,83.3,89.0,
  93.1,82.7,88.9,88.6,84.1,87.5,
  85.4,73.0,84.2,80.3,79.6,86.0,
  87.9,85.1,79.4,80.7,89.3,81.5,
  82.4,88.5,87.0,85.4,85.9,79.1,
  83.6,79.4,81.3,80.5,80.9,79.3,
  80.4,88.2,82.3,88.0,90.0,83.6,
  81.4,84.8,81.0,81.2,79.1,83.8,
  80.5,77.1,84.4,90.4,82.9,83.0,
  87.9,78.9,81.4,83.5,82.2,79.0,
  84.2,83.0,87.6,81.7,91.3,87.4)
```

```
> replicate = as.factor(replicate)
> block = as.factor(block)
> variety = as.factor(variety)
> Anova1 = aov( yield ~ replicate + variety +
  replicate:block)
> summary(Anova1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
replicate	3	101.3	33.75	10.206	3.37e-05 ***
variety	17	543.2	31.95	9.661	1.04e-09 ***
replicate:block	8	213.0	26.63	8.052	1.44e-06 ***
Residuals	43	142.2	3.31		

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # This Anova1 table gives not the correct Sum of Squares
> # for variety for the test.

> Anova.Test = aov( yield ~ block + variety)
> summary(Anova.Test)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
block	11	283.3	25.75	7.787	3.36e-07 ***
variety	17	574.2	33.78	10.213	4.37e-10 ***
Residuals	43	142.2	3.31		

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # In this Anova.Test table the correct test for
> # variety is given.

> # Intra-block analysis

> model.1 = lm( yield ~ replicate + block + variety)
> Anova2 = aov(model.1)
> summary(Anova2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
replicate	3	101.3	33.75	10.21	3.37e-05 ***
block	8	182.0	22.75	6.88	8.56e-06 ***
variety	17	574.2	33.78	10.21	4.37e-10 ***
Residuals	43	142.2	3.31		

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Note:** To get the estimate of a treatment mean with the Least Squares Method, we use the R-package lsmeans.

```
> library(lsmeans)
NOTE: A nesting structure was detected in the fitted model:
      block %in% replicate
> LS.rg = ref.grid(model.1)
NOTE: A nesting structure was detected in the fitted model:
      block %in% replicate
> LSM.V = lsmeans(LS.rg, "variety")
> LSM.V
```

variety	lsmean	SE	df	lower.CL	upper.CL
1	82.7	0.977	43	80.7	84.6
2	87.4	0.866	43	85.6	89.1
3	81.8	0.977	43	79.9	83.8
4	86.6	0.978	43	84.6	88.5
5	88.9	0.978	43	86.9	90.9
6	88.6	0.975	43	86.6	90.6
7	82.7	0.978	43	80.8	84.7
8	88.5	0.980	43	86.5	90.4
9	84.1	0.978	43	82.1	86.0
10	81.5	0.978	43	79.5	83.5
11	77.1	0.978	43	75.1	79.0
12	85.7	1.142	43	83.4	88.0
13	81.1	0.978	43	79.1	83.1
14	81.9	0.980	43	80.0	83.9
15	84.7	0.979	43	82.8	86.7
16	81.0	0.980	43	79.0	82.9
17	84.5	0.979	43	82.6	86.5
18	81.4	0.977	43	79.4	83.3

Results are averaged over the levels of: block, replicate  
Confidence level used: 0.95

```
> contrast(LSM.V, method = "pairwise")
contrast estimate    SE df t.ratio p.value
```

1 - 2	-4.7145	1.29	43	-3.647	0.0582
1 - 3	0.8155	1.36	43	0.601	1.0000
1 - 4	-3.9129	1.37	43	-2.859	0.3160
1 - 5	-6.2423	1.43	43	-4.372	0.0081
1 - 6	-5.9431	1.39	43	-4.268	0.0109
1 - 7	-0.0716	1.43	43	-0.050	1.0000
1 - 8	-5.8022	1.36	43	-4.274	0.0107
1 - 9	-1.3999	1.40	43	-1.002	0.9999
1 - 10	1.1679	1.43	43	0.815	1.0000
1 - 11	5.5979	1.43	43	3.918	0.0288
1 - 12	-3.0143	1.50	43	-2.013	0.8469
1 - 13	1.5574	1.40	43	1.115	0.9995
1 - 14	0.7246	1.40	43	0.519	1.0000
1 - 15	-2.0941	1.36	43	-1.545	0.9814
1 - 16	1.6794	1.36	43	1.238	0.9982
1 - 17	-1.8778	1.36	43	-1.384	0.9938
1 - 18	1.2888	1.39	43	0.927	1.0000
2 - 3	5.5300	1.29	43	4.276	0.0107
2 - 4	0.8016	1.36	43	0.590	1.0000
2 - 5	-1.5278	1.30	43	-1.177	0.9990
2 - 6	-1.2286	1.32	43	-0.928	1.0000
2 - 7	4.6429	1.30	43	3.560	0.0720
2 - 8	-1.0877	1.29	43	-0.841	1.0000
2 - 9	3.3146	1.33	43	2.498	0.5440

2 - 10	5.8825	1.33	43	4.423	0.0070
2 - 11	10.3124	1.33	43	7.766	<.0001
2 - 12	1.7002	1.40	43	1.212	0.9986
2 - 13	6.2719	1.35	43	4.631	0.0038
2 - 14	5.4392	1.33	43	4.091	0.0180
2 - 15	2.6204	1.26	43	2.074	0.8157
2 - 16	6.3940	1.32	43	4.830	0.0020
2 - 17	2.8367	1.30	43	2.190	0.7489
2 - 18	6.0033	1.29	43	4.644	0.0036
3 - 4	-4.7284	1.40	43	-3.387	0.1084
3 - 5	-7.0578	1.35	43	-5.210	0.0006
3 - 6	-6.7586	1.43	43	-4.725	0.0028
3 - 7	-0.8871	1.40	43	-0.635	1.0000
3 - 8	-6.6177	1.36	43	-4.880	0.0018
3 - 9	-2.2154	1.39	43	-1.594	0.9752
3 - 10	0.3524	1.43	43	0.247	1.0000
3 - 11	4.7824	1.43	43	3.334	0.1223
3 - 12	-3.8298	1.51	43	-2.543	0.5133
3 - 13	0.7419	1.36	43	0.547	1.0000
3 - 14	-0.0909	1.37	43	-0.066	1.0000
3 - 15	-2.9096	1.39	43	-2.094	0.8049
3 - 16	0.8639	1.40	43	0.618	1.0000
3 - 17	-2.6933	1.43	43	-1.884	0.9023
3 - 18	0.4733	1.36	43	0.349	1.0000
4 - 5	-2.3294	1.39	43	-1.676	0.9612
4 - 6	-2.0302	1.35	43	-1.498	0.9862
4 - 7	3.8413	1.36	43	2.835	0.3297
4 - 8	-1.8893	1.40	43	-1.351	0.9952
4 - 9	2.5131	1.36	43	1.853	0.9135
4 - 10	5.0809	1.36	43	3.749	0.0450
4 - 11	9.5108	1.39	43	6.842	<.0001
4 - 12	0.8987	1.55	43	0.581	1.0000
4 - 13	5.4703	1.36	43	4.033	0.0211
4 - 14	4.6376	1.36	43	3.415	0.1016

4 - 15	1.8188	1.44	43	1.267	0.9977
4 - 16	5.5924	1.40	43	4.001	0.0230
4 - 17	2.0351	1.40	43	1.457	0.9895
4 - 18	5.2017	1.43	43	3.642	0.0589
5 - 6	0.2992	1.40	43	0.214	1.0000
5 - 7	6.1707	1.36	43	4.553	0.0047
5 - 8	0.4401	1.36	43	0.324	1.0000
5 - 9	4.8424	1.35	43	3.577	0.0692
5 - 10	7.4102	1.36	43	5.463	0.0003
5 - 11	11.8402	1.43	43	8.282	<.0001
5 - 12	3.2280	1.52	43	2.130	0.7850
5 - 13	7.7997	1.36	43	5.756	0.0001
5 - 14	6.9669	1.40	43	4.977	0.0013
5 - 15	4.1482	1.40	43	2.971	0.2585
5 - 16	7.9217	1.44	43	5.511	0.0002
5 - 17	4.3645	1.44	43	3.040	0.2266
5 - 18	7.5311	1.39	43	5.431	0.0003
6 - 7	5.8715	1.39	43	4.235	0.0120
6 - 8	0.1409	1.37	43	0.103	1.0000
6 - 9	4.5432	1.39	43	3.261	0.1436
6 - 10	7.1110	1.36	43	5.240	0.0006
6 - 11	11.5410	1.35	43	8.525	<.0001
6 - 12	2.9288	1.50	43	1.956	0.8731
6 - 13	7.5005	1.43	43	5.256	0.0005
6 - 14	6.6677	1.36	43	4.917	0.0016
6 - 15	3.8490	1.40	43	2.758	0.3747
6 - 16	7.6226	1.39	43	5.483	0.0003
6 - 17	4.0653	1.35	43	3.001	0.2442
6 - 18	7.2319	1.43	43	5.060	0.0010
7 - 8	-5.7306	1.43	43	-4.007	0.0227
7 - 9	-1.3283	1.36	43	-0.979	0.9999
7 - 10	1.2395	1.35	43	0.915	1.0000
7 - 11	5.6695	1.36	43	4.180	0.0140
7 - 12	-2.9427	1.51	43	-1.954	0.8739

7 - 13	1.6290	1.39	43	1.172	0.9991
7 - 14	0.7962	1.36	43	0.587	1.0000
7 - 15	-2.0225	1.43	43	-1.410	0.9925
7 - 16	1.7510	1.43	43	1.225	0.9984
7 - 17	-1.8062	1.40	43	-1.293	0.9971
7 - 18	1.3604	1.40	43	0.974	0.9999
8 - 9	4.4024	1.36	43	3.246	0.1485
8 - 10	6.9702	1.40	43	4.992	0.0012
8 - 11	11.4001	1.44	43	7.931	<.0001
8 - 12	2.7880	1.55	43	1.796	0.9318
8 - 13	7.3596	1.39	43	5.291	0.0005
8 - 14	6.5269	1.44	43	4.531	0.0051
8 - 15	3.7081	1.36	43	2.733	0.3896
8 - 16	7.4817	1.43	43	5.214	0.0006
8 - 17	3.9244	1.40	43	2.802	0.3485
8 - 18	7.0910	1.35	43	5.238	0.0006
9 - 10	2.5678	1.39	43	1.850	0.9144
9 - 11	6.9978	1.40	43	5.008	0.0012
9 - 12	-1.6144	1.55	43	-1.038	0.9998
9 - 13	2.9573	1.36	43	2.182	0.7540
9 - 14	2.1245	1.43	43	1.485	0.9874
9 - 15	-0.6942	1.43	43	-0.486	1.0000
9 - 16	3.0793	1.44	43	2.144	0.7768
9 - 17	-0.4779	1.37	43	-0.349	1.0000
9 - 18	2.6887	1.36	43	1.982	0.8613
10 - 11	4.4300	1.36	43	3.264	0.1427
10 - 12	-4.1822	1.55	43	-2.699	0.4110
10 - 13	0.3895	1.36	43	0.287	1.0000
10 - 14	-0.4433	1.39	43	-0.318	1.0000
10 - 15	-3.2621	1.37	43	-2.384	0.6223
10 - 16	0.5115	1.40	43	0.365	1.0000
10 - 17	-3.0458	1.43	43	-2.130	0.7850
10 - 18	0.1209	1.39	43	0.087	1.0000
11 - 12	-8.6122	1.50	43	-5.754	0.0001



11 - 13	-4.0405	1.40	43	-2.892	0.2982
11 - 14	-4.8733	1.36	43	-3.594	0.0663
11 - 15	-7.6920	1.40	43	-5.502	0.0002
11 - 16	-3.9185	1.36	43	-2.888	0.3006
11 - 17	-7.4757	1.36	43	-5.511	0.0002
11 - 18	-4.3091	1.37	43	-3.147	0.1830
12 - 13	4.5717	1.55	43	2.942	0.2727
12 - 14	3.7389	1.45	43	2.572	0.4939
12 - 15	0.9202	1.50	43	0.613	1.0000
12 - 16	4.6937	1.45	43	3.231	0.1533
12 - 17	1.1365	1.45	43	0.782	1.0000
12 - 18	4.3031	1.55	43	2.779	0.3618
13 - 14	-0.8328	1.40	43	-0.596	1.0000
13 - 15	-3.6515	1.40	43	-2.614	0.4656
13 - 16	0.1221	1.37	43	0.089	1.0000
13 - 17	-3.4352	1.44	43	-2.393	0.6160
13 - 18	-0.2686	1.36	43	-0.198	1.0000
14 - 15	-2.8187	1.43	43	-1.967	0.8682
14 - 16	0.9548	1.36	43	0.704	1.0000
14 - 17	-2.6024	1.39	43	-1.873	0.9065
14 - 18	0.5642	1.44	43	0.393	1.0000
15 - 16	3.7736	1.36	43	2.774	0.3648
15 - 17	0.2163	1.39	43	0.155	1.0000
15 - 18	3.3829	1.35	43	2.497	0.5446
16 - 17	-3.5573	1.36	43	-2.625	0.4588
16 - 18	-0.3907	1.40	43	-0.279	1.0000
17 - 18	3.1666	1.40	43	2.265	0.7020

Results are averaged over the levels of: block, replicate

P value adjustment: tukey method for comparing a family of 18 estimates

**Note:** Interblock analysis with recovery of information.  
Model with replicate and variety is fixed and blocks in the replicates are random.  
We use the R-packages `lme4` and `lmerTest`.

```
> library(lme4)
> MODEL = lmer( yield ~ replicate + variety + (1|block))
> MODEL
Linear mixed model fit by REML ['lmerMod']
Formula: yield ~ replicate + variety + (1 | block)
REML criterion at convergence: 254.338
Random effects:
Groups      Name          Std.Dev.
block      (Intercept)  2.249
Residual                    1.818
Number of obs: 72, groups:  block, 12
```

Fixed Effects:

(Intercept)	replicate2	replicate3	replicate4	variety2	variety3
84.3773	-2.4114	-3.0725	-2.0948	4.8983	-0.6510
variety4	variety5	variety6	variety7	variety8	variety9
3.9886	6.4540	5.8929	0.4159	5.7099	1.5954
variety10	variety11	variety12	variety13	variety14	variety15
-0.9867	-5.3017	3.3040	-1.2680	-0.5179	2.1998
variety16	variety17	variety18			
-1.4235	2.0399	-1.0348			

```
> library(lmerTest)
```

```
> MODEL.1 = lmer( yield ~replicate + variety + (1|block))
```

```
> lsm = ls_means(MODEL.1)
```

```
> lsm
```

Least Squares Means table:

