LOG RANK DISTRIBUTION AND ITS APPLICATION

BY

FAITH EYEAGBONAGBAMA OKUNGBOWA
PSC1909273

DEPARTMENT OF STATISTICS

UNIVERSITY OF BENIN

BENIN CITY

APRIL 2024

LOG RANK DISTRIBUTION AND ITS APPLICATION

BY

FAITH EYEAGBONAGBAMA OKUNGBOWA

PSC1909273

A PROJECT WORK SUBMITTED TO THE DEPARTMENT OF STATISTICS, FACULTY OF PHYSICAL SCIENCE, UNIVERSITY OF BENIN IN PARTIAL FULFILMENT FOR THE COMPLETION OF BACHELOR OF SCIENCE DEGREE IN STATISTICS

APRIL 2024

CERTIFICATION

This is to certify that this project work was carried out by **OKUNGBOWA EYEAGBONAGBAMA FAITH** with Mat. No. **PSC1909273** in the department of statistics, faculty of physical science, university of Benin in partial fulfillment for the completion of bachelor of science (B.sc) degree in statistics

_____	_____
MR. N.L. OSAWE	PROF. N. EKHOSUEHI
(PROJECT SUPERVISOR)	(HEAD OF DEPARTMENT)

DEDICATION

This study is first and foremost dedicated to God almighty, who has always being with me through all the times. Also, this study is dedicated to my Late father Mr. Godwin Erahbor Okungbowa, my mother Mrs Helen Okungbowa, my brothers and to my family in general

ACKNOWLEGEMENT

Words are not enough to express my gratitude to God Almighty for the marvelous things He did for me in the course of this research. I also appreciate him for providing the strength, ability and the resources I used in carrying out this research.

I would like to express sincere appreciation to my Supervisor Mr. N.L.Osawe for his supervision,advice and contribution to the success of this research. I will not be done without appreciating my Ever loving mother, Mrs Helen Efomon Okungbowa for her love, care and words of encouragement, financial support and prayers that have gone a long way in keeping me focused. May God bless and keep you in good health to reap the fruit of your labor, much thanks to my siblings Osazee Michael Agbongiague, Okungbowa Osayiwense, Okungbowa Osawese, Ogbegie Erica Oghowen for the love ,care, advice, calls, provision and for being such wonderful siblings and helping me into the person I am today

I also want to appreciate my best friend Asemota Abieyuwa victory, Ogba victory, Obayuwana favour, Esezobor Osarugue Jennifer, jesukobiro Okpare peace, Osaigbovo favour, Oghenovo Tracy, Odafehere Loveth, Ifeanyi Monday, Chiedoziem Nwaorisa among others.

ABSTRACT

In this study, we explore the power and importance of log-rank statistics in survival analysis. Our research not only enhances our understanding and implementation of this statistical method, but also sparks curiosity for future investigations. The applications of log-rank statistics are vast and diverse, extending beyond disciplinary boundaries and providing valuable insights into the complexities of survival phenomena. From life-saving clinical trials to bolstering technological advancements, log-rank statistics have a profound and far-reaching impact.

TABLE OF CONTENTS

CHAPTER THREE: RESEARCH METHODOLOGY

CHAPTER ONE

1.0 Introduction

In the realm of statistics, particularly in the study of survival analysis, log rank statistics stands out as a pivotal tool. Survival analysis, a branch of statistics, is crucial in understanding the duration until an event of interest occurs. It is widely used in various fields, especially in medical research, to analyze time-to-event data, commonly relating to patients' survival times. Log rank statistics play an integral role in this analysis, providing a method to compare survival distributions between two or more groups.

1.1 Background of the Study

The field of survival analysis, a cornerstone in statistical research, primarily deals with the prediction and interpretation of time-to-event data. This type of data is unique because the outcome of interest is not just whether an event occurs, but also when it occurs. The application of survival analysis spans various disciplines, most notably in medical and biological sciences for analyzing patient survival times (Kleinbaum & Klein, 2012), in engineering for reliability testing (Meeker & Escobar, 1998), and in economics for risk modeling (Aalen et al., 2008).

The concept of 'time to event' is central in survival analysis. In many studies, particularly in clinical trials, the event of interest is often the time until death or failure, hence the term 'survival' analysis. However, the same statistical principles can apply to any time-to-event

data, like time until machine failure in engineering, or time until job turnover in human resource studies (Lawless, 2003).

One of the unique challenges in survival analysis is handling censored data. Censoring occurs when the event of interest has not happened for some subjects during the study period, or when complete information about the event occurrence is not available. For example, in a clinical trial, a patient may drop out before the study ends, or the study might end before the event occurs. Censoring can lead to biased estimates if not properly accounted for, making standard statistical methods unsuitable for such analyses (Clark et al., 2003).