	Estimate	Std. Error	df	t value	lower	upper	Pr(> t )	
replicate1	85.7837	1.3693	7.6	62.648	82.5966	88.9707	1.379e-11	***
replicate2	83.3722	1.3671	7.5	60.984	80.1865	86.5580	1.923e-11	***
replicate3	82.7111	1.3671	7.5	60.501	79.5254	85.8968	2.041e-11	***
replicate4	83.6889	1.3671	7.5	61.216	80.5032	86.8746	1.868e-11	***
variety1	82.4826	1.1656	38.2	70.765	80.1235	84.8417	< 2.2e-16	***
variety2	87.3809	1.0769	33.1	81.141	85.1901	89.5717	< 2.2e-16	***
variety3	81.8316	1.1655	38.2	70.211	79.4726	84.1906	< 2.2e-16	***
variety4	86.4711	1.1660	38.2	74.163	84.1113	88.8310	< 2.2e-16	***
variety5	88.9366	1.1658	38.2	76.287	86.5770	91.2962	< 2.2e-16	***
variety6	88.3755	1.1643	38.2	75.902	86.0189	90.7321	< 2.2e-16	***
variety7	82.8985	1.1657	38.2	71.115	80.5391	85.2579	< 2.2e-16	***
variety8	88.1924	1.1673	38.2	75.553	85.8299	90.5550	< 2.2e-16	***
variety9	84.0780	1.1659	38.2	72.116	81.7183	86.4377	< 2.2e-16	***
variety10	81.4959	1.1657	38.2	69.910	79.1365	83.8553	< 2.2e-16	***
variety11	77.1809	1.1662	38.2	66.183	74.8206	79.5412	< 2.2e-16	***
variety12	85.7866	1.3031	44.2	65.833	83.1608	88.4124	< 2.2e-16	***
variety13	81.2145	1.1660	38.2	69.654	78.8546	83.5745	< 2.2e-16	***
variety14	81.9647	1.1672	38.2	70.225	79.6023	84.3270	< 2.2e-16	***
variety15	84.6823	1.1664	38.2	72.600	82.3215	87.0431	< 2.2e-16	***
variety16	81.0591	1.1671	38.2	69.452	78.6968	83.4213	< 2.2e-16	***
variety17	84.5225	1.1665	38.2	72.456	82.1615	86.8836	< 2.2e-16	***
variety18	81.4478	1.1654	38.2	69.890	79.0891	83.8065	< 2.2e-16	***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Confidence level: 95%

Degrees of freedom method: Satterthwaite

## Author details

L. Rob Verdooren<sup>1\*</sup> and Dieter Rasch<sup>2,3\*</sup>

1 Experimental Design and their Analysis at the Agricultural University of Wageningen, The Netherlands

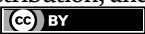
2 Statistical Department of the “Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere” [Research Institute for the Biology of Farm Animals] in Dummerstorf (Near Rostock), Germany

3 Statistics at the Agricultural University of Wageningen, The Netherlands

\*Address all correspondence to: [l.r.verdooren@outlook.com](mailto:l.r.verdooren@outlook.com) and [d\\_rasch@t-online.de](mailto:d_rasch@t-online.de)

## IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Fisher RA. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*. 1926;**33**: 503-513
- [2] Kuehl RO. *Statistical Principles of Research Design and Analysis*. Belmont, California: Duxbury Press; 1994
- [3] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 1955; **50**:1112-1118
- [4] Dunnett CW. New tables for multiple comparisons with a control. *Biometrics*. 1964;**20**:482-491
- [5] Fisher RA. An examination of the different possible solutions of a problem in incomplete blocks, *annals of Eugenics*. 1940;**10**:52-75
- [6] Bose RC. A note on the resolvability of balanced incomplete block designs. *Sankya, A*. 1942;**6**:105-110
- [7] Rasch D, Teuscher F, Verdooren LR. A conjecture about BIBDs. *Communications in Statistics-Simulation and Computation*. 2014;**43**:1526-1537
- [8] Rasch D, Herrendörfer G. *Statistische Versuchsplanung*. VEB Deutscher Verlag der Wissenschaften. 1982
- [9] Rasch D, Herrendörfer G. *Experimental design: Sample size determination and block designs*. D. Reidel Publishing Company; 1986
- [10] Rasch D, Verdooren R. *Angewandte Statistik mit R für Agrarwissenschaften*. Heidelberg: Springer; 2023
- [11] Fisher RA, Yates F. *Statistical Tables for Biological, Agricultural and Medical Research*. 1st ed. Edinburgh, London, UK: (6th edition 1963): Oliver and Boyd. 1938
- [12] Rasch D, Herrendörfer G. *Statystyczne planowanie doświadczeń*. Warszawa: Wydawnictwo Naukowe PWN; 1991
- [13] Rasch DAMK. Statistical experimental design. In: Ocerin JMC, editor. *Design of experiments and statistical education in agriculture*. Proc. 3rd EU HARMA Meeting. Cordoba; 1996. pp. 1-95
- [14] Rasch D. Determination of the Size of an Experiment, *MODA 5—Advances in Model-Oriented Data*. Dordrecht: D. Reidel Publishing Company; 1998
- [15] Teuscher F, Rasch D. On the existence and generation of non-trivial BIBDs. In: *Accepted for Communications in Statistics - Simulation and Computation*. 2024
- [16] Cochran WG, Cox GM. *Experimental Designs*. Second ed. New York, London, Sydney: John Wiley & Sons, Inc.; 1957
- [17] Patterson HD, Williams ER. A new class of resolvable incomplete block designs. *Biometrika*. 1976;**63**:83-92
- [18] Patterson HD, Thompson R. Maximum likelihood estimation of components of variance. In: Corsten LCA, Postelnicu T, editors. *Proceedings of the 8th International Biometric Conference*, Constanta, Editura Academiei Republic Socialiste Romania, Bukarest, 384 pages, 1975. pp. 197-207
- [19] John JA, Williams ER. *Cyclic and Computer Generated Designs*. 2nd ed.

London, Glasgow, Weinstein, New York, Tokyo, Melbourne, Madras: Chapman & Hall; 1995

[20] John JA, Wolock FW, David HA. Cyclic Designs. Applied Mathematics Series. National Bureau of Standards; 1972. p. 62

[21] Lamacraft RR, Hall WB. Tables of incomplete cyclic block designs:  $r = k$ . Australian Journal of Statistics. 1982;**24**: 350-360

[22] CycDesign. A Package for the Computer Generation of Experimental Designs. 2014. Available from: <http://www.vsnl.co.uk/software/cycdesign/>

[23] Williams ER. Iterative analysis of generalized lattice designs. Australian and New Zealand Journal of Statistics. 1977;**19**:39-42

[24] Patterson HD, Williams ER, Hunter EA. Block designs for variety trials. Journal of Agricultural Science. Cambridge. 1978;**90**:395-400

[25] Patterson HD, Silvey V. Statutory and recommended list trials of crop varieties in the United Kingdom (with discussion). Journal of the Royal Statistical Society, Series A. 1980;**143**: 219-252

[26] Cullis BR, Smith AB, Coombes NE. On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological and Environmental Statistics. 2006;**11**(4): 381-393

[27] Williams ER, John JA, Whitaker D. Construction of more flexible and efficient p-rep designs. Australian and New Zealand Journal of Statistics. 2014; **56**:89-96

[28] Verdooren LR. Use of alpha-designs in oil palm breeding trials. American Journal of Theoretical and Applied Statistics. 2019;**8**:136-143. DOI: 10.11648/j.ijepe.20190804.12

[29] Verdooren R, Soh AC, Mayes S, Roberts J. Chapter 12 field experimentation. In: Soh AC, Mayes S, Roberts JA, editors. Oil Palm Breeding, Genetics and Genomics. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group; 2017



# Perspective Chapter: Using Effect Sizes to Study the Survival Difference between Two Groups

*Huan Wang, Li Sheng and Dechang Chen*

## Abstract

Statistical tests are often used to detect the difference in survival between two groups. Small  $p$ -values, say less than 0.05, are commonly used to declare significant differences. The problem is that  $p$ -values do not tell how much the differences are. An alternative is to use effect sizes to detect the difference in survival between two groups. Effect sizes provide numerical numbers to quantify the differences. In this study, we reviewed the effect size  $ES_G$  that was developed recently by Wang, H., Chen, D., Pan, Q. et al. The effect size  $ES_G$  is not only unaffected by the change in sample sizes but also applicable no matter if hazards are proportional. We presented some applications of the effect size in comparing different groups of patients with prostate cancer. The results showed that the effect size  $ES_G$  performed well in detecting and quantifying the difference in survival between two groups.

**Keywords:** prostate cancer, survival analysis, effect size,  $p$ -values, sample size

## 1. Introduction

To assess differences in survival between two groups (populations), the most common practice is to perform a hypothesis test and report the  $p$ -value. A small  $p$ -value indicates a statistically significant difference in survival between the two groups, while a large  $p$ -value could indicate the opposite. Thus,  $p$ -values do reflect differences in survival to some extent.

However, because the  $p$ -value is susceptible to variations in sample sizes, it is not an adequate measure of the difference in survival. When performing a test, the value of the test statistic and  $p$ -value are calculated using samples. If the sample size is small, a large insignificant  $p$ -value may be produced; if the sample size is large, a small significant  $p$ -value may be obtained. Thus, different sample sizes can yield inconsistent conclusions. The  $p$ -values can be calibrated according to sample sizes. But in general,  $p$ -values are used without regard to sample sizes, and such  $p$ -values are not appropriate measures of survival differences. For the same reason, the value of the test statistic is not suitable either for measuring differences in survival. A natural

question is: Which measure, other than the test statistic value and  $p$ -value, better describes the survival difference between two groups?

The effect size may be a good choice for such a measure [1, 2]. An effect size is a quantitative measure of the magnitude of a difference between two groups and is not affected by changes in the sample size. An effect size differs from a  $p$ -value in that the former is a direct measure of the strength of the effect (difference) while the latter is a measure of how likely the observed difference is due to chance [3].

In the absence of censoring, there are extensive studies about effect sizes and many effect sizes are available, e.g., correlation coefficient, odds ratio, relative risk, and Cohen's  $d$  [4], etc. However, there are not many studies on effect sizes assessing the difference in survival for time-to-event data. The theory behind effect sizes with the presence of censoring is more complicated than that without censoring. It is not trivial at all to obtain effect sizes in cases where censoring occurs.

Below we briefly review the effect sizes associated with censoring. The hazard ratio is one commonly used effect size for right-censored data. Hazard ratios come from the Cox modeling [5] based on the assumption of proportional hazards. If the assumption of proportional hazards is violated, a hazard ratio can fail to capture the relative difference in survival between two groups [6]. The average hazard ratio (AHR) [7] and the restricted mean survival time (RMST) [8] are two types of estimates of effect sizes without the assumption of proportional hazards. However, it is not easy to interpret AHR because of its complex definition. The use of RMST appears to be limited by the difficulty in selecting the appropriate time period for calculating estimates. Recently, Wang et al. [9] proposed to use the weighted differences in hazards as effect sizes for studying the survival difference between two groups. Their proposed effect sizes can be applied with or without the proportional hazards assumption. In this study, we investigate survival differences between two groups by using  $ES_G$ , one of the effect sizes in [9]. Advantages of using  $ES_G$  include its good performance and ease of computation and interpretation.

This study is based on the work in [9] and [10]. It is organized in the following way. Section 2 reviews the effect size  $ES_G$ , its estimate, its properties, and its partition rule. Section 3 illustrates some applications of the effect size in comparing different cohorts of patients with prostate cancer. We conclude in Section 4.

## 2. The effect size and its estimate

### 2.1 Definition of the effect size

Suppose we would like to compare the survival difference between Group 1 and Group 2. For the Group  $i$  ( $i = 1, 2$ ), we use the following notations:

- $\lambda_i(t)$  – the hazard function
- $S_i(t)$  – the survival function of the failure time
- $S_i^*(t)$  – the survival function of the censoring time
- $\pi_i(t)$  – set to be  $S_i(t)S_i^*(t)$ , which is the probability of being at risk at time  $t$

We have the following effect size [9].



$$ES_G = \int_0^{\infty} \pi_1(t)\pi_2(t)(\lambda_1(t) - \lambda_2(t)) dt. \quad (1)$$

The effect size  $ES_G$  is derived on the basis of the Gehan-Wilcoxon test statistic [11]. An interpretation of the effect size comes directly from the formula (1). In fact, the formula states that  $ES_G$  is a weighted difference between two hazard functions  $\lambda_1(t)$  and  $\lambda_2(t)$  with the weight equal to  $\pi_1(t)\pi_2(t) = S_1^*(t)S_2^*(t)S_1(t)S_2(t)$ . The term “weighted” is used here because of the integration in (1). It is seen that the weight at time  $t$ , i.e.,  $S_1^*(t)S_2^*(t)S_1(t)S_2(t)$  is the probability that the observed times (either failure time or censoring time) of both groups exceed  $t$ .

It is important to note that the effect size  $ES_G$  represents the weighted difference in hazards for the largest time period of study for which censoring is possible for both groups. Therefore,  $ES_G$  can be employed to compare the two groups for the time period of study which is designed to compare the two groups. For example, consider the scenario where the study is terminated at time  $\tilde{t}$ . Since  $S_1^*(t) = S_2^*(t) = 0$  for  $t > \tilde{t}$ , so that  $\pi_1(t) = \pi_2(t) = 0$  for  $t > \tilde{t}$ , we have  $ES_G = \int_0^{\tilde{t}} \pi_1(t)\pi_2(t)(\lambda_1(t) - \lambda_2(t))dt$ . Then it is seen that  $ES_G$  only computes the weighted difference before time  $\tilde{t}$  and thus the effect size  $ES_G$  only compares the two groups before time  $\tilde{t}$ .

If we switch the positions of  $\lambda_1(t)$  and  $\lambda_2(t)$  in (1), then the resulting effect size will be negative of the effect size defined in (1). Because of this, it is often convenient to talk about the absolute value of the effect size, that is,  $|ES_G|$ . Therefore, using  $\lambda_1(t) - \lambda_2(t)$  or  $\lambda_2(t) - \lambda_1(t)$  is not of main concern.

In practice, it is impossible to compute  $ES_G$  in (1). However, it is easy to compute an estimate of the effect size using sample data.

## 2.2 Estimate of the effect size

Suppose there are two samples of survival data from the two groups under study. Sample 1 from Group 1 has a size  $n_1$ , and Sample 2 from Group 2 has a size  $n_2$ . Let  $n = n_1 + n_2$ . Combine the two samples and let  $t_1, \dots, t_J$  be the distinct failure times in increasing order from the pooled sample. For any  $j(1 \leq j \leq J)$ , we use the following notations:

- $D_{1j}$  – the number of subjects who failed at  $t_j$  in sample 1
- $D_{2j}$  – the number of subjects who failed at  $t_j$  in sample 2
- $Y_{1j}$  – the number of subjects who were at risk at  $t_j$  in sample 1
- $Y_{2j}$  – the number of subjects who were at risk at  $t_j$  in sample 2
- $Y_j$  – the total number of subjects who were at risk at  $t_j$  in both samples

Define [9].

$$\hat{ES}_G \equiv \frac{n}{n_1 n_2} \sum_{j=1}^J \frac{Y_j}{n} \left( \frac{D_{1j} Y_{2j}}{Y_j} - \frac{D_{2j} Y_{1j}}{Y_j} \right). \quad (2)$$

Then  $\hat{ES}_G$  is an estimate of the effect size  $ES_G$ .

### 2.3 Properties of the effect size

The effect size  $ES_G$  has many nice properties. Below we list some of them [9].

- a.  $ES_G$  is equal to the probability that a randomly selected subject from Group 2 can be observed to live longer than a randomly selected subject from Group 1 minus the probability that a randomly selected subject from Group 1 can be observed to live longer than a randomly selected subject from Group 2.
- b.  $ES_G$  lies inside the interval  $[-1, 1]$ .
- c. A positive (negative) effect size  $ES_G$  implies a “higher” (“lower”) hazard in Group 1 than in Group 2.
- d. The effect size depends on the censoring survival functions, i.e.,  $S_1^*(t)$  and  $S_2^*(t)$ .
- e. If the assumption of proportional hazards holds, i.e., the hazard ratio  $\frac{\lambda_1(t)}{\lambda_2(t)}$  equals constant  $r$ , then  $ES_G \approx \frac{r-1}{r+1}$  for light censoring in both groups.
- f. If the integrand in (1) is absolutely integrable,  $\hat{ES}_G$  converges (in probability) to  $ES_G$  as  $n \rightarrow \infty$ .