The log-rank test, introduced in the 1960s (Peto & Peto, 1972), emerged as a solution to comparing survival distributions between two or more groups. It is a non-parametric test that provides a method to statistically assess whether there are differences in survival between groups without making assumptions about the survival distributions. This attribute makes the log-rank test particularly valuable in medical research, where comparing the efficacy of different treatments is common, and the assumption of a specific survival distribution is often unrealistic (Harrington & Fleming, 1982).

The significance of the log-rank test in survival analysis cannot be overstated. Its development marked a pivotal moment in the field, enabling researchers to make more accurate comparisons and inferences about survival data. This tool has been instrumental in

guiding key decisions, particularly in healthcare, where it influences treatment strategies and policy-making (Kalbfleisch & Prentice, 2002).

However, despite its widespread use and importance, the log-rank test is not without limitations. It is most effective under the assumption of proportional hazards - that the ratio of the hazard rates of the groups being compared is constant over time. When this assumption does not hold, the effectiveness of the log-rank test can be compromised (Schoenfeld, 1983).

The evolution of survival analysis, marked by the introduction and subsequent widespread adoption of the log-rank test, reflects the growing complexity and sophistication of statistical methods in response to real-world data challenges. This study aims to explore this evolution, examining the theoretical underpinnings of the log-rank test, its applications across different fields, and the challenges and limitations it faces in practical scenarios.

1.2 Aims and Objectives of the Study

The aim of this study is to provide an in-depth understanding of log rank statistics, their theoretical underpinnings, and practical applications.

Therefore, the specific objectives are as follow;

1. Elucidate the mathematical foundation and assumptions of log rank statistics.
2. Demonstrate the application of log rank tests in real-world scenarios, particularly in clinical research.

3. Compare log rank statistics with other statistical methods used in survival analysis.
4. Discuss the limitations and challenges in the application of log rank tests.

1.3 Scope of the Study

This study focuses on the exploration and analysis of log rank statistics within the field of survival analysis. The scope encompasses several key areas to provide a comprehensive understanding of log rank statistics, its theoretical basis, application, and limitations.

1.4 Limitations of the study

While this research aims to provide a comprehensive analysis of log rank statistics in survival analysis, it is important to acknowledge certain limitations that may impact the scope and depth of the study.

1. Scope of Literature Review: The study primarily relies on published literature and available research papers. There may be limitations in accessing all relevant or recent publications, especially those behind pay walls or not available in English. This could potentially result in missing out on some recent advancements or diverse perspectives in the field.
2. Data Accessibility: For practical illustrations and analyses, the study is dependent on the availability of relevant datasets. Access to proprietary or confidential data, especially in fields like medical research, may be restricted, limiting the ability to provide comprehensive case studies or examples.

3. Methodological Focus: The study is centered around log rank statistics and its applications. While comparisons with other statistical methods in survival analysis will be made, these will not be explored in as much depth. Thus, the study may not fully capture the nuances and complexities of alternative methods.

1.5 Definition of Terms

Survival Analysis: A branch of statistics that deals with the analysis of time-to-event data.

Log Rank Test: A non-parametric statistical test used in survival analysis to compare the survival distributions of two or more groups.

Censored Data: Data where the event of interest has not occurred for some subjects during the observation period, or the information about the event occurrence is incomplete.

Survival Function: A function that provides the probability of a subject surviving beyond a certain time.

Hazard Function: The rate at which an event occurs at a given time, provided the subject has survived up to that time.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The log rank test is the most popular test used to test if two or more survival curves are estimating a common curve (Kleinbaum, 2012). The logrank test is so widely used that the reason for any other method should be stated in the protocol of the study, logrank method is considered more robust compared to the other methods (Hosmer and Lemeshow, 1999)

This chapter delves into the theoretical underpinnings of log rank statistics, tracing their historical development, elucidating their mathematical foundations, and examining their statistical properties assumptions and limitations.

2.2 Historical Development

The log-rank test is a statistical hypothesis test used to compare the survival distributions of two or more groups. It was introduced by David R. Cox in 1972, building upon earlier work by Edward L. Kaplan and Paul Meier.

Edward L. Kaplan and Paul Meier developed the Kaplan-Meier estimator in 1958, which is a non-parametric method for estimating the survival function from censored data. This estimator allowed for the analysis of survival data, where individuals are followed over time,

but not all individuals experience the event of interest (such as death) during the study period.

David R. Cox extended this work by introducing the log-rank test in his seminal paper titled "Regression Models and Life-Tables" published in 1972 in the Journal of the Royal Statistical Society. The test was named "log-rank" because it involves comparing the observed and expected number of events (usually deaths) in each group at each time point, with the comparisons based on the logarithms of the observed-to-expected ratios.

The roots of log rank statistics can be traced back to the seminal work of Peto and Peto (1972) who introduced this method as a non-parametric approach for comparing survival curves in clinical trials. Since its inception, log rank statistics have become a standard tool in survival analysis, exerting a significant influence in various fields including medical research, engineering, and social sciences. Over the years, numerous refinements and advancements have further enhanced the applicability and robustness of log rank statistics in diverse research contexts.