Property (a) provides another interpretation of the effect size  $ES_G$ . From property (a), we see that  $ES_G$  does not directly compare failure times between groups but rather compares observed times. Property (b), coming directly from (a), gives the range of the effect size. So we know  $|ES_G|$  ranges from 0 to 1. Property (c) explains the sign of the effect size. Property (d), following from formula (1), emphasizes the fact that sizes of censoring survival functions impact the magnitudes of the effect size. If light censoring occurs for both groups, i.e.,  $S_1^*$  and  $S_2^*$  are close to 1, the effect size and hazard ratio depend on each other (approximately) and the relationship is described by property (e). Property (f) states that the effect size and its estimate will be sufficiently close for large samples.

### 2.4 Partition of values of the effect size

The effect size  $ES_G$  quantifies the survival difference between the two groups. In many cases, a single value of the effect size is not enough and we would like to know if an effect size is sufficiently large to be (practically) meaningful. For instance, in a clinical setting, one may need to evaluate the clinical meaningfulness of the magnitude of an effect size. Therefore, we need certain rules to determine if an effect size is small, medium, or large. This involves partitioning the values of the effect size.

Assume that the failure times in the two groups are exponentially distributed. Also, assume that the censoring times in the two groups are exponentially distributed. Then using the widely used rule of thumb on the magnitude of Cohen's  $d$ , we have **Table 1** [9] which shows a list of small, medium, and large effect sizes for selected censoring rates  $CR_i$  in group  $i$ . The rate  $CR_i$  can be estimated by the corresponding observed censoring rate. When an estimate of the effect size is available, we can use the table to determine if the effect size is small, medium, or large. Here are the steps. First, locate

CR <sub>1</sub>	Norms	CR <sub>2</sub>									
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
0%	Small	0.13	0.12	0.11	0.10	0.09	0.08	0.07	0.06	0.04	0.02
	Medium	0.31	0.29	0.27	0.24	0.22	0.19	0.16	0.12	0.09	0.04
	Large	0.47	0.44	0.40	0.36	0.32	0.27	0.22	0.17	0.12	0.06
10%	Small	0.12	0.11	0.11	0.10	0.09	0.08	0.07	0.05	0.04	0.02
	Medium	0.30	0.28	0.26	0.24	0.21	0.18	0.15	0.12	0.08	0.04
	Large	0.46	0.43	0.39	0.35	0.31	0.27	0.22	0.17	0.12	0.06
20%	Small	0.12	0.11	0.10	0.09	0.09	0.08	0.07	0.05	0.04	0.02
	Medium	0.29	0.27	0.25	0.23	0.20	0.18	0.15	0.12	0.08	0.04
	Large	0.44	0.41	0.38	0.34	0.30	0.26	0.22	0.17	0.12	0.06
30%	Small	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04	0.02
	Medium	0.27	0.25	0.24	0.22	0.20	0.17	0.15	0.12	0.08	0.04
	Large	0.42	0.40	0.36	0.33	0.29	0.26	0.21	0.17	0.12	0.06
40%	Small	0.10	0.09	0.09	0.08	0.08	0.07	0.06	0.05	0.04	0.02
	Medium	0.25	0.24	0.22	0.21	0.19	0.16	0.14	0.11	0.08	0.04
	Large	0.40	0.38	0.35	0.32	0.28	0.25	0.21	0.16	0.11	0.06
50%	Small	0.09	0.09	0.08	0.08	0.07	0.06	0.06	0.05	0.03	0.02
	Medium	0.23	0.22	0.21	0.19	0.17	0.16	0.13	0.11	0.08	0.04
	Large	0.37	0.35	0.33	0.30	0.27	0.24	0.20	0.16	0.11	0.06
60%	Small	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.04	0.03	0.02
	Medium	0.20	0.19	0.18	0.17	0.16	0.14	0.12	0.10	0.07	0.04
	Large	0.34	0.32	0.30	0.28	0.25	0.22	0.19	0.15	0.11	0.06
70%	Small	0.06	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.03	0.02
	Medium	0.17	0.17	0.16	0.15	0.14	0.13	0.11	0.09	0.07	0.04
	Large	0.29	0.28	0.26	0.24	0.22	0.20	0.17	0.14	0.10	0.06
80%	Small	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.02
	Medium	0.13	0.13	0.12	0.12	0.11	0.10	0.09	0.08	0.06	0.04
	Large	0.23	0.22	0.21	0.20	0.19	0.17	0.15	0.13	0.09	0.05
90%	Small	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01
	Medium	0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.05	0.03
	Large	0.14	0.14	0.13	0.13	0.12	0.11	0.11	0.09	0.07	0.05

**Table 1.**  
 Small, medium, and large effect sizes  $ES_G$ . The censoring rate in group  $i$  is denoted by  $CR_i$  ( $i = 1, 2$ ).

the triplet of numbers according to the censoring rates. Then use the midpoints of adjacent numbers in the triplet to construct three consecutive and disjoint intervals for small, medium, and large effect categories. And finally, the decision is made by checking which interval contains the effect size. For instance, for  $CR_1 = 10\%$  and

$CR_2 = 20\%$ , the triplet consists of 0.11, 0.26, 0.39. Using the midpoint 0.19 of 0.11 and 0.26 and midpoint 0.33 of 0.26 and 0.39, we construct three consecutive and disjoint intervals:  $[0, 0.19)$ ,  $[0.19, 0.33)$ , and  $[0.33, 1]$ . Then our rule of thumb is: for  $CR_1 = 10\%$  and  $CR_2 = 20\%$ , the effect size  $ES_G$  is small if  $|ES_G| \in [0, 0.19)$ , medium if  $|ES_G| \in [0.19, 0.33)$ , and large if  $|ES_G| \in [0.33, 1]$ . Therefore, if an estimate  $\hat{ES}_G = 0.45$ , we can say that the effect size  $ES_G$  is large.

**Table 1** clearly shows that for a given effect size, its category (small, medium, or large) depends on the censoring rates  $CR_i$ . Two effect sizes with the same numerical number can have two different categories (e.g., one is small and the other is large) because of the different censoring rates. And two effect sizes with different numerical numbers can have the same category. Therefore, using the only numerical value of an effect size, one cannot determine if this effect size is small, medium, or large. If two comparisons have comparable censoring rates, one could compare two effect sizes by using only their numerical values.

Note that **Table 1** is obtained under the assumption that both failure and censoring times follow exponential distributions. This assumption may be violated in practice. In cases where this assumption does not hold, the rule resulting from the table serves as a simple and useful reference.

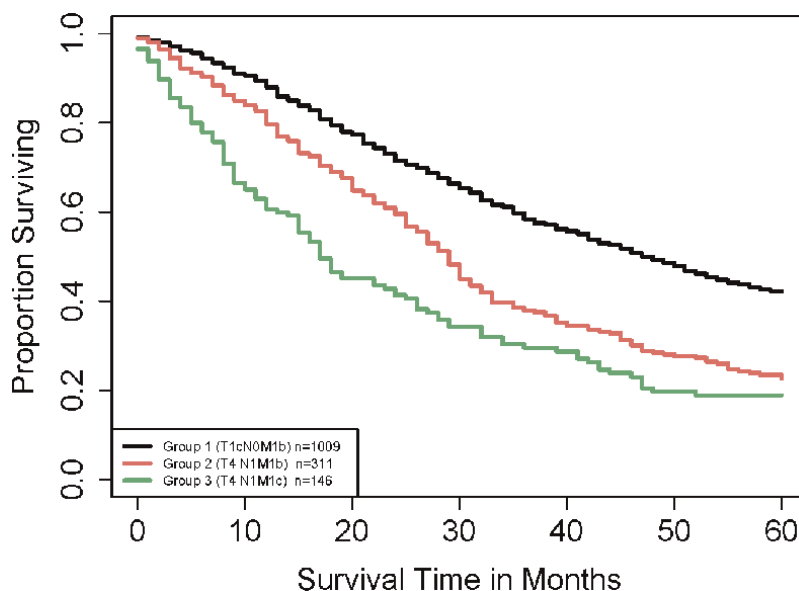
### 3. Examples

In this section, we present some applications of the effect size  $ES_G$  in comparing survival times of patients with prostate cancer. Disease-specific survival data with a primary diagnosis of prostate cancer during 2013–2015 were obtained from 17 databases of the surveillance, epidemiology, and end results Program (SEER) of the National Cancer Institute [12]. The years of diagnosis were chosen to ensure at least 5 years of follow-up and suitable sample sizes. The primary tumor (9 levels: T1a, T1b, T1c, T2a, T2b, T2c, T3a, T3b, T4), regional lymph nodes (2 levels: N0 and N1), and distant metastasis (4 levels: M0, M1a, M1b, M1c) were considered with definitions according to the AJCC Cancer Staging Manual, 7th edition [13]. Combinations of the primary tumor, regional lymph nodes, and distant metastasis are used to define groups in this study. For instance, T1aN0M0 defines a group of survival times for the patients whose tumor size is T1a, lymph node status is N0, and distant metastasis status is M0. For each possible group, the corresponding SEER dataset represents a sample.

#### 3.1 Example 1

This example illustrates how to use  $ES_G$  to examine differences in survival between groups. We consider three groups: Group 1, Group 2, and Group 3 defined by T1cN0M1b, T4N1M1b, and T4N1M1c, respectively. The SEER data provides us with three samples for Groups 1, 2, and 3, with sample sizes of 1009, 311, and 146, respectively. **Figure 1** shows the Kaplan–Meier [14] curves based on the three samples. This figure clearly indicates that the difference in survival between Group 1 and Group 2 is smaller than that between Group 1 and Group 3.

Calculation shows that  $\hat{ES}_G$  between Group 1 and Group 2 is  $-0.207$  and  $\hat{ES}_G$  between Group 1 and Group 3 is  $-0.374$ . Since the absolute values of  $-0.207$  are smaller than that of  $-0.374$ , assuming Group 2 and Group 3 have a similar censoring



**Figure 1.**  
 Kaplan-Meier curves of T1cN0M1b, T4N1M1b, T4N1M1c in Example 1.

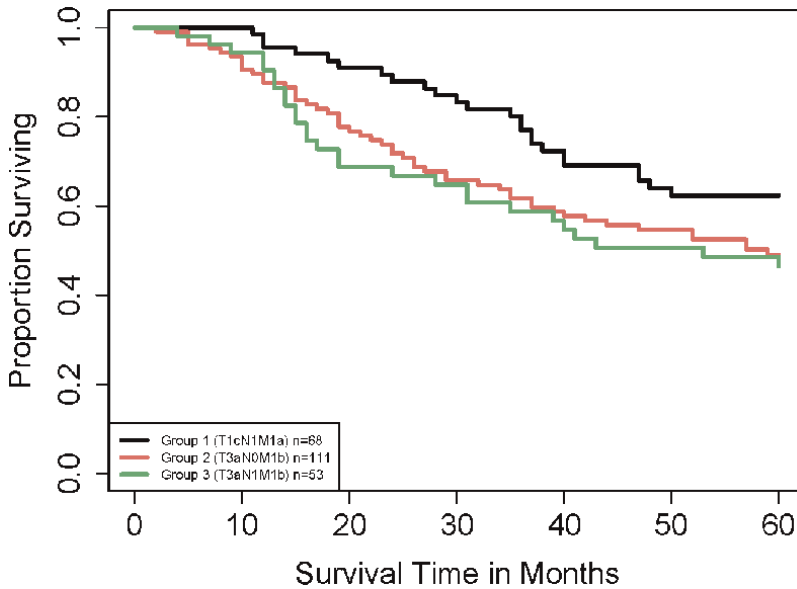
rate, we conclude that the survival difference between Group 1 and Group 2 is smaller than that between Group 1 and Group 3, which is consistent with the observation in **Figure 1**. Furthermore, **Table 1** can be used to give a stronger comparison. Since the estimated censoring rates in Groups 1, 2, and 3 are 43.7%, 27.3%, and 21.2%, respectively, it follows from **Table 1** that  $ES_G$  between Group 1 and Group 2 is medium and  $ES_G$  between Group 1 and Group 3 is large.

On the other hand, we could use statistical tests to examine the differences between groups. For instance, the  $p$ -values of the Gehan-Wilcoxon test between Group 1 and Group 2 and between Group 1 and Group 3 are, respectively,  $4.8 \times 10^{-11}$  and  $9.9 \times 10^{-22}$ . These two  $p$ -values show that both the differences in survival between Group 1 and Group 2 and between Group 1 and Group 3 are significant. However, it is hard for us to use the  $p$ -values to imagine how much the differences are without looking at **Figure 1**. Furthermore, since  $9.9 \times 10^{-22}$  is smaller  $4.8 \times 10^{-11}$ , we tend to conclude that the survival difference between Group 1 and Group 3 is larger than that between Group 1 and Group 2. But, it is hard for us to imagine the discrepancy between the two differences without looking at the survival curves.

This example demonstrates that even though  $p$ -values and effect sizes can be used to compare groups,  $p$ -values are in general less informative than effect sizes.

### 3.2 Example 2

As shown in **Example 1**, we compute values of  $ES_G$  (and censoring rates) and then use them to differentiate differences between groups. However,  $p$ -values may not be sufficient for us to do so, as illustrated in this example. Similar to **Example 1**, we consider three groups: Group 1, Group 2, and Group 3 defined by T1cN1M1a, T3aN0M1b, and T3aN1M1b, respectively. Note that these groups are different from those in **Example 1**. The SEER samples for Groups 1, 2, and 3 have sizes of 68, 111, and



**Figure 2.**  
Kaplan–Meier curves of  $T_{1c}N_{1M1a}$ ,  $T_{3a}N_{0M1b}$ ,  $T_{3a}N_{1M1b}$  in Example 2.

53, respectively. **Figure 2** shows the Kaplan–Meier curves of the three samples. This figure indicates that the difference in survival between Group 1 and Group 2 is smaller than that between Group 1 and Group 3.

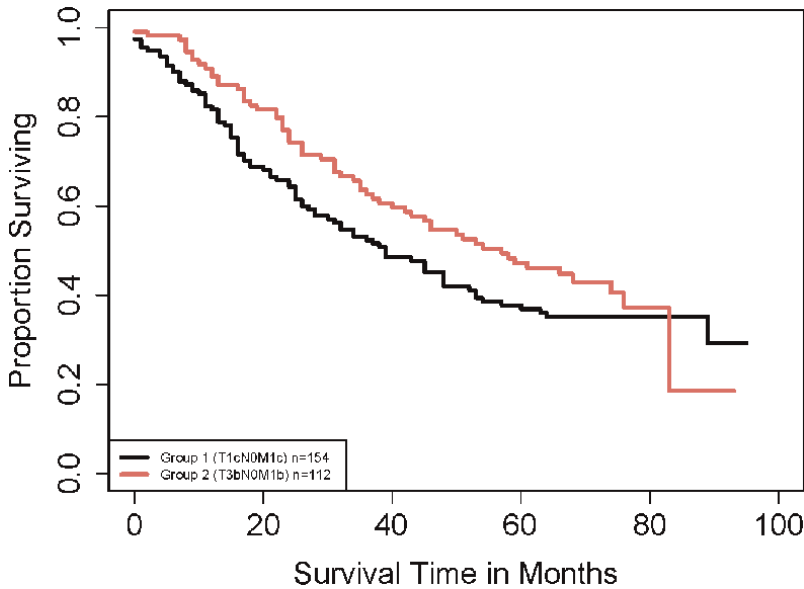
Calculation shows that  $ES_G$  estimates between Group 1 and Group 2 and between group 1 and group 3 are  $-0.168$  and  $-0.198$ , respectively. The estimated censoring rates in Groups 1, 2, and 3 are 58.8%, 48.6%, and 47.2%, respectively, so, from **Table 1**,  $ES_G$  between Group 1 and Group 2 is medium and  $ES_G$  between group 1 and group 3 is large. Thus, the difference in survival between Group 1 and Group 2 is smaller than that between Group 1 and Group 3, which is consistent with the observation in **Figure 2**.

If using the Gehan-Wilcoxon test to examine the differences between groups, the  $p$ -values of the test between Group 1 and Group 2 and between Group 1 and Group 3 are, respectively, 0.0259 and 0.0271. Since 0.0259 is smaller than 0.0271, with our common rule that a smaller  $p$ -value shows more significance, we would conclude that the survival difference between Group 1 and Group 2 is bigger than that between Group 1 and Group 3. Unfortunately, this conclusion contradicts our observation in **Figure 2**.

This example demonstrates a) a smaller  $p$ -value does not always mean a more significant result (See more related simulation results in [9].); and b) effect sizes can differentiate differences between groups when  $p$ -values fail to do so.

### 3.3 Example 3

The proposed effect sizes can be applied no matter if hazards are proportional. Here we present one example with non-proportional hazards, which  $ES_G$  can be applied to study the difference between two groups, while the hazard ratio approach fails to do so. We consider the following two groups: Group 1 and Group 2 defined to



**Figure 3.**  
*Kaplan–Meier curves of T1cN0M1c and T3bN0M1b in Example 3.*

be T1cN0M1c and T3bN0M1b, respectively. The samples for Groups 1 and 2 have sizes of 154 and 112, respectively. **Figure 3** shows the Kaplan–Meier curves of the two samples. This figure indicates a clear difference in survival between Group 1 and Group 2.

The Grambsch–Therneau test [15] for examining the proportional hazard assumption gives a  $p$ -value of 0.012, suggesting that the ratio of hazard rates would depend on time and thus would not be an appropriate effect size. Regardless of the violation of the proportional hazards assumption, the use of the Cox proportional hazards model would provide an estimated hazard ratio (Group 1 over Group 2) of 1.264 with a wide confidence interval (CI) (95% CI: 0.913 to 1.751). These estimates are not very informative when assessing if the two groups differ in survival. Therefore, the hazard ratio approach should not be used to examine the survival difference between Group 1 and Group 2.

In comparison, the  $p$ -value of the Gehan–Wilcoxon test is 0.026, which shows a significant difference in survival between Group 1 and Group 2. With effect sizes, we have  $\hat{ES}_G = 0.14$  (95% CI: 0.02 to 0.26, a bootstrap CI [16] based on 100,000 bootstrap samples). Note that the CI does not contain 0, so the two groups differ in survival. Furthermore, the positive effect size indicates that Group 1 has a shorter survival than Group 2. Since the censoring rates in Group 1 and Group 2 are 42.2% and 44.6%, respectively, from **Table 1**, we see that there is a medium effect size between the two groups.

This example demonstrates an application of the effect size  $ES_G$  to measure the difference in survival between two groups where the proportional hazards assumption does not hold. With non-proportional hazards, the traditional hazard ratio in general can not serve as an effect size.

As shown above, the effect size  $ES_G$  has direct applications in practice. It is particularly useful when repeated comparisons of survival differences are required.

For instance, when integrating additional variables/factors into the TNM staging system for cancer, assessing the survival difference is needed for many pairs of groups and two groups can be merged if the effect size assessing their survival difference is small. See [17, 18] for studies that applied the effect size  $ES_G$  and the Ensemble Algorithm for Clustering Cancer Data [19–24] to update and improve the staging system for thyroid and ovarian cancers.

#### **4. Conclusion**

We have reviewed the effect size  $ES_G$  and its estimate for comparing the survival difference between two groups. The effect size  $ES_G$  quantifies the survival difference between two groups over the time period of investigation. One can claim a small or big difference in survival according to the effect size and the censoring rates. This is different from checking the  $p$ -value of a statistical test, which may not provide any insight into the size of the survival difference and could cause misunderstanding.  $ES_G$  can be applied no matter if hazards are proportional. This is different from the use of the hazard ratio, the traditionally used effect size. Applications of hazard ratios require the assumption of proportional hazards. With non-proportional hazards, hazard ratios can fail to detect and quantify the difference between two groups. We have also used  $ES_G$  to compare different groups of patients with prostate cancer. The results have shown that  $ES_G$  is a promising effect size for studying differences in survival between two groups.

There is a need for further research and refinement for our study, which currently focuses on unadjusted comparison. Our future research endeavors will delve into the exploration of incorporating methodologies that can be used to adjust important variables, such as matching, stratification, and inverse probability weighting. This future work will broaden the scope of scenarios in which  $ES_G$  can be effectively utilized.

#### **Acknowledgements**

This study is based upon work supported by John P. Murtha Cancer Center Research Program under the Grant No. 64349-MCC Comprehensive Research.

#### **Disclaimer**

The contents, views, or opinions expressed in this publication or presentation are those of the authors and do not necessarily reflect the official policy or position of the Uniformed Services University of the Health Sciences, the Department of Defense (DoD), or Departments of the Army, Navy, or Air Force, or the U.S. Food and Drug Administration. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.



## Author details

Huan Wang<sup>1</sup>, Li Sheng<sup>2</sup> and Dechang Chen<sup>3\*</sup>

1 Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, USA


2 Department of Mathematics, Drexel University, Philadelphia, USA

3 Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, USA

\*Address all correspondence to: [dechang.chen@usuhs.edu](mailto:dechang.chen@usuhs.edu)

## IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;**567**:305-307
- [2] Wasserstein R, Schirm A, Lazar N. Moving to a world beyond " $p < 0.05$ ". *American Statistician*. 2019;**73** (Supplement 1):1-19
- [3] Sullivan G, Feinn R. Using effect size—Or why the P value is not enough. *Journal of Graduate Medical Education*. 2012;**4**(3):279-282
- [4] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Revised ed. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.; 1977
- [5] Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972; **34**(2):187-202
- [6] Uno H, Claggett B, Tian L, Inoue E, Gallo P, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 2014; **32**(22):2380
- [7] Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*. 2009;**28**(19):2473-2489
- [8] Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*. 2011;**30**(19):2409-2421
- [9] Wang H, Chen D, Pan Q, Hueman MT. Using weighted differences in hazards as effect sizes for survival data. *Journal of Statistical Theory and Practice*. 2022;**16**(1):12
- [10] Wang H. *Development of Prognostic Systems for cancer Patients*. USA: The George Washington University; 2020
- [11] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965;**52** (1-2):203-224. DOI: 10.2307/233382
- [12] Surveillance, Epidemiology, and End Results (SEER) Program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) Research Data (2000-2020), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission. Available from: <https://seer.cancer.gov/>
- [13] Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A. *AJCC Cancer Staging Manual*. 7th ed. New York: Springer-Verlag; 2010
- [14] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958;**53**:457-481
- [15] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;**81**(3):515-526
- [16] Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press; 1997. Available from: <https://www.amazon.com/Bootstrap-Application-Statistical-Probabilistic-Mathematics/dp/0521574714>
- [17] Yang CQ, Gardiner L, Wang H, Hueman MT, Chen D. Creating prognostic systems for well-differentiated thyroid cancer using machine learning. *Frontiers in Endocrinology*. 2019;**10**:288

[18] Grimley PM, Liu Z, Darcy KM, Hueman MT, Wang H, Sheng L, et al. A prognostic system for epithelial ovarian carcinomas using machine learning. *Acta obstetrica et gynecologica Scandinavica*. 2021;**100**(8):1511-1519

[19] Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. *Journal of Biomedicine & Biotechnology*. 2009; **2009**:632786. DOI: 10.1155/2009/632786. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2702512/>

[20] Hueman MT, Wang H, Yang CQ, Sheng L, Henson DE, Schwartz AM, et al. Creating prognostic systems for cancer patients: A demonstration using breast cancer. *Cancer Medicine*. 2018; **7**(8):3611-3621

[21] Hueman M, Wang H, Henson D, Chen D. Expanding the TNM for cancers of the colon and rectum using machine learning: A demonstration. *ESMO Open*. 2019;**4**(3):e000518

[22] Hueman M, Wang H, Liu Z, Henson D, Nguyen C, Park D, et al. Expanding TNM for lung cancer through machine learning. *Thoracic Cancer*. 2021;**12**(9):1423-1430

[23] Yang CQ, Wang H, Liu Z, Hueman MT, Bhaskaran A, Henson DE, et al. Integrating additional factors into the TNM staging for cutaneous melanoma by machine learning. *PLoS One*. 2021;**16**(9):e0257949

[24] Wang H, Liu Z, Yang J, Sheng L, Chen D. Using machine learning to expand the Ann Arbor staging system for Hodgkin and Non-Hodgkin lymphoma. *BioMedInformatics*. 2023; **3**(3):514-525



# Perspective Chapter: Linear Regression and Logistic Regression Models

*Dilip Kumar Ghosh*

## Abstract

In this chapter, we have discussed the detailed concept of simple linear regression and logistic regression analysis. Further we have discussed the procedure of computing regression coefficients, standard error, t test, Z test, p value and 95% confidence intervals for simple linear regression and logistic regression analysis. We also explained that for testing the simple linear regression coefficient, we use t test, whereas, for testing the logistic regression coefficient, we use Z test. Several examples on medical data are considered and various related statistics were computed using manually, R studio package, and Jamovi.

**Keywords:** regression model, logistic function, odds ratio, scatter diagram, regression, coefficients, estimators, predicted value

## 1. Introduction

The method of linear regression is used in predicting the value of one variable based on the value of another variable. The variable you need to predict is known as dependent variable, whereas the variable used to predict the other variable's value is known as independent variable. This method contains one or more than one explanatory variables. If it contains only one explanatory variable is called simple linear regression, otherwise, multiple linear regression model. The regression coefficients involved in linear equation is estimated using the least squares method of estimation. Once regression coefficients are estimated, you can fit a model that predicts the value of dependent variable. Linear regression is used to establish a linear relation between response and explanatory variable in biological, behavioral, environmental, social sciences, business etc. Montgomery et al. [1], Rencher and Bruce [2], Swaminathan [3] and Lane [4] discussed regression analysis with several examples. Further, Noce and McKeown [5] discussed a logistic modeling of factors influencing internet use. While Seo et al. [6] discussed the relations between physical activity and behavioral.

Logistic regression is a statistical method that is used to establish a relationship between one dependent variable and one or more than one explanatory variables, where dependent variable is dichotomous.

## 2. Simple linear regression

Suppose random samples  $(x_{i1}, x_{i2}, \dots, x_{in}, y_1)$  of size  $n$ , where  $i = 1, 2, \dots, n$ ; is drawn from a population. The random variables  $(x_1, x_2, \dots, x_n)$  are generally known as predictor variables. However, depending upon situation and with practical point of view, these random variables are also known with different names. The different names are independent variables, covariates, regressor and explanatory variables. Variable  $y$  is called as response variable. Sometimes variable  $y$  is also known as dependent variable or outcome variable. Suppose we are willing to establish a linear regression between response variable  $y$  and explanatory variables  $(x_1, x_2, \dots, x_n)$ , then it could be represented by a model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + e \quad (1)$$

where,  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  are parameters and is known as regression coefficients,  $e$  is random error which is distributed normally with mean zero and variance  $\sigma^2$ .

For example, the effect of age, weight, height and walking habit on systolic blood pressure. This model is called multiple regression models as predicted variables are more than one.

Suppose we are interested in bi-variate regression model, where  $Y$  is response variable and  $X$  is predicted variable. This model is called simple linear regression model or general linear model. This model is represented by

$$Y = \beta_0 + \beta_1 X + e \quad (2)$$

Where,  $Y$  is response variable,  $X$  is predicted variable,  $\beta_0$  is intercept,  $\beta_1$  is regression coefficient and  $e$  is random variable; and  $e$  is distributed normally with mean zero and variance  $\sigma^2$ . For example, the effect of age on systolic blood pressure.

## 3. Scatter diagram

A scatter diagram is a two-dimension graph involving the magnitude of the response variable ( $Y$ ) and predicted variable ( $X$ ). Scatter diagram provides a rough idea about the relationship between response and predicted variables. Following are the various steps for drawing a scatter diagram:

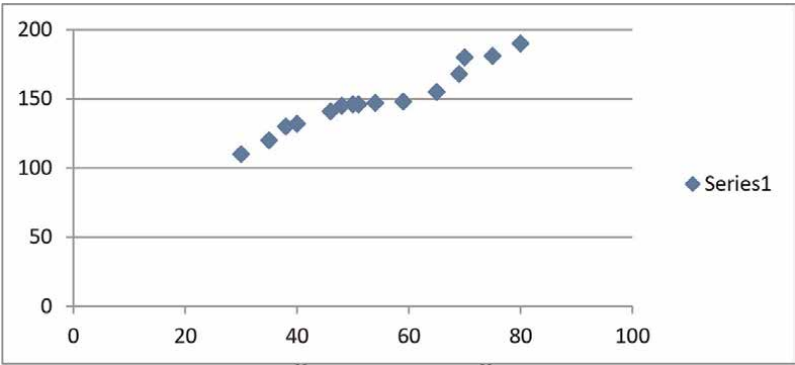
- i. Select the horizontal axis ( $X$ ) and vertical axis ( $Y$ )
- ii. Take response variable on  $Y$ -axis and predicted variable on  $X$ -axis.
- iii. Tick the point at the corresponding area of  $(X, Y)$ .

In the scatter diagram, if the observations are approximately scattered around the straight line, it shows a linear relationship between response and predicted variables. Once the relationship is established, one can use simple linear regression model to know the relationship between the variables. However, if the observations are not scattered around the straight line, it does not show a linear relationship between the two variables. In such situation, one can use transformation or non-linear regression method to find the best fitted regression model.

**Example 1:** A sample of 15 men of age group 30–70 was collected to investigate the effect of weight of the patients on the sugar level of the diabetic patients. The data on the blood sugar level (mg/dl) and weight (in kg) of 15 men are shown in **Table 1**.

S. no.	Blood sugar level	Weight of the patients (in kg)
1	146	50
2	145	48
3	141	46
4	168	69
5	132	40
6	190	80
7	180	70
8	130	38
9	181	75
10	148	59
11	110	30
12	147	54
13	146	51
14	120	35
15	155	65

**Table 1.**  
*Blood sugar level and duration of walk of 15 men.*



**Figure 1.**  
*Scatter diagram.*

Draw the scatter diagram and give your interpretation about the data.  
From **Figure 1**, it is clear that approximately all the 15 observations are scattered around the straight line. Hence, there is linear relationship between blood sugar level and weight of the person.

### 3.1 Assumptions underlying linear regression model

For applying any statistical method, first of all, we should study the assumptions underlying it. So we shall discuss the assumptions of simple linear regression models. Following are the assumptions:

- i. The regression model is assumed to be linear in parameters.
- ii. The error term  $e$  is assumed to be normally distributed with mean 0 and variance  $\sigma^2$ , i.e.,  $e \sim N(0, \sigma^2)$ .