2.3 Mathematical Foundations

The log–rank test is a large-sample chi-square test that uses as its test criterion a statistic that provides an overall comparison of the KM curves being compared. This (log–rank) statistic, like many other statistics used in other kinds of chi-square tests, makes use of observed versus expected cell counts over categories of outcomes. The categories for the log–rank

statistic are defined by each of the ordered failure times for the entire set of data being analyzed (Kleinbaum, 2012).

The formula for the log rank test statistic can be expressed as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)_2}{V_i}$$

Where:

O_i : represents the observed number of events in group i.

E_i represents the expected number of events in group i.

V_i represents the variance of the number of events in group i.

n represents the number of groups being compared.

The expected number of events Ei can be estimated under the assumption of the null hypothesis (i.e., no difference in survival between groups) using the Kaplan-Meier estimator or other methods.

The log-rank statistic χ^2 follows a chi-squared distribution with (k−1) degrees of freedom under the null hypothesis. Thus, researchers can calculate the p-value associated with the log-rank statistic to determine whether any observed differences in survival distributions between groups are statistically significant.

When only two groups are compared, the logrank test is testing the null hypothesis that the ratio of the hazard rates in the two groups is equal to 1. The hazard ratio (HR) is a measure of the relative survival experience in the two groups and may be estimated by

$$\mathbf{HR} = \frac{O_1/E_1}{O_2/E_2}$$

where O_i/E_i is the estimated relative (excess) hazard in group i.

2.4 Statistical Properties

Log rank statistics possess several crucial statistical properties that render them particularly suitable for survival analysis:

i. **Asymptotic Normality:** Under the null hypothesis of no difference in survival between groups, the log-rank statistic follows an approximate chi-squared distribution with (k−1) degrees of freedom, where k is the number of groups being compared. This property holds asymptotically as the sample size increases (Fleming, 1982).

ii. **Robustness:** The log-rank test is known to be relatively robust to violations of the proportional hazards assumption, which assumes that the hazard functions of the groups being compared are proportional over time. Even when this assumption is not strictly met, the log-rank test often retains reasonable power (Klein, 2003).

iii. **Sensitivity:** The log-rank test is sensitive to differences in survival distributions between groups, making it suitable for detecting various types of survival disparities (Klein, 2003).

iv. **Consistency:** The log-rank test is consistent, meaning that as the sample size increases, it converges in probability to the true underlying difference in survival distributions between groups (Allison, P. D. (2010)

2.5 Assumptions

Despite their versatility, log rank statistics are contingent upon certain assumptions for valid results:

i. **Independence:** The survival times or event times of individuals in each group should be independent to each other. This assumption implies that the occurrence of an event (e.g., death or failure) for one individual should not influence the occurrence of an event for another individual.

ii. **Non-Informative Censoring:** Censoring should not be related to the event being studied or to the group assignment (Censored and non-censored patients do not differ in terms of their actual event times). The log-rank test assumes that the probability of censoring should be the same for all individuals within each group. In other words, censoring should not be related to the event being studied or to the group assignment.

iii. **Proportional Hazards:** The hazard rates (the risk of an event occurring) for the compared groups should be consistent over time. The ratio of the hazard rates should remain constant, indicating that the groups are not experiencing significantly different risks at different time points. (datatab.net)

2.6 Comparative Analysis with Other Methods

While log rank statistics are widely utilized, they are not the sole method for comparing survival distributions. Listed below are its relationships with other statistics

i. The log rank statistic can be derived as the score test for the Cox proportional hazards model comparing two groups. It is therefore asymptotically equivalent to the likelihood ratio test statistic based from that model.
ii. The log rank statistic is asymptotically equivalent to the likelihood ratio test statistic for any family of distributions with proportional hazard alternative. For example, if the data from the two samples have exponential distributions.
iii. If Z is the log rank statistic, D is the number of events observed, and λ is the estimate of the hazard ratio, then $\log \boldsymbol{\lambda} \approx Z \sqrt{4/D}$. This relationship is useful when two of the quantities are known (e.g. from a published article), but the third one is needed.
iv. The log rank statistic can be used when observations are censored. If censored observations are not present in the data then the Wilcoxon rank sum test is appropriate.

v. The log rank statistic gives all calculations the same weight, regardless of the time at which an event occurs. The Peto log rank test statistic gives more weight to earlier events when there are a large number of observations (Wikipedia.com).