### 3.2 Estimation of parameters

For regression model (2),  $\beta_1$  is the parameter and is constant and unknown which can be estimated using least squares method of estimation. Model (2) can be rewritten as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \text{ where } i = 1, 2, \dots, n. \quad (3)$$

$$E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (4)$$

On differentiating the error sum of squares (E) with respect to  $\beta_0$  and  $\beta_1$  and then equating them to zero, we can obtain the least squares estimator of  $\beta_0$  and  $\beta_1$ .

$$\frac{\delta E}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (5)$$

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (6)$$

$$\text{So, } \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i \quad (7)$$

$$\text{Similarly, } \frac{\delta E}{\delta \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (8)$$

$$\sum_{i=1}^n (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \quad (9)$$

$$\text{So, } \sum_{i=1}^n y_i x_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \quad (10)$$

On solving (7) and (10), we get

$$\hat{\beta}_1 \left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] = n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i \quad (11)$$

$$\text{Hence, } \hat{\beta}_1 = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (12)$$

$$\text{Or, } \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - n \bar{y} \bar{x}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (13)$$

Eq. (12) can easily be written as.



$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (14)$$

On dividing the numerator and denominator of the above equation by  $n$ , we have,

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (15)$$

From (7), we have  $n\bar{y} = n\hat{\beta}_0 + n\hat{\beta}_1 \bar{x}$

$$\text{So, } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (16)$$

Here,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the least square estimator of intercept  $\beta_0$  and slope  $\beta_1$ . In the regression model, slope is called regression coefficient. Hence, on ward  $\beta_1$  will be called regression coefficient. Thus, for the linear regression method, the fitted regression model is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (17)$$

In the matrix notation, the linear model can be written as

$$Y = X\beta + e \quad (18)$$

Where,  $Y$  is a vector of  $n \times 1$  observations,  $X$  is a matrix of  $n \times 2$ ,  $\beta$  is a vector of  $2 \times 1$  parameters, and  $e$  is the random error of  $n \times 1$ .

Using least squares method of estimation, the normal equation is obtained as

$$X'Y = X'X\beta \quad (19)$$

On multiplying both side of (19) by  $(X'X)^{-1}$ , we have

$$(X'X)^{-1}(X'Y) = (X'X)^{-1}(X'X)\beta \quad (20)$$

$$\text{Hence, } \hat{\beta} = (X'X)^{-1}(X'Y) \quad (21)$$

Where,  $\hat{\beta}$  is the estimate of the regression coefficient  $\beta$ .

In that case, the fitted regression model is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (22)$$

**Example 2:** A samples of 15 men of age group 30–70 was collected to investigate the effect of weight (in kg) of the patients on the blood pressure level of the diabetic patients. The data on the blood pressure level (mm/hg) and weight (in kg) of 15 men are shown in **Tables 2** and **3**.

Where, residual =  $(y - \text{predicted value of } y)$ .

$$\begin{aligned} \text{Regression coefficient, } \hat{\beta}_1 &= \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{15 \times 130284 - 2153 \times 864}{15 \times 52622 - (864)^2} = \frac{1954260 - 1860192}{789330 - 746496} = \frac{94068}{42834} \end{aligned} \quad (23)$$

S. no.	Blood pressure level (mm/hg)	Weight (in kg)
1	125	50
2	123	48
3	120	46
4	181	73
5	105	40
6	190	80
7	185	75
8	118	45
9	175	74
10	168	69
11	110	43
12	130	54
13	128	51
14	116	44
15	179	72

**Table 2.**  
*Blood pressure level and weight of 15 men.*

y	x	y square	x square	xy	Predicted value of y	Residual	Residual sum of squares
125	50	15,625	2500	6250	126.8429	-1.84293	3.396390985
123	48	15,129	2304	5904	122.4507	0.549282	0.301710716
120	46	14,400	2116	5520	118.0585	1.941494	3.769398952
181	73	32,761	5329	13,213	177.3534	3.646632	13.29792494
105	40	11,025	1600	4200	104.8819	0.11813	0.013954697
190	80	36,100	6400	15,200	192.7261	-2.72611	7.431675732
185	75	34,225	5625	13,875	181.7456	3.25442	10.59124954
118	45	13,924	2025	5310	115.8624	2.1376	4.56933376
175	74	30,625	5476	12,950	179.5495	-4.54947	20.69771368
168	69	28,224	4761	11,592	168.5689	-0.56894	0.323697275
110	43	12,100	1849	4730	111.4702	-1.47019	2.161452755
130	54	16,900	2916	7020	135.6274	-5.62735	31.66711304
128	51	16,384	2601	6528	129.039	-1.03904	1.079595809
116	44	13,456	1936	5104	113.6663	2.333706	5.446183694
179	72	32,041	5184	12,888	175.1573	3.842738	14.76663534
2153	864	322,919	52,622	130,284	2153		119.5140309

**Table 3.**  
*Predicted and residual value.*

Predictor	Estimate	SE	t	P
Intercept	17.03763	3.3607	5.07	<0.001
$\beta$	2.196106	0.0567	38.70	<0.001

**Table 4.**  
Model coefficient.

Hence,  $\hat{\beta}_1 = 2.196106$ .

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 143.5333 - 2.196106 \times 57.6 \\ &= 17.03763.\end{aligned}$$

The fitted regression model is given by

$$\hat{Y} = 17.03763 + 2.196106 X \quad (24)$$

$r = 0.996$  and  $R^2 = 0.991$ .

From model coefficient (**Table 4**), it is obvious that regression coefficient of predictor weight is highly significant as  $p < 0.001$ . Also  $R^2$  is very close to 1. This concludes that as the weight of the patient increases, blood pressure increases. That is, weight is under control, the blood pressure is normal or under normal. In this example, weight ranges from 69 to 80, the blood pressure ranges from 179 to 190. However, weight ranges from 40 to 54, the blood pressure ranges from 105 to 130. Again, when the weight ranges from 69 to 80, the blood pressure ranges from 168 to 190. Thus, higher the weight, higher is blood pressure.

### 3.3 Regression coefficient using R studio package

```
>y = c(125,123,120,181,105,190,185,118,175,168,110,130,128,116,179)
>x = c(50,48,46,73,40,80,75,45,74,69,43,54,51,44,72)
>result = data.frame(y,x)
>z = lm(y ~ x,result)
>summary(z)
```

## 4. Forecasting or predicted value of Y

Using the fitted model (24), we can easily forecast the blood pressure level corresponding to its weight. Suppose, weight of a patient is 82 kg then its blood pressure will be obtained from  $\hat{Y} = 17.03763 + 2.196106 \times 82 = 197.1183$ .

*Residuals:* We can obtain the residuals of all the 15 patients using  $(Y - \hat{Y})$ . This value is shown in **Table 3**.

*R Squares:* We can compute  $R^2$  using the residual sum of squares from the expression

$$R^2 = 1 - \frac{\text{Residual sum of squares}}{\text{total sum of squares of } y} \quad (25)$$

Where, residual sum of squares =  $\sum (Y - \hat{Y})^2 = 119.5140309$  and is shown in

**Table 3.**

Total sum of squares of y =  $\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = 322,919 - \frac{(2153)^2}{15} = 13891.73333$ .

So,  $R^2 = 1 - \frac{119.5140309}{13891.73333} = 0.9914$ .

We can also obtain Regression coefficients using the Matrix form as following:

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 50 & 48 & 46 & 73 & 40 & 80 & 75 & 45 & 74 & 69 & 43 & 54 & 51 & 44 & 72 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 15 & 864 \\ 864 & 52622 \end{bmatrix}, X'Y = \begin{bmatrix} 2153 \\ 130284 \end{bmatrix} \text{ and}$$

$$(X'X)^{-1} = \begin{bmatrix} 1.22851006 & -0.0201708923 \\ -0.0201708923 & 0.0003501891 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 17.037634 \\ 2.196106 \end{bmatrix}. \hat{\beta}_0 = 17.037634 \text{ and } \hat{\beta}_1 = 2.196106.$$

#### 4.1 R studio program to obtain the estimate of regression coefficients

```
>X = matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,50,48,46,73,40,80,75,45,74,69,43,54,51,
44,72),ncol = 2)
> X' = t(X)
> X'X = X'% * %X
>Y = c(125,123,120,181,105,190,185,118,175,168,110,130,128,116,179)
>m = solve(X'X)
> X'Y = X'% * %Y
>n=X'Y
>β=m% * %n
```

## 5. Regression lines

Let the regression model is denoted by  $Y = a + bX$ , then the two types of regression lines are following:

i. Regression line of Y on X and is denoted by  $b_{YX}$

ii. Regression line of X on Y and is denoted by  $b_{XY}$

$$b_{YX} = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = 2.196106 \quad (26)$$

$$\text{Regression line of X on Y and is denoted by } b_{XY} = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2} = 0.4514 \quad (27)$$

## 6. Expectation and variance of estimators

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1 \quad (28)$$

This show that least squares estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are unbiased estimators of  $\beta_0$  and  $\beta_1$ , respectively.

$$V(\hat{\beta}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\text{sum of squares of } x} \right] \quad (29)$$

$$\text{And } V(\hat{\beta}_1) = (X'X)^{-1} \sigma^2 = \frac{\sigma^2}{\text{sum of squares of } x}. \quad (30)$$

Where,  $\sigma^2$  is unknown. It is estimated from the given data as

$\hat{\sigma}^2 = \left( \frac{y_i - \hat{y}_i}{n-2} \right)^2$ , and sum of squares of x is obtained as

Sum of squares of x =  $\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$ ; where n is number of observations.

Thus, when  $\sigma^2$  is unknown,  $V(\hat{\beta}_0)$  and  $V(\hat{\beta}_1)$  is determined as

$$V(\hat{\beta}_0) = \hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\text{sum of squares of } x} \right] \text{ and} \quad (31)$$

$$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\text{sum of squares of } x}. \quad (32)$$

Again, when  $\sigma^2$  is unknown, standard error of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are determined as

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\text{sum of squares of } x} \right]} \text{ and} \quad (33)$$

$$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{\text{sum of squares of } x}}. \quad (34)$$

**Example 3:** Consider the example 2, compute the estimate of  $V(\hat{\beta}_0)$ ,  $V(\hat{\beta}_1)$  and its standard error.

Solution: From example 2, we have  $\hat{\beta}_0 = 17.037634$  and  $\hat{\beta}_1 = 2.196106$ . Sum of squares of X =  $\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = 52,622 - \frac{864^2}{15} = 52,622 - 49766.4 = 2855.6$ .

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{119.5140}{13} = 9.1934.$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{864}{15} = 57.6.$$

Estimates of the  $V(\hat{\beta}_0)$  and  $V(\hat{\beta}_1)$  can be determined from

$$\begin{aligned}
 V(\hat{\beta}_0) &= \hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\text{sum of squares of } x} \right] \\
 &= 9.1934 \left[ \frac{1}{15} + \frac{(57.6)^2}{2855.6} \right]
 \end{aligned} \tag{35}$$

$$\text{So, } V(\hat{\beta}_0) = 11.2941844.$$

$$V(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{\text{sum of squares of } x} = \frac{9.1934}{2855.6} = 0.003219428 \tag{36}$$

$$SE(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{11.2941844} = 3.360682133,$$

$$SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{0.003219428} = 0.056740004.$$

## 7. Testing of hypothesis of estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$

For testing the hypothesis that the sample comes from the population for which the value of  $\beta_0$  is equal to 0. That is,

Our Null hypothesis  $H_0 : \beta_0 = 0$

Against the alternate hypothesis  $H_1 : \beta_0 \neq 0$ .

Under the null hypothesis, for testing  $H_0 : \beta_0 = 0$ , we define the test statistics given by

$$t = \frac{\hat{\beta}_0 - E(\hat{\beta}_0)}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} \tag{37}$$

$$= \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\text{sum of squares of } x} \right]}}, \text{ where } \sigma^2 \text{ is unknown.} \tag{38}$$

Similarly, for testing the hypothesis that the sample comes from the population for which the value of  $\beta_1$  is equal to 0. That is,

Our Null hypothesis  $H_0 : \beta_1 = 0$ .

Against the alternate hypothesis  $H_1 : \beta_1 \neq 0$ .

Thus, Under the null hypothesis, for testing  $H_0 : \beta_1 = 0$ , we define the test statistics given by

$$t = \frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\text{sum of squares of } x}}} \tag{39}$$

where  $\sigma^2$  is unknown, and statistics  $t$  follows student's  $t$  distribution with  $(n - 2)$  degrees of freedom.

**Example 4:** Consider the data given in example 2. Find the effect of weight on blood pressure of 15 patients, test the null hypothesis for testing the significance of the regression coefficients  $\hat{\beta}_0$  and  $\hat{\beta}_1$  at 5% level of significance.

Solution: From example 2, we have,

$$\hat{\beta}_0 = 17.037634, SE(\hat{\beta}_0) = 3.360682133, \text{ and}$$

$$\hat{\beta}_1 = 2.196106, SE(\hat{\beta}_1) = 0.056740004.$$

Under the null hypothesis, for testing  $H_0 : \beta_0 = 0$ , we have,

$$t = \frac{\hat{\beta}_0 - E(\hat{\beta}_0)}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} \quad (40)$$

where  $\sigma^2$  is unknown.

$$t = \frac{17.037634}{3.360682133} = 5.069695.$$

Under the null hypothesis, for testing  $H_0 : \beta_1 = 0$ , we have

$$t = \frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}, \text{ where } \sigma^2 \text{ is unknown.} \quad (41)$$

$$t = \frac{2.196106}{0.056740004} = 38.70472.$$

With  $\alpha = 0.05$ , for two sided  $\alpha$ , the tabulated value of  $t$  at 13 degrees of freedom with 5% level of significance is 2.160. In case of  $H_0 : \beta_0 = 0$ , the calculated value of  $t$  ( $=5.069695$ ) is greater than tabulated value of  $t$ , so test is significant. That is, we reject the null hypothesis. Hence, we may conclude that the value of  $\beta_0$  is not equal to zero.

Similarly, In case of  $H_0 : \beta_1 = 0$ , the calculated value of  $t$  ( $=38.70472$ ) is greater than tabulated value of  $t$ , so test is highly significant. That is, we reject the null hypothesis. Hence, we may conclude that the value of  $\beta_1$  is not equal to zero. Thus, the fitted simple regression model is highly significant. In other word, we can say as the weight of the patient increases, the chance of blood pressure may increase.

Alternatively, we can also test the significance of regression coefficient  $\beta_1$  using the analysis of variance **Table 5**.

Sources of variation	Degrees of freedom	Sum of squares	Mean squares	F-Ratio	P-value
Regression	1	$\sum (\hat{y}_i - \bar{y})^2$	SSR/df = MSR	MSR/MSE	
Error	(n - 2)	$\sum (y_i - \hat{y}_i)^2$	SSE/df = MSE		
Total	(n - 1)	$\sum (y_i - \bar{y})^2$			

**Table 5.**  
Analysis of variance.

y	x	Predicted value of y ( $\hat{y}_i$ )	$\sum (\hat{y}_i - \bar{y})^2$	$\sum (y_i - \hat{y}_i)^2$	$\sum (y_i - \bar{y})^2$
125	50	126.8429	278.568451	3.396390985	343.4832089
123	48	122.4507	444.475264	0.301710716	421.6164089
120	46	118.0585	648.965129	3.769398952	553.8162089
181	73	177.3534	1143.797	13.29792494	1403.753609
105	40	104.8819	1493.93304	0.013954697	1484.815209
190	80	192.7261	2419.93256	7.431675732	2159.154209
185	75	181.7456	1460.17834	10.59124954	1719.487209
118	45	115.8624	765.678707	4.56933376	651.9494089
175	74	179.5495	1297.16479	20.69771368	990.1532089
168	69	168.5689	626.78347	0.323697275	598.6194089
110	43	111.4702	1028.04315	2.161452755	1124.482209
130	54	135.6274	62.5039822	31.66711304	183.1502089
128	51	129.039	210.083689	1.079595809	241.2834089
116	44	113.6663	892.038047	5.446183694	758.0826089
179	72	175.1573	1000.07497	14.76663534	1257.886809
2153	864	2153	13772.2206	119.5140309	13891.73333

**Table 6.**  
Regression, error and total sum of squares.

Sources of variation	Degrees of freedom	Sum of squares	Mean squares	F-Ratio	P-value
Regression	1	13772.2206	13772.2206	1498.057	<0.001
Error	13	119.5140309	9.193387		
Total	14	13891.73333			

**Table 7.**  
Analysis of variance.

**Example 5:** Consider the data given in example 2. Find the effect of weight on blood pressure of 15 patients, test the null hypothesis for testing the significance of the regression coefficients  $\hat{\beta}_1$  at 5% level of significance using analysis of variance **Table 5**.

Using **Table 3**, we have **Table 6**.

Using **Table 6**, we can obtain the ANOVA table as shown in **Table 7**.

From **Table 7**, we can observe the p value for regression coefficient is less than 0.001 as well calculated value of F is very large. Which is greater than tabulated value of F with (1, 13) degrees of freedom at 5% level of significance, where tabulated value of F is 4.67. This shows that the test is significant and hence rejects the null hypothesis.

## 7.1 R studio program for obtaining ANOVA table

```
>y = c(125,123,120,181,105,190,185,118,175,168,110,130,128,116,179)
```



```
>x = c(50,48,46,73,40,80,75,45,74,69,43,54,51,44,72)
>result = data.frame(y,x)
> av. = aov(y ~ x,result)
> summary(av)
```

## 8. Confidence interval of estimated regression coefficients

Now we discuss how to obtain 100% confidence interval of estimated regression coefficients. In fact we are interested for confidence interval for regression coefficient  $\beta_1$  only. In fact we generally compute the confidence interval to determine the range. In case of regression coefficient  $\beta_1$ , we wish to determine the lower and upper limit of the  $\beta_1$ . We can obtain the 100% confidence interval of the  $\beta_1$  as

Confidence interval for  $\beta_1 = \hat{\beta}_1 \pm t_{(n-2),\alpha} SE(\hat{\beta}_1)$  for two tailed test.

**Example 6:** Consider the data given in example 2. Find the effect of weight on blood pressure of 15 patients. Obtain 100% confidence interval of regression coefficients  $\hat{\beta}_1$ .

Solution: For this data from example 2, we have,

$$\hat{\beta}_1 = 2.196106, SE(\hat{\beta}_1) = 0.056740004 \text{ and } t_{(n-2),\alpha} = 2.160.$$

$$\begin{aligned} \text{Confidence interval for } \beta_1 &= 2.196106 \pm 2.160 \times 0.056740004 \\ &= 2.196106 \pm 0.122558409 \end{aligned}$$

So, lower confidence limit of  $\beta_1$  is  $2.196106 - 0.122558409 = 2.073547489$ .

Upper confidence limit of  $\beta_1$  is  $2.196106 + 0.122558409 = 2.318664306$ .

Thus, the confidence limit of  $\beta_1$  is ranges from 2.073547489 to 2.318664306.

If one is interested for determining the confidence limit of  $\beta_0$ , the same procedure can be used.

## 9. Logistic regression

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regressions with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest

We can also say that Logistic regression is used for predicting binary outcomes on the basis of one or more predictor variables. The concept of logistic regression is similar to the ordinary multiple linear regression. In this method we are willing to fit a best model which can determine the relationship between a response variable and one or more explanatory variables. As in the case of ordinary linear regression, the form of the model is linear with respect to the regression parameters, the same is true for the logistic regression. The only difference between the two regression is following: in logistic regression the response variable is binary (also called dichotomous), whereas in ordinary linear regression it is continuous. Logistic regression can also be called as a predictive algorithm in which by using explanatory variables one can predict the dependent variable, just like Linear

regression, but with a simple difference that the dependent variable in the logistic regression should be considered as a categorical variable.

## 10. Logistic function

For understanding the logistic regression, first we must determine logistics function. Let us consider the equation of the best fit model in the simple linear regression as

$$y = \beta_0 + \beta_1 x \quad (42)$$

where,  $y$  is response variable and  $x$  is explanatory variable.

Let us replace  $y$  by probability  $P$  which is given as

$$P = \beta_0 + \beta_1 x \quad (43)$$

In (43) value of  $P$  may be negative in some case and in some other cases, value of  $P$  may be more than one. However, value of  $P$  ranges from 0 to 1 only. This is a contradiction. To overcome this problem, we can take odds of  $P$  instead of probability. Odds of probability is defined as  $\text{odd} = \frac{P}{1-P}$ . That is, odds of probability is defined as the ratio of the probability of success and the probability of failure.

So, from (43), we have

$$\frac{P}{1-P} = \beta_0 + \beta_1 x. \quad (44)$$

As we are aware that odds will be always positive, that is odds are ranging from 0 to infinity. Again, to overcome this problem we take the log transformation because by considering log transformation it will range from  $-\infty$  to  $+\infty$ .

$$\text{Log} \left( \frac{P}{1-P} \right) = \beta_0 + \beta_1 x \quad (45)$$

On taking exponential on both sides of (45), we have

$$\text{Exp} \left[ \text{Log} \left( \frac{P}{1-P} \right) \right] = \exp[\beta_0 + \beta_1 x] \quad (46)$$

That is,  $\left( \frac{P}{1-P} \right) \log(e) = e^{\beta_0 + \beta_1 x}$

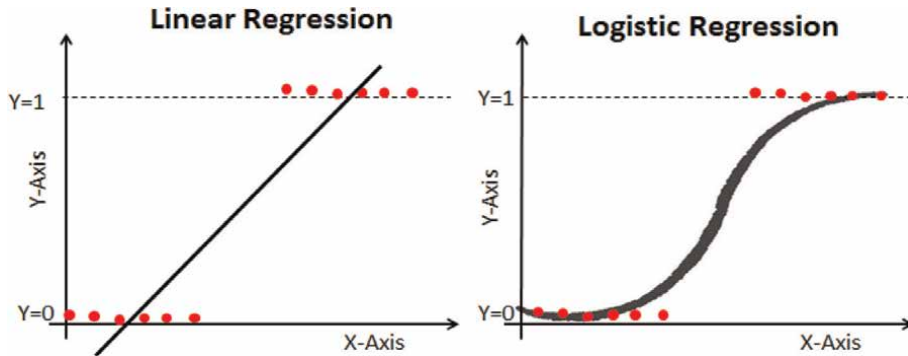
$$\left( \frac{P}{1-P} \right) = e^{\beta_0 + \beta_1 x} \quad (47)$$

$$P = (1-P) e^{\beta_0 + \beta_1 x}$$

$$P + P e^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1 x}$$

$$P(1 + e^{\beta_0 + \beta_1 x}) = e^{\beta_0 + \beta_1 x}$$

$$\text{So, } P = \frac{e^{\beta_0 + \beta_1 x}}{(1 + e^{\beta_0 + \beta_1 x})} \quad (48)$$



**Figure 2.**  
Shape of linear and logistic regression.

On dividing the numerator and denominator of (48) by  $e^{\beta_0 + \beta_1 x}$ , we have

$$P = \frac{(e^{\beta_0 + \beta_1 x}) / (e^{\beta_0 + \beta_1 x})}{(1 + e^{\beta_0 + \beta_1 x}) / (e^{\beta_0 + \beta_1 x})} = \frac{1}{(1 + e^{\beta_0 + \beta_1 x}) / e^{\beta_0 + \beta_1 x}} \quad (49)$$

$$\text{Thus, } P = \frac{1}{[1 + e^{-(\beta_0 + \beta_1 x)}]} \quad (50)$$

We can call (50) as a logistic function.

If we consider only one explanatory variable then the graph of simple linear regression will give a straight line however, the graph of logistic regression will be of S shape. This is shown in **Figure 2**.

Logistic regression can also be called as a predictive algorithm in which by using explanatory variables one can predict the dependent variable, just like Linear Regression, but with a simple difference that the dependent variable in the logistic regression should be considered as a categorical variable.

**Example 7:** A random sample of 300 women were selected, where 300 women are either suffering with cancer or not. The response of yes will be asked from the 300 women. It is found that 225 women responded yes. The response of yes out of n sample number follows binomial distribution with parameters n and p. Obtain the odds ratio.

**Solution:** Out of 300 women 225 have responded yes for cancer. So, the sample proportion is

$$\hat{P} = \frac{225}{300} = 0.75.$$

Sample proportion cannot be used for finding logistic regression and hence, we need the odds. Where odds is the ratio of proportion for two outcomes. One outcomes is “yes” and the other outcomes is “no”. Proportion of yes is 0.75, hence, proportion of no is  $1 - \hat{P} = 1 - 0.75 = 0.25$ .

$$\text{Odd ratio of yes and no of women cancer} = \frac{\hat{P}}{1 - \hat{P}} = \frac{0.75}{0.25} = 3.$$

Hence, odds are 3 to 1 that woman has cancer yes to no. Similarly, we can also say odds are 1 to 3, that is, women has cancer no to yes is 1 to 3.

**Example 8:** The sample proportion of women who were detected as cancer patient is 65%, whereas the sample proportion of men detected as cancer patient is 45%.

In this sample of young adult, it can be observe that the sample proportion of women detected as cancer patient is 20% higher than the sample proportion of men detected as cancer. Now we wish to analyze this data using logistic regression. In this example the predictive variable is sex which is a categorical variable. So we need to use a numeric code. The better way is to use a indicator saying whether the adult is women or not. The indicator function is defined as

$$x = \begin{cases} 1 & \text{if the person is women} \\ 0, & \text{if the person is men} \end{cases} \quad (51)$$

Since, the response is given in proportion, so we transform it into odds. There will be two odds, one for women and other for men.

$$\text{Odds for women are given as } \frac{\hat{P}}{1 - \hat{P}} = \frac{0.65}{1 - 0.65} = \frac{0.65}{0.35} = 1.8571.$$

$$\text{Similarly, odds for men is given as } \frac{\hat{P}}{1 - \hat{P}} = \frac{0.45}{1 - 0.45} = \frac{0.45}{0.55} = 0.8182.$$

Now we can build the logistic regression model by considering  $\log(\text{odds})$  as the linear function of the explanatory variable. Hence, logistics model is defined as

$$\log\left(\frac{\hat{P}}{1 - \hat{P}}\right) = \beta_0 + \beta_1 x, \quad (52)$$

where  $x$  is explanatory variable,  $p$  is the binomial proportion and  $\beta_0, \beta_1$  are the parameters of the logistic regression model.

Here, there are only two values of  $x$  and hence write two equations: one for women and other for men.

$$\text{For women, } \log\left(\frac{\hat{P}}{1 - \hat{P}}\right) = \beta_0 + \beta_1 \times 1 \quad (53)$$

$$\text{And for men, } \log\left(\frac{\hat{P}}{1 - \hat{P}}\right) = \beta_0 + \beta_1 \times 0 \quad (54)$$

Because, there is a  $\beta_1$  in the equation of women as  $x = 1$ . This is missing in the equation of men as  $x = 0$ .

Therefore, the logistic regression model for women and men are following:

$$\text{Log}(1.8571) = \beta_0 + \beta_1 \quad (55)$$

$$\text{Log}(0.8182) = \beta_0$$

$$\beta_0 + \beta_1 = 0.6190 \quad (56)$$

$$\beta_0 = -0.20065 \quad (57)$$

On solving (24) and (25), we have

$$-0.20065 + \beta_1 = 0.6190 \quad (58)$$

Hence,  $\beta_1 = 0.6190 + 0.20065 = 0.81965$ .

Now the fitted logistic regression model is given by

$$\text{Log}(\text{odds}_{\text{women}}) = \beta_0 + \beta_1 \quad (59)$$

$$\text{So, } \text{odds}_{\text{women}} = e^{\beta_0 + \beta_1} \quad (60)$$

$$\text{Similarly, } \text{odds}_{\text{men}} = e^{\beta_0} \quad (61)$$

$$\frac{\text{odds}_{\text{women}}}{\text{odds}_{\text{men}}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} = e^{0.81965} = 2.269705.$$

$$\text{odds}_{\text{women}} = 2.269705 \times \text{odds}_{\text{men}}.$$

That is, we can say that odds of women are 2.269705 times odds of men.

Note that, if we have indicator function as

$$x = \begin{cases} 0 & \text{if the person is women} \\ 1 & \text{if the person is men} \end{cases} \quad (62)$$

Then the sign of  $\beta_1$  will be negative. That is,

$$\text{Log}(\text{odds}_{\text{women}}) = \beta_0. \text{So, } \text{odds}_{\text{women}} = e^{\beta_0} \quad (63)$$

$$\text{Similarly, } \text{odds}_{\text{men}} = e^{\beta_0 + \beta_1} \quad (64)$$

$$\frac{\text{odds}_{\text{women}}}{\text{odds}_{\text{men}}} = \frac{e^{\beta_0}}{e^{\beta_0 + \beta_1}} = e^{-\beta_1} = e^{-0.81965} = 0.440586$$

$$\text{odds}_{\text{women}} = 0.440586 \times \text{odds}_{\text{men}}. \quad (65)$$

Therefore, we can say odds of women are 0.440586 times odds of men.

**Example 9:** Hemoglobin contain of 20 patients corresponding to their age was collected at a hospital to know the relationship between hemoglobin and age. The collected observations are shown in **Table 8**.

If we use a simple linear regression to find the effect of age on the response variable Hemoglobin, we obtain the following statistics using software (**Table 9**).

Hb(g/Dl)	Age (Year)	Anemic(1 = yes, 0 = no)
11.2	15	1
11.3	21	1
11.5	23	1
16.3	25	0
16.5	26	0
10.1	28	1

Hb(g/dl)	Age (Year)	Anemic(1 = yes, 0 = no)
9.9	30	1
17.1	32	0
17.2	34	0
17.9	36	0
10.1	38	1
11.6	40	1
18.3	43	0
18.6	46	0
18.9	54	0
19.2	56	0
19.6	58	0
19.9	60	0
16.9	62	0
17.2	69	0

**Table 8.**  
Level of hemoglobin and its corresponding age.

Predictor	Estimate	SE	t	P
Intercept	9.382	1.7875	5.25	<0.001
Age	0.153	0.0420	3.64	0.002

**Table 9.**  
Model coefficient and 95% confidence.

Here regression coefficient is significant. That is, there is significant effect of age on hemoglobin.

As we are aware that the amount of hemoglobin in whole blood is expressed in grams per deciliter (g/dl). The normal Hb level for males is 14 to 18 g/dl; and for females is 12 to 16 g/dl. When the hemoglobin level is low, the patient has *anemia*.

If we are interested to know whether the patient is suffering with anemia then we have to use the logistic regression method. For this we have to transform the Hemoglobin data into presence or absence of Anemia. Since, the data belongs to women patients, so if the value is less than 12, the code is 1, that is women has Anemia, while the value is more than 12, the code is 0 (no Anemia). This is shown in column 3 of the **Table 8**. Now we fit a logistic regression between presence/absence of anemia and actual age using the software Jamovi. The following statistics is obtained (**Table 10**).