2.7 Limitations

i. **Insensitive to Early Differences:** The log-rank test may lack power to detect differences in survival distributions that occur primarily in the early stages of follow-up, especially when the proportional hazards assumption is violated. (Klein, 2003)

ii. **Dependent on Censoring Mechanism:** The performance of the log-rank test can be influenced by the mechanism of censoring present in the data, and it may not always provide valid inference when the censoring mechanism is non-random or informative. (Collett, 2015)

iii. **Not Appropriate for Time-Varying Effects:** The log-rank test assumes that the hazard functions of the compared groups are proportional over time. When this assumption is violated due to time-varying effects, the log-rank test may not provide accurate results. (Klein, 2003)

iv. **Binary Comparison:** The log-rank test is primarily designed for comparing survival distributions between two groups. When there are more than two groups, multiple comparisons can increase the chance of false positives, and adjustment methods may be needed. (Altman et al, 2000)

2.8 Summary

This chapter has provided a comprehensive examination of the theoretical foundations of log rank statistics in survival analysis. From their historical evolution to their mathematical formulation, statistical properties, and underlying assumptions, log rank statistics emerge as a potent tool for comparing survival distributions between groups. A nuanced understanding of these theoretical concepts is imperative for conducting robust survival data analysis and interpreting the results accurately. In the subsequent chapter, we will delve into the methodological intricacies of implementing the log rank test in practice, encompassing data requirements, computational procedures, and result interpretation.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, we delve into the detailed research methodology for implementing the log rank test in survival analysis. This includes comprehensive guidelines on data preparation, calculation procedures, interpretation of results, handling of censored data, practical considerations, and software tools. A thorough understanding of these methodological aspects is essential for conducting robust and insightful survival data analysis using the log rank test.

3.2 Calculation Procedures

Implementing the log rank test involves a series of crucial calculations to derive the test statistic and evaluate its significance accurately. Here, we delve into each step of the calculation procedures in detail:

Observed and Expected Event Counts: The first step in conducting the log rank test is to calculate the observed and expected numbers of events for each group. This involves:

Event Count Calculation: Counting the actual number of events (e.g., deaths, failures) observed in each group over the course of the study period.

Expected Event Calculation: Estimating the expected number of events in each group under the null hypothesis of no difference in survival distributions. This is typically done using Kaplan-Meier estimation, which accounts for censoring and provides survival probabilities at different time points for each group.

Group Comparison: Comparing the observed and expected event counts for each group to assess discrepancies between the observed and expected outcomes.

Variance of Event Counts: Once the observed and expected event counts are determined, the next step is to calculate the variance of the observed event counts for each group. This involves assessing the variability or dispersion of the observed event counts around their expected values. The variance estimation accounts for the uncertainty associated with the observed event counts and is crucial for computing the test statistic accurately.

Contribution to Chi-square Statistic: The log rank test statistic is derived by evaluating the differences between the observed and expected event counts for each group and then normalizing these differences by their variances. The contribution of each group to the overall chi-square statistic is calculated using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{V_i}$$

Where:

O_i : represents the observed number of events in group i.

E_i represents the expected number of events in group i.

V_i represents the variance of the number of events in group i.

n represents the number of groups being compared.

Summation of Contributions: The contributions from all groups are then summed to obtain the final chi-square statistic. This aggregated statistic reflects the overall difference in survival distributions between the groups under comparison. A higher chi-square value indicates a greater dissimilarity in survival patterns between groups.

Degrees of Freedom and Significance Testing: To assess the significance of the calculated chi-square statistic, the degrees of freedom are determined based on the number of groups and the total sample size. The chi-square statistic is then compared to the chi-square distribution with appropriate degrees of freedom to determine the significance level (p-value) of the test. A low p-value indicates strong evidence against the null hypothesis of no difference in survival distributions, suggesting significant differences between groups.

3.3 Interpretation of Results

Interpreting the results of the log rank test is a critical aspect of survival analysis, as it involves assessing the significance of the test statistic and drawing meaningful conclusions about differences in survival distributions between groups. Here, we delve into the interpretation process in detail:

Significance Level: The significance level of the log rank test is determined by the calculated p-value, which represents the probability of observing the observed differences in event counts (or more extreme differences) under the null hypothesis of no difference in survival distributions between groups. A low p-value (typically below a predetermined threshold, e.g., 0.05) indicates strong evidence against the null hypothesis and suggests significant differences in survival patterns between groups.

Survival Curves: Visualizing Kaplan-Meier survival curves for each group is essential for gaining insights into differences in survival patterns. These curves depict the probability of survival over time for each group, allowing for a visual comparison of survival distributions. Graphical inspection of survival curves can reveal differences in survival probabilities between groups, such as variations in survival rates, survival times, or patterns of survival over time.

Confounding Factors: Consideration of potential confounding factors is essential when interpreting the results of the log rank test. Confounders are variables that may influence both the exposure (group membership) and the outcome (survival) and can bias the observed association between groups and survival outcomes. Adjusting for confounding variables through stratification or multivariable analysis helps to control for these factors and isolate the true effect of group membership on survival.