Predictor	Estimate	SE	t(Z)	Odds ratio	P	95% confidence interval	
						Lower	Upper
Intercept	3.993	2.1239	1.88	54.216	0.060	0.844	3483.577
Age	-0.130	0.0629	-2.06	0.878	0.039	0.776	0.994

**Table 10.**  
Model coefficients and 95% confidence intervals.

Since p value of regression coefficient (Age) is less than 0.05 and hence test is significant. That is, as age increases, chance of Anemia decreases.

## 11. Testing of hypothesis and confidence intervals of logistic regression coefficient

For testing the hypothesis that the sample comes from the population for which the value of logistic regression coefficient  $\beta_1$  is equal to 0. That is, Our Null hypothesis  $H_0 : \beta_1 = 0$  against the alternate hypothesis  $H_1 : \beta_1 \neq 0$ .

Under the null hypothesis, for testing  $H_0 : \beta_1 = 0$ , we define the test statistics given by

$$Z = \frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-0.130}{0.0629} = -2.06. \quad (66)$$

We can obtain the 95% confidence interval of the logistic regression coefficient  $\beta_1$  as

Confidence interval for  $\beta_1 = e^{\hat{\beta}_1 \pm Z SE(\hat{\beta}_1)}$ . That is, lower limit =  $e^{\hat{\beta}_1 - Z SE(\hat{\beta}_1)}$  and upper limit =  $e^{\hat{\beta}_1 + Z SE(\hat{\beta}_1)}$ .

For Example 9, lower limit =  $e^{-0.130 - 1.96 \times 0.0629} = 0.776$  and upper limit =  $e^{-0.130 + 1.96 \times 0.0629} = 0.993$ .

## 12. Conclusions

The main objective of this chapter is to discuss about simple linear regression and logistic regression analysis. Here, we have explained how to estimate regression coefficients, its standard error, testing of hypothesis of regression coefficients, 95% confidence intervals for simple linear regression and logistic regression model. All the statistics calculated manually is verified using R studio package and Jamovi package.

## Acknowledgements

Author is thankful to the management committee members of the IntechOpen, the referee and the editor of this edited book for providing me an opportunity to share my works.


## **Author details**

Dilip Kumar Ghosh  
Marwadi University, Rajkot, Gujarat, India

\*Address all correspondence to: ghosh\_dkg@yahoo.com

## **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 



## References

- [1] Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. Wiley. © 2019-2021 Pluripotent Limited
- [2] Rencher AC, Bruce Schaalje G. Linear Models in Statistics. New Jersey: John Wiley; 2008
- [3] Swaminathan S. Regression Detailed View. Published in Towards Data Science; 2018. Available from: [https://scholar.google.com.vn/citations?view\\_op=view\\_citation&hl=vi&user=K8vtbzAAAAAJ&citation\\_for\\_view=K8vtbzAAAAAJ:d1gkVwhDpl0C](https://scholar.google.com.vn/citations?view_op=view_citation&hl=vi&user=K8vtbzAAAAAJ&citation_for_view=K8vtbzAAAAAJ:d1gkVwhDpl0C)
- [4] Lane DM. Introduction to Linear Regression, Chapter 14 Regression. Available from: <https://onlinestatbook.com/2/regression/intro.html>
- [5] Noce AA, McKeown L. A new benchmark for internet use: A logistic modeling of factors influencing internet use in Canada, 2005. Government Information Quarterly. 2008;25:462-476
- [6] Seo D-C et al. Relations between physical activity and behavioral and perceptual correlates among Midwestern college students. Journal of American College Health. 2007;56:187-197



# QoLMiss: Package for Repeatedly Measured Quality of Life of Cancer Patients Data

*Ankita Pal, Satyajit Pradhan, Aseem Mishra,  
Pankaj Chaturvedi and Atanu Bhattacharjee*

### Abstract

Quality of Life (QoL) has become increasingly important in cancer clinical trials. The R package QoLMiss is a package developed for computing the different sub-domains of the European Organization for Research and Treatment of Cancer (EORTC) questionnaire and also finding a survival link with these sub-domains. This package contains the scale scoring and survival outcomes of the other domains obtained from QoL data. The scale scores are also evaluated if there is the presence of missing data in repeatedly measured QoL data. The cancer specific QLQ are also considered scoring and survival analysis.

**Keywords:** quality of life, cancer, QLQ-C30, functional scales, symptom scales, global health status, hazard ratio, R package

### 1. Introduction

Current oncology focuses not only on pharmacological treatment but also on a fuller understanding of the experiences of patients and their families. This will help in prioritizing the allocation of resources and planning and providing holistic care that will measurably affect the quality of life [1]. The therapeutic efficacy of a controlled clinical trial is most often measured in terms of patients' survival. In cancer trials, several types of survival times are used, which include disease-free, relapse-free, local recurrence-free survival, LR recurrence-free survival, progression-free survival, disinfection-free survival, and overall survival [2].

Nearly every cancer treatment that intends a cure in some way interferes with a patient's bodily integrity. Quality of life (QoL) or a person's well-being encompasses a broad range of variables describing the patient's subjective reactions and perceptions to their environment as long as any treatment fails to expand the lives of patients exceptionally. The alternative preference is given to the increase in QoL [3]. Understanding the social consequences of disease is very important for any treatment protocol and acknowledging the fact medical intervention aims to increase the length and QoL. For these reasons, the quality, effectiveness, and efficiency of health care are often evaluated by their impact on a patient's QoL [4].

The statistical analysis of QoL is challenging, so a few assumptions are to be considered: (i) QoL is a subjective construct that is indirectly observed and measured, (ii) It is multiple dimensional based on different characteristics of physical and psychological well-being. (iii) QoL is time-dependent, which reflects a person's experiences.

The QoL in cancer is a multidimensional concept that is dynamic, referring to the patient's day-to-day life—balancing between the present situation and the ideal situation at a given time [5]. It is a specific and multidimensional type of patient-reported outcomes (PROs) that encompasses the patients' social, financial, psycho-social, and physical activities [6, 7]. After their completion of treatment, the QoL for cancer patients is related both directly and indirectly to health, disease, disability, and impairment. The more significant symptoms have been associated with, the higher levels of emotional suffering and poor physical and societal functioning, and global QoL.

A first-generation core questionnaire, the EORTC QLQ-C36, was developed in 1987 [8]. It is studied through validated questionnaires that the patients fill at different time points. The European Organization for Research and Treatment of Cancer (EORTC) has developed a questionnaire, named the QLQ-C30, which helps assess Health-Related Quality of Life (HRQoL) in cancer patients with 30 questions and some extra questions related to disease-specific treatment measurements. The EORTC QLQ-C30 is a widely used and well-validated instrument that is designed to assess health-related quality of life in patients with cancer.

EORTC QLQ-C30 scales are scored on a 4-point response scale, ranging from not at all to very much, except the last two questions, which are scored on a 7-point response scale. Statistically, the most exciting feature of QoL evaluation is considering its time-dependent structure. Whenever any patient faces a diagnosis of a fatal disease or a distorting treatment, their well-being may be affected and hence decline. In other words, it will be the process of surviving the disease and its treatment that reflects in their future QoL. Traditional clinical trials that measure the time of a fatal event, of course, take into account the time factor in treatment comparisons. Statistical analytical procedures for these tests use survival analysis methods [9].

The aim of this paper is to prepare and present R package functions that can easily work with different sub-domains of the EORTC questionnaires, both for QLQ-C30 and the cancer-specific QLQs. In addition, the presence of missing data in repeatedly measured QoL data is quite often. This package and functions are also presented to impute the valid missing observations and cover the analytical support to work with QoL data. The missing observations were imputed with the minimum value of the questions. Survival analysis is also performed for all sub-domains from the cancer-specific quality of life questionnaires.

## **2. Methodology**

The QLQ-C30 in EORTC questionnaire provides functional scales, symptoms scales, and global health status. The functional scales include five functions, they are, Physical Functioning (PF), Role Functioning (RF), Cognitive Functioning (CF), Emotional Functioning (EF), and Social Functioning (SF). The symptom scales include nine symptoms, are, Fatigue, Nausea and Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties. Each of the multi-item scales includes a different set of items, in other words, there are several

items included on every scale. All of the scales have scores ranging from 0 to 100. A high functional, symptom, and global scale score represent a healthy level of functioning, a high level of symptomatology or problems, and a high QoL, respectively. The principle of the scoring scales is the same for all the domains; that is, a linear transformation is used to standardize the raw scores [8].

The procedure for computing the domain-wise scale scores [8] is explained by scale items as  $I_1, I_2, \dots, I_n$ . For instance, in the case of QLQ-C30 the scale items will be  $I_1, I_2, \dots, I_{30}$ . The scoring manual [8] has a detailed explanation for all the domains and sub-domains.

For all scales, the **RawScores**, is the mean of the component items given as,

$$RawScore = RS = (I_1 + I_2 + \dots + I_n)/n \quad (1)$$

For **Functional Scales**, the formula for calculating scores is,

$$Score = \left[ 1 - \frac{(RS - 1)}{range} \right] \times 100 \quad (2)$$

For **Symptom scales/items** and **Global health scales/QoL**, the formula for calculating scores is,

$$Score = \left[ \frac{(RS - 1)}{range} \right] \times 100 \quad (3)$$

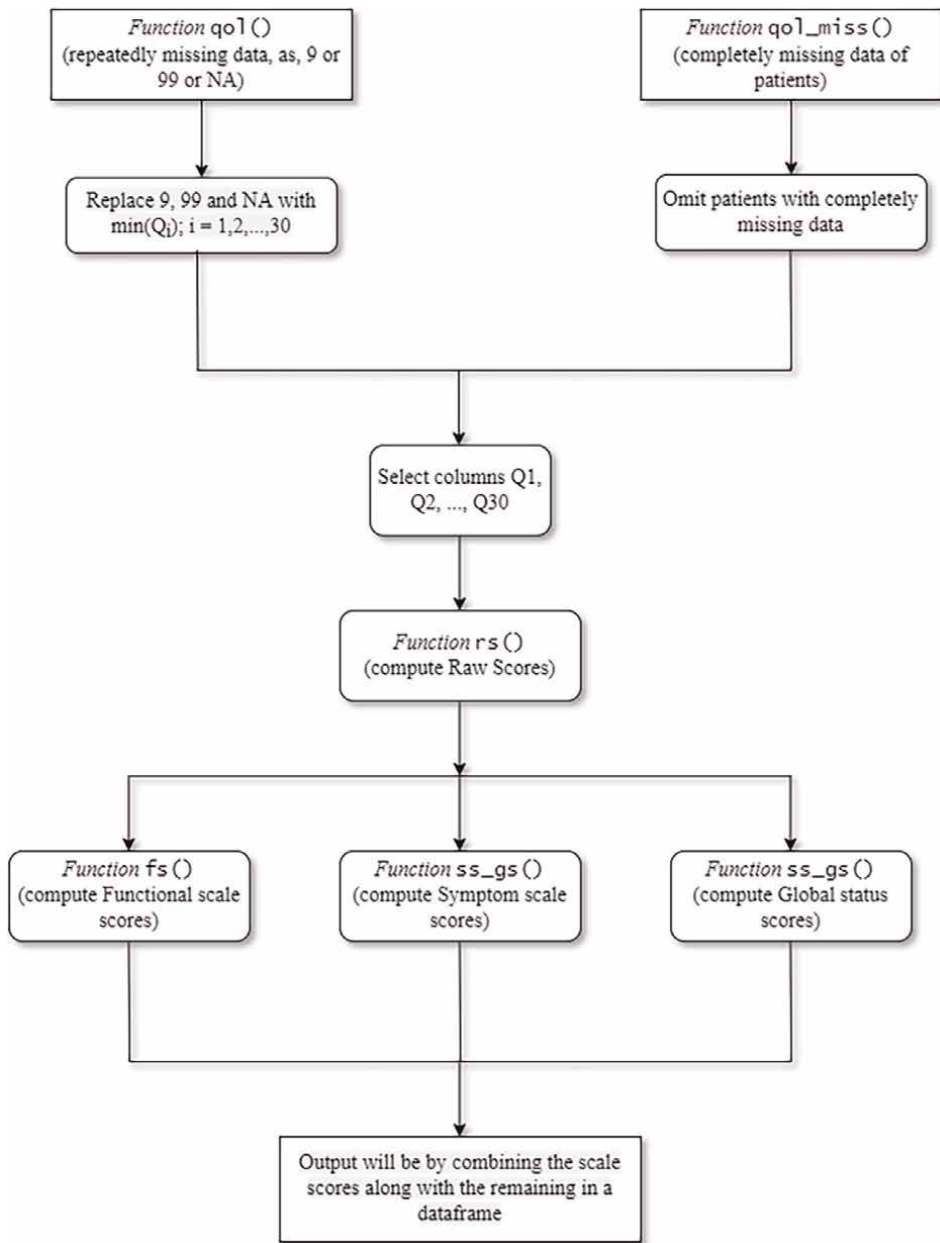
*Range* is the difference between the possible maximum and the minimum response to individual items; most items take values from 1 to 4, giving a range = 3. The QLQ-C30 has been designed so that all items on any scale take the same range of values. The exceptions are the items contributing to the global health status/QoL, which contains two items taking values from 1 to 7, giving a range = 6.

A function was developed in R using the above formulae. The purpose of this function is to convert the item-wise values into domain-wise scores and generate a comprehensive dataset.

Separate functions were prepared for Raw Score, Functional Scales Score, Global health status/QoL, and Symptom Scales, and then all these functions were collated under one single function. This function aims to take the entire data as the input and consider only those columns that contain the data of the 30 questionnaires by considering it as the revised data. Further, a nested function was formed, which contains three functions for calculating the domain-wise scale scores.

The first function `rs()` was prepared for calculating the Raw Score. Similarly, other scoring values were created. The Functional Scales Score `fs()` uses the formula of the functional scale score mentioned above for calculating the scores for all the scales under this domain. The combined function `ss_gs()` for Global health status/QoL and Symptom Scales calculates all the items under these two domains since both these scales use the same formula for calculating the scores. The three functions, `rs`, `fs`, and `ss_gs` were combined under one single function named as `qol` and `qol_miss` depending on the type of data that is used. The `qol` function can work with both complete data and data with some missing information because this function will first check for any missing information. If missing information is found, it will replace them with the minimum value. On the other hand, if no missing information is found, it will continue to calculate the scores. On the other hand, if the data contains missing

information for all the questions for a particular patient, that is, if row(s) have complete missing values then the `qol_miss` function can be used. In such a case, the row(s) with completely missing data of patients will be omitted and then the score calculations will be performed. A flowchart of the process for calculating all the scale scores using these functions available in the `QoLMiss` package is represented below (Figure 1).



**Figure 1.**  
Flowchart about creation of algorithms to calculate the domain-wise scale scores from the QoL data.

The domain-wise scale scores are also calculated from the cancer-specific questionnaires, such as, lung, head and neck, breast, ovarian, and thyroid. The functions are named as `lc_qol` for QLQ-LC13, `hnc_qol` for QLQ-HN35, `brc_qol` for QLQ-BR23, `ovc_qol` for QLQ-OV28, and `thyc_qol` for QLQ-THY34. These functions works similarly as the functions, `qol` and `qol_miss` mentioned in the above flowchart.