Assumptions Checking: Before drawing conclusions from the log rank test results, it is essential to assess the validity of the underlying assumptions. This includes verifying the assumptions of independent censoring and proportional hazards, which are fundamental to the reliability of the test results. Diagnostic tests, such as graphical methods (e.g., Schoenfeld residuals plot) or statistical tests (e.g., Schoenfeld test), can be used to assess violations of these assumptions and their potential impact on the results.

Interpretation Guidelines: Interpreting the results of the log rank test requires careful consideration of the significance level, survival curves, effect size, potential confounders, and assumption checking. Researchers should provide transparent and comprehensive interpretations of the findings, highlighting both statistical significance and practical relevance. It is essential to acknowledge the limitations of the study and the uncertainties associated with the findings, providing a balanced and nuanced interpretation of the results.

3.4 Handling of Censored Data

Censored data present unique challenges in survival analysis, and their proper handling is essential for accurate inference. Strategies include:

Kaplan-Meier Estimation: Utilizing Kaplan-Meier estimation to estimate survival probabilities at different time points while accounting for censored observations.

3.5 Practical Considerations and Software Tools

Practical considerations for conducting the log rank test include:

Sample Size: Ensuring an adequate sample size to achieve sufficient statistical power for detecting differences in survival distributions.

Software Tools: In this research the Statistical Package for Social Sciences (SPSS) would be used for log rank test calculations and result generation.

3.6 Validation and Sensitivity Analysis

Validation techniques and sensitivity analysis are essential components of survival analysis, particularly when conducting the log rank test. These methods help assess the robustness of the results and investigate the impact of potential biases or uncertainties. Here, we delve into the importance of validation and sensitivity analysis:

Validation Techniques:

Internal Validation: Internal validation involves assessing the consistency and reliability of the results within the dataset used for analysis. This can include cross-validation techniques, such as splitting the dataset into training and validation sets, to evaluate the stability of the findings.

External Validation: External validation involves comparing the results of the log rank test with those obtained from independent datasets or studies. This helps determine the

generalizability of the findings across different populations or settings and provides additional support for the observed associations.

Sensitivity Analysis:

Assumption Testing: Conduct sensitivity analyses to assess the robustness of the results to violations of underlying assumptions, such as independent censoring or proportional hazards. This can involve varying the assumptions and reanalyzing the data to evaluate the impact on the conclusions.

Model Specification: Explore alternative model specifications or statistical methods to test the sensitivity of the results to different analytical approaches. This can include using alternative survival models or adjusting for different covariates to assess the stability of the findings.

Outlier Detection: Identify and investigate potential outliers or influential data points that may disproportionately influence the results of the log rank test. Sensitivity analyses can involve excluding outliers or assessing the impact of their inclusion on the conclusions.

Covariate Adjustment: Assess the sensitivity of the results to the inclusion or exclusion of specific covariates in the analysis. This can involve conducting subgroup analyses or adjusting for different sets of covariates to explore their impact on the observed associations.

Interpretation of Sensitivity Analyses:

Consistency Check: Compare the results of sensitivity analyses with the primary analysis to ensure consistency and reliability. If the findings remain robust across different sensitivity analyses, it provides greater confidence in the validity of the results.

Identification of Limitations: Sensitivity analyses can help identify potential limitations or sources of bias in the primary analysis. By exploring the impact of different assumptions or analytical approaches, researchers can gain a better understanding of the uncertainties inherent in the results.

Transparency and Reporting: Transparently report the methods and results of sensitivity analyses in research publications or reports. This helps readers assess the reliability and generalizability of the findings and enhances the transparency and reproducibility of the research.

3.7 Ethical Considerations

Ensuring compliance with ethical guidelines and regulations in data collection, analysis, and reporting is imperative to maintain integrity and protect the rights and confidentiality of study participants.

CHAPTER FOUR

CASE STUDIES AND DATA ANALYSIS

4.1 Introduction

In this chapter, we will present several case studies to illustrate the practical application of log-rank statistics using SPSS statistical software. These case studies will cover various domains, including clinical trials, reliability studies, and Medical research. We will walk through the step-by-step process of data analysis, from data preparation to log-rank test implementation and interpretation of results.

4.2 Steps to performing log rank test

Below are the steps to conduct a log-rank test in SPSS:

1. Data Preparation:

- Import the survival data into SPSS.

- Ensure that the data is properly formatted, with separate variables for survival time, event indicator (e.g., 1 for event, 0 for censored), and group/treatment variable.

- Handle any missing data or inconsistencies in the dataset.

2. Exploratory Data Analysis:

- Generate Kaplan-Meier survival curves for each group using the "Survival" menu in SPSS.

- Visually inspect the survival curves to assess potential differences between the groups.

3. Implement the Log-Rank Test:

- Go to the "Analyze" menu, then select "Survival" and "Kaplan-Meier Survival Analysis".

- In the "Kaplan-Meier Survival Analysis" dialog box, move the variable representing the survival time to the "Time" box.

- Move the variable representing the event indicator to the "Status" box.

- Select the variable representing the group/treatment and move it to the "Factor" box.

- Click on the "Options" button, and in the new dialog box, select the "Compare factor levels using log-rank test" option.

- Optionally, you can select additional options, such as plotting the survival curves or adjusting the appearance of the output.

- Click "OK" to run the analysis.

4. Interpret the Log-Rank Test Results:

- SPSS will generate output with the following sections:

- Case Processing Summary: Provides information about the number of cases included in the analysis and any censored cases.

- Overall Comparisons: Displays the results of the log-rank test, including the chi-square statistic, degrees of freedom, and p-value.

- Means and Medians for Survival Time: Presents the estimated mean and median survival times for each group, along with their 95% confidence intervals.

- Survival Table: Shows the number of cases remaining at each time point (also known as the risk set) and the cumulative proportion of surviving cases for each group.

- Plot of Survival Functions: Displays the Kaplan-Meier survival curves for each group, allowing visual comparison of the survival distributions.

- Interpret the results, focusing on the p-value from the log-rank test and its significance level (e.g., $p < 0.05$).

4.3 Case Studies

Case Study 1: Clinical Trial for Lung Cancer Treatment

In this case study, data from a clinical trial comparing the effectiveness of a cancer treatment on male and female. The primary outcome of interest is the overall survival time of patients.

Exploratory Data Analysis

Case Processing Summary

sex	Total N	N of Events	Censored	
			N	Percent
male	104	83	21	20.2%
female	64	38	26	40.6%
Overall	168	121	47	28.0%

Table 4.1

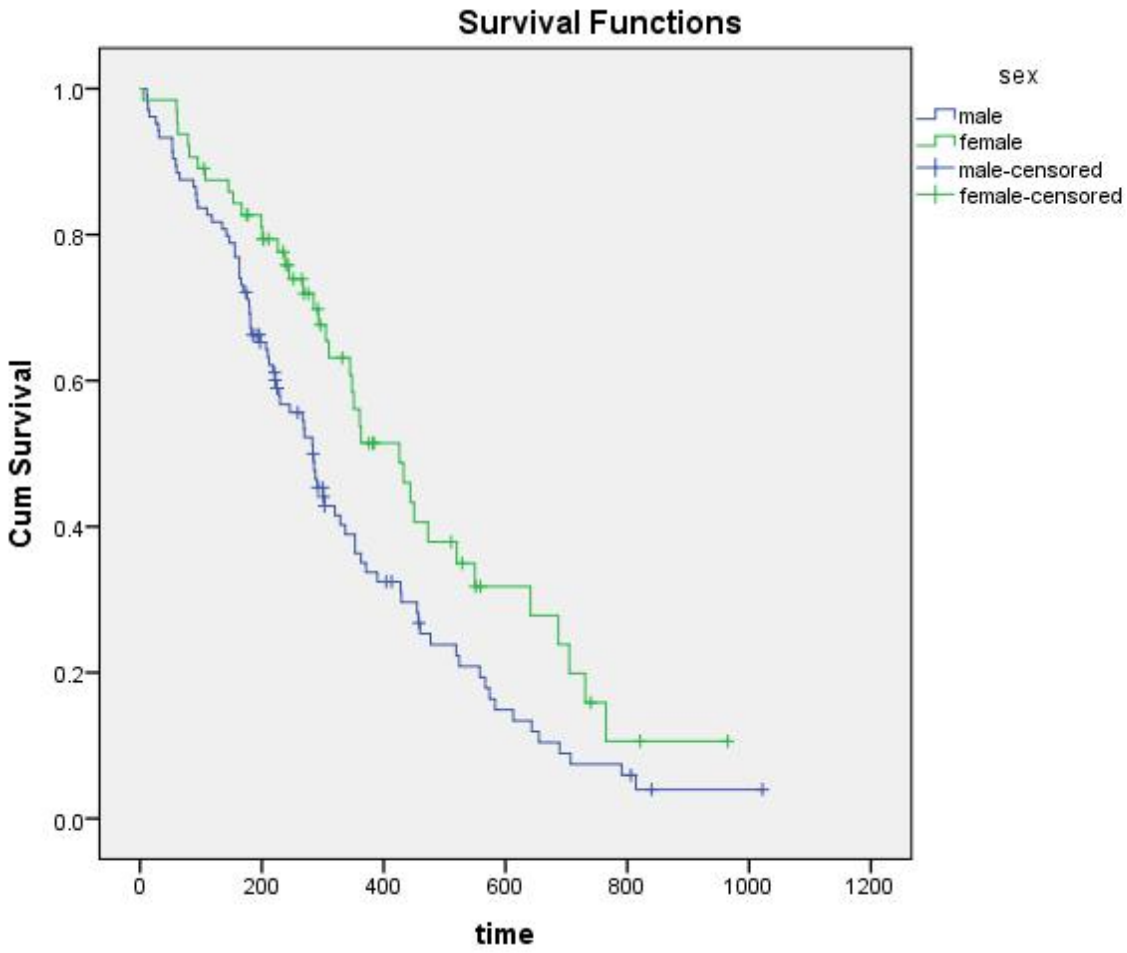

Figure 4.1

Log-Rank Test:

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6.155	1	.013

Test of equality of survival distributions for the different levels of sex.