Another set of functions is prepared for determining the survival outcome for each and every scale scores. The hazard ratio (95% CI) is calculated for all scales, with the help of the function `coxph()` from the survival package. The values of the hazard ratio will help in understanding the survival relationship of the patients with the domain-wise scale scores.

The functions that are prepared for determining the survival relationship are named as `surv_c30`, `surv_c30_miss`, `surv_lc13`, `surv_hn35`, `surv_br23`, `surv_ov18`, and `surv_thy34`. The first step in all these functions is to divide the data according to the two arms, which will help in comparing the survival outcomes between the two arms. The different scales are considered as the covariates and univariate analysis of each of these scales is performed using the Cox-Proportional Hazard model. This analysis provides the results of the hazard ratio (95% CI) for each of the scales.

Hence the survival functions take the entire dataset as its input, provided the data consists columns such as 'time', 'event', and 'arm'. The column named 'time' should contain the survival time of the patients. The column named 'event' should contain the status of the patient, indicating with the value 0 if the patient is alive and 1 if death has occurred. Another column named 'arm' should contain the arm to which the patient has been randomized. This data is then passed to the respective QoL function for obtaining the domain-wise scale scores, which are then passed to the function `coxph()` from the survival package for obtaining the hazard ratio (95% CI). Therefore, all the prepared survival functions returns a dataframe containing the hazard ratios along with the 95% CIs for each and every scale. A flowchart of the process for performing the analysis using these functions is provided below (**Figure 2**).

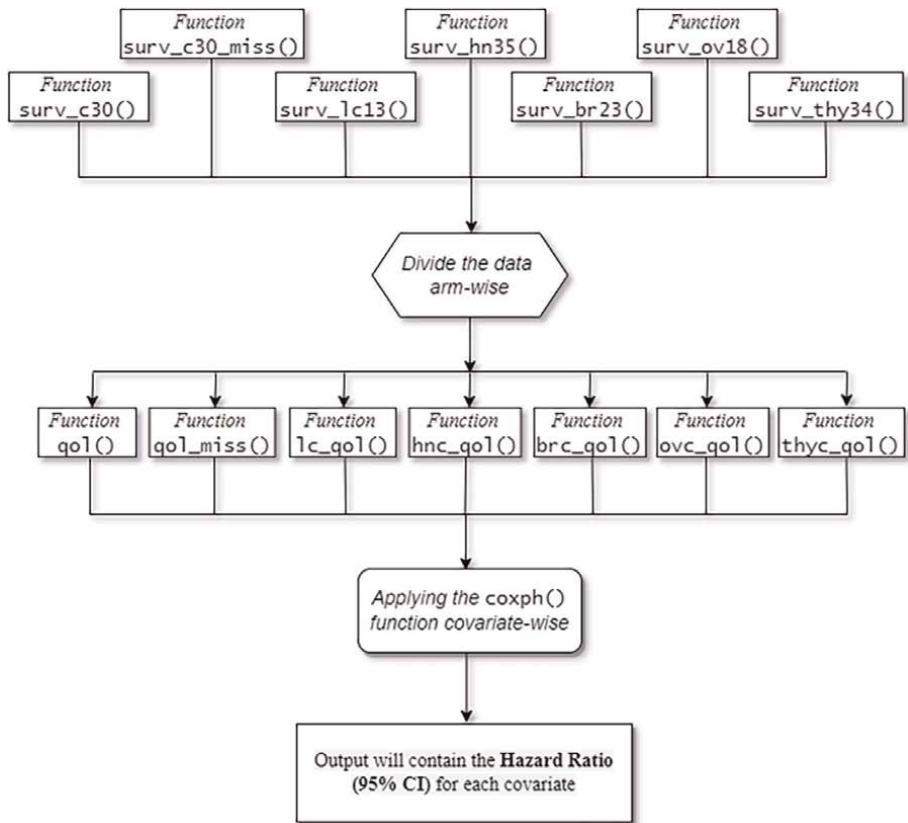
### 3. Simulation

The first step in all `qol` functions was to select only the 30 columns containing the data of the 30 questions from the complete response values.

We prefer to use simulated data and find the results and analyze from this data. So, data were simulated from the Poisson distribution with mean ( $\lambda = 2.5$ ). A random data set of size 100 was simulated for each of the 30 questions. A complete data is available considering that it contained the information of 100 patients. The only condition that is required to be checked is that the column names of the 30 questions needed to be mentioned as Q1, Q2, ..., Q30.

In some cases, it can occur that there is missing information in the data denoted as NA. Among these 30 questions, most items take values from 1 to 4, and the last two questions take values from 1 to 7, so it can be the case that the value entered is 9 or 99. Thus, another data was simulated which contained missing information for some patients, that is, the values occurred as NA or 9 or 99.

Other cases can occur that no information is obtained for any particular patient; that is, it may be obtained that for all the 30 questions, the data is available as NA. So, the third type of data was simulated in which for some patients no information was open, and the data is represented as NA.



**Figure 2.**  
*Flowchart about creation of algorithms to perform the survival analysis from the QoL data.*

For using the survival functions, a data is needed which contains three columns time to event (denoted as time), status of the patient (denoted as event) and type of treatment (denoted as arm). Therefore, the survival functions, `surv_c30`, `surv_c30_miss`, `surv_lc13`, `surv_hn35`, `surv_br23`, `surv_ov18`, and `surv_thy34`, will take a dataset as input which contains the 30 questions, time, event and arm.

#### 4. Results

After the simulated data is passed to any of the following functions `qol`, `qol_miss`, `lc_qol`, `hnc_qol`, `brc_qol`, `ovc_qol`, or `thyc_qol`, depending on the type of data, the domain-wise scores are calculated with the help of the function `fs` for Functional Scale scores, and the combined function `ss_gs` for Global health status/QoL and Symptom Scale scores. A matrix of dimension  $100 \times 30$  is obtained where the 30 domain-wise scale scores are for each of the 100 patients. These values are replaced in the data set with the 30 questions and returned as the final result.

Suppose some of the values entered in the data is NA or 9 or 99. In that case, this data will be passed to the `qol` function to check for any missing information, and if found, these values will be replaced with the minimum value of that particular question, which is generally obtained as 1. Thus, complete data will be obtained without any missing or



incorrect values and can for calculating the domain-wise scale scores. The cancer specific functions, `lc_qol`, `hnc_qol`, `brc_qol`, `ovc_qol`, and `thyc_qol` will also perform similarly as the `qol` function, depending on the cancer-specific QoL questionnaire data.

In case there is no information available for a patient, that is, the scale values are available as NA, this data will be passed to the `qol_miss` function, and the information of this particular patient will be completely ignored, in other words, the information of this patient will be removed from the data. After the required changes have been made in the data, then the domain-wise scale scores will be calculated.

For performing the univariate survival analysis considering the domain-wise scale scores as the covariates, the simulated data is passed to any of the following functions `surv_c30`, `surv_c30_miss`, `surv_lc13`, `surv_hn35`, `surv_br23`, `surv_ov18`, or `surv_thy34`, depending on the type of data. These data will again be passed to any of the `qol` functions as required for obtaining the domain-wise scale scores. These outputs are passed to the `coxph()` function from the survival package for obtaining the hazard ratios (95% CI) for each of the domain-wise scale scores.

## 5. Illustration

A simulated data was obtained from Poisson Distribution with a mean of  $2.5(=\lambda)$ . This is complete data without any missing information, so after passing the data into the `qol` function, no modifications are required. The final data frame is obtained containing the domain-wise scale scores.

Similarly, data were simulated in which there were some values were obtained as NA or 9 or 99. It is also possible to work with the `qol` function. The values NA or 9 or 99 were replaced with the minimum value for that particular question. After this modification of the data, the domain-wise scale scores were calculated, and a data frame was returned containing the domain-wise scale scores.

Lastly, the third type of data was simulated in which, for some patients, there was no information available. The information of patients was represented as NA. After passing this data into the `qol_miss` function, the information of these patients was removed from the data frame. The `qol(x)` was used to run the function and in the place of `x` the simulated data named as `c30_df` was passed as input in the `qol` function, that is, `qol(c30_df)`. A small part of output is provided below.

```
> # Load the simulated data
> data("c30_df")
> # Display head of the scale scores dataframe
> head(qol(c30_df))
```

ID	time	event	arm	QL	PF	RF	EF	CF	SF
1	498	0	2	8.33	46.67	50.00	58.33	66.67	100.00
2	91	0	1	16.67	40.00	50.00	50.00	50.00	33.33
3	13	1	1	8.33	53.33	50.00	41.67	50.00	50.00
4	707	0	2	16.67	60.00	33.33	41.67	83.33	50.00
5	993	1	2	66.67	33.33	66.67	66.67	83.33	50.00
6	23	0	1	25.00	73.33	16.67	33.33	83.33	16.67
ID	FA	NV	PA	DY	SL	AP	CO	DI	FI

1	44.44	0.00	33.33	0.00	0.00	33.33	33.33	100.00	100.00
2	55.56	16.67	83.33	66.67	0.00	33.33	0.00	100.00	100.00
3	33.33	50.00	33.33	33.33	66.67	66.67	66.67	33.33	33.33
4	11.11	0.00	50.00	66.67	0.00	33.33	0.00	66.67	66.67
5	44.44	50.00	66.67	100.00	66.67	33.33	0.00	100.00	66.67
6	66.67	50.00	16.67	0.00	66.67	0.00	33.33	100.00	33.33

The data that was simulated for testing the `qol` and `qol_miss` functions also contained three more columns containing information for time to event (denoted as time), status of the patient (denoted as event) and type of treatment (denoted as arm), which will help in illustrating the `surv_c30`, `surv_c30_miss`, `surv_lc13`, `surv_hn35`, `surv_br23`, `surv_ov18`, and `surv_thy34` functions. An illustration is given using the `surv_c30(x)` function and in the place of `x` the simulated data named as `c30_df` was passed as input in the `surv_c30` function, that is, `surv_c30(c30_df)`. The output as obtained is provided below.

```
> # Load the simulated data
> data("c30_df")
> # Display the Hazard Ratios (95% CI)
> surv_c30(c30_df)
```

	HR	Lower 95% CI	Upper 95% CI
QL	1.030	1.030	1.020
PF	1.010	1.020	0.999
RF	1.010	1.010	1.000
EF	1.010	1.010	1.010
CF	1.010	1.010	1.000
SF	1.000	1.000	1.000
FA	0.994	0.991	0.997
NV	1.010	1.020	1.010
PA	0.986	0.992	0.980
DY	1.010	1.010	1.010
SL	1.010	1.010	1.000
AP	1.010	1.000	1.010
CO	1.000	1.000	1.000
DI	0.993	0.994	0.993
FI	0.975	0.979	0.972

## 6. Discussion

The application of QoL assessment is unavoidable in cancer care [10, 11]. There is not enough ready-to-use functions for calculating the scale scores, so we prepared the method and package `QoLMiss` to work with the QoL data for cancer patients. It can

quickly provide domain-wise score computation. The implementation of the domain-wise scores in a QoL allows in finding the significant symptoms affecting the patients achieve the goal of understanding the well-being of the patients in QoL analysis in oncology clinical trials.

The QoLMiss will be updated as new modules are developed by the EORTC group. The QoLMiss package will be completed over time, by applying some Cox analyses algorithms of QoL score data. The current package will then be expanded by the addition of Cox-Proportional Hazard models. Our future endeavor will also involve in preparing functions for Bayesian Survival analysis. The package will be upgraded over time, by applying some new imputation techniques.

Further research can be performed by exploring different missing data imputation techniques in scenarios where missing data are Missing at Random (MAR), Missing not at Random (MNAR), Missing Completely at Random (MCAR) and then evaluating the domain-wise scores. Future endeavor can be to use the different scale scores to further analyze the quality of life of cancer patients. The QoLMiss package repository can be obtained using the link: <https://github.com/apstat/QoLMiss-Package>.

## Author details

Ankita Pal<sup>1\*</sup>, Satyajit Pradhan<sup>1</sup>, Aseem Mishra<sup>1</sup>, Pankaj Chaturvedi<sup>2,3</sup>  
and Atanu Bhattacharjee<sup>3,4</sup>

1 Mahamana Pandit Madan Mohan Malaviya Cancer Centre, Tata Memorial Centre, India

2 Centre for Cancer Epidemiology, Tata Memorial Centre, India


3 Homi Bhabha National Institute, Mumbai, India

4 Section of Biostatistics, Centre for Cancer Epidemiology, Tata Memorial Centre, India

\*Address all correspondence to: [apalstat97@gmail.com](mailto:apalstat97@gmail.com)

## IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Adler NE, Page AEK. Cancer Care for the Whole Patient: Meeting Psychosocial Health Needs. Washington, DC, USA: National Academies Press/Institute of Medicine (US) Committee on Psychosocial Services to Cancer Patients/Families in a Community Setting; 2008
- [2] Olschewski M, Schumacher M. Statistical analysis of quality of life data in cancer clinical trials. *Statistics in Medicine*. 1990;**9**:749-763
- [3] Wood-Dauphinee S, Troidl H. Endpoints for clinical studies: Conventional and innovative variables. In: Troidl H, Spitzer WO, McPeck B, Mulder DS, McKneally MF, editors. *Principles and Practice of Research: Strategies for Surgical Investigators*. New York: Springer; 1986. pp. 53-68
- [4] Carr AJ, Gibson B, Robinson PG. Is quality of life determined by expectations or experience? *BMJ*. 2001; **322**(7296):1240-1243. DOI: 10.1136/bmj.322.7296.1240
- [5] Aaronson NK, Cull A, Kaasa S. Sprangers MAG for the EORTC Study Group on Quality of Life. The European Organization for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: An update. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2nd ed. New York, NY, USA: Raven Press; 1996. pp. 179-189
- [6] Reale ML, De Luca E, Lombardi P, Marandino L, Zichi C, Pignataro D, et al. Quality of life analysis in lung cancer: A systematic review of phase III trials published between 2012 and 2018. *Lung Cancer*. 2019;**139**(2020):47-54
- [7] Jitender S, Mahajan R, Rathore V, Choudhary R. Quality of life of cancer patients. *Journal of Experimental Therapeutics & Oncology*. 2018;**12**(3): 217-221
- [8] EORTC QLQ-C30 Scoring Manual. 3rd ed. Brussels: EORTC; 2001. ISBN: 2-9300 64-22-6
- [9] Cox DR, Oakes D. *Analysis of Survival Data*. London: Chapman and Hall; 1984
- [10] Nayak MG, George A, Vidyasagar MS, et al. Quality of life among cancer patients. *Indian Journal of Palliative Care*. 2017;**23**(4):445-450. DOI: 10.4103/IJPC.IJPC\_82\_17
- [11] Hassen AM, Taye G, Gizaw M, Hussien FM. Quality of life and associated factors among patients with breast cancer under chemotherapy at Tikur Anbessa specialized hospital, Addis Ababa, Ethiopia. 2019. DOI: 10.1371/journal.pone.0222629





*Edited by B. Santhosh Kumar*

This one-of-a-kind volume explores the most recent advancements in biostatistics and its applications. Among the chapters included are both original contributions and reviews of the most recent advancements in the field. Written by researchers from the pharmaceutical industry, universities, and the research and development divisions of the government, this book offers a balanced representation of the research community. It provides advanced graduate students and new researchers with a complete scholarly assessment of several research frontiers in biostatistics and will help these readers make significant contributions to the development of the subject by exploring new approaches and making discoveries.

Published in London, UK

© 2024 IntechOpen

© Tetiana Lazunova / iStock

**IntechOpen**