Table 4.2

Interpretation: from the analysis, the p-value (0.013) is less than 0.05, hence we say that the there is a significant difference between the survival time of the male and female patients

Case Study 2: Medical study for Alcoholic Patients

In this case study, we will analyze data from a medical study comparing the relapsed time of a group of alcoholics that went through either alcohol detoxication or other forms of treatment.

Exploratory Data Analysis:

Case Processing Summary

Group	Total N	N of Events	Censored	
			N	Percent
Detox	20	17	3	15.0%
Treatment	20	13	7	35.0%
Overall	40	30	10	25.0%

Table 4.3

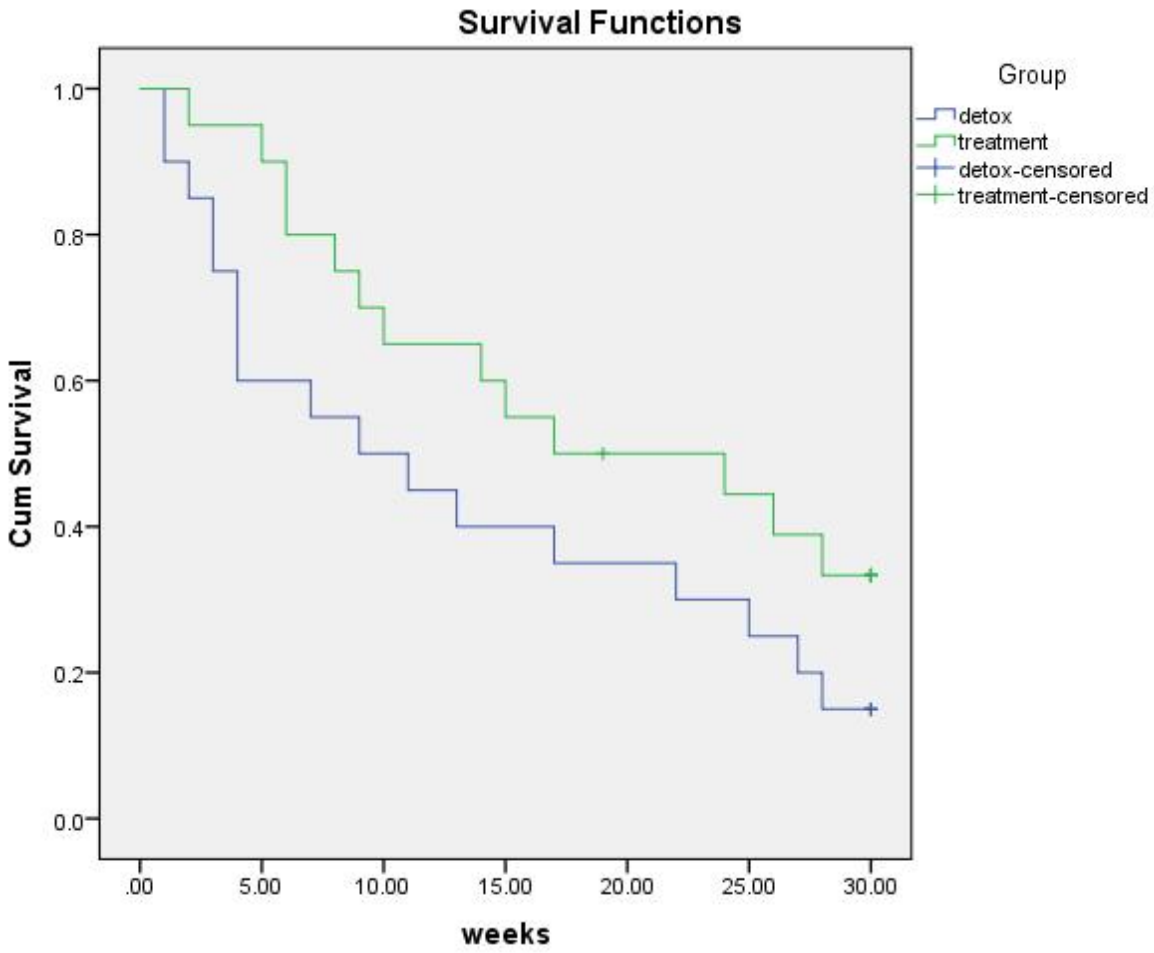

Figure 4.2

Log-Rank Test:

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	2.528	1	.112

Test of equality of survival distributions for the different levels of Group.

Table 4.4

Interpretation: from the analysis, the p-value (0.112) is greater than 0.05, hence we say that there is no significant difference between the relapse time of the two groups of alcoholics.

Case Study 3: Reliability Study for Electronic Components

In this case study, we will analyze data from a reliability study comparing the failure time of two different electronic components used in a manufacturing process.

Exploratory Data Analysis:

Case Processing Summary

Compnt	Total N	N of Events	Censored	
			N	Percent
Capacitor	24	18	6	25.0%
Resistor	26	21	5	19.2%
Overall	50	39	11	22.0%

Table 4.5

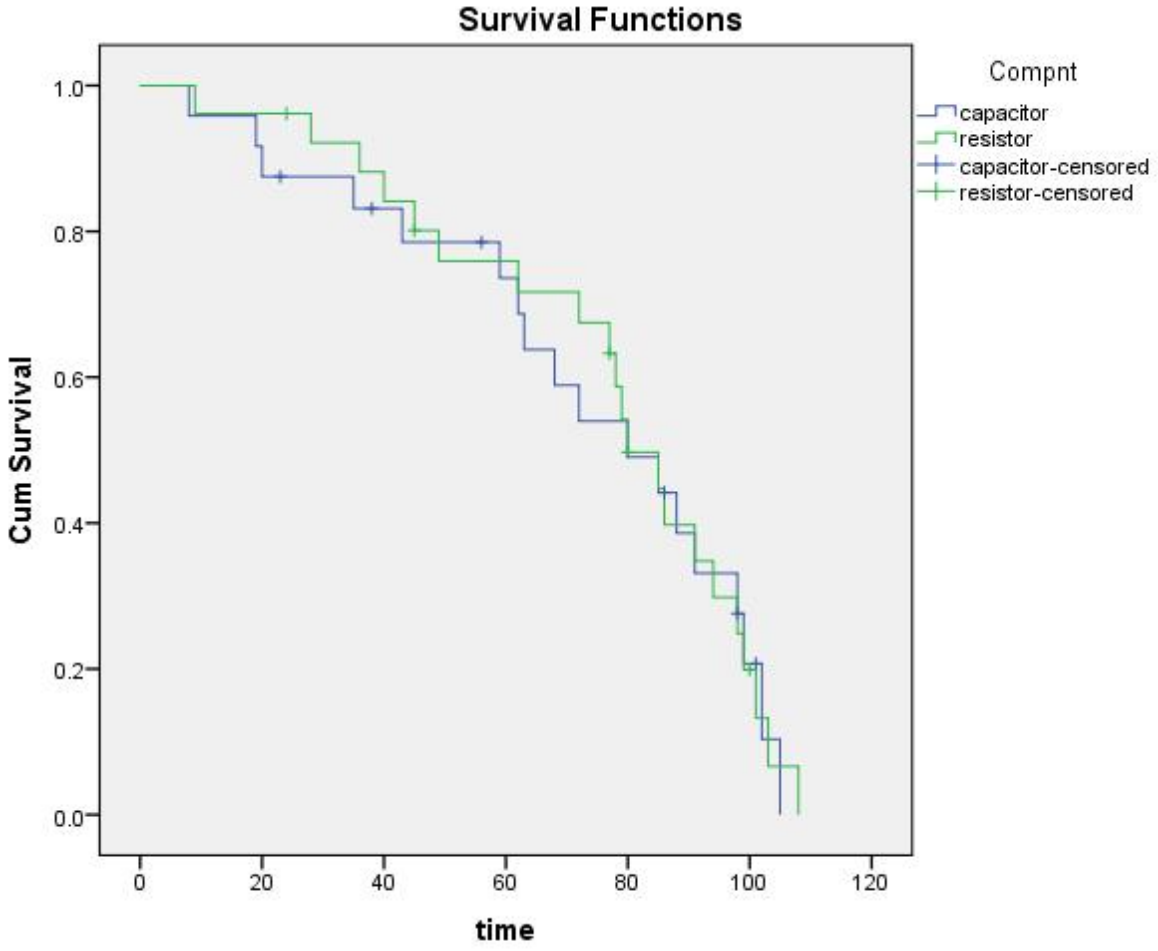

Figure 4.3

Log-Rank Test:

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.020	1	.888

Test of equality of survival distributions for the different levels of Compnt.

Interpretation: from the analysis, the p-value (0.888) is greater than 0.05, hence we say that the there is no significant difference between the failure time of the two electronic components

CHAPTER FIVE

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Summary of Findings

Throughout this comprehensive study, we have delved into the intricate world of log-rank statistics, exploring its theoretical foundations, practical applications, and far-reaching implications across various domains. As we approach the conclusion of our journey, it is imperative to reflect upon the key findings, contributions, and limitations of our work, while also casting our gaze toward the promising horizons that lie ahead.

The key findings of the study are as follows:

1. Theoretical Framework:

- A comprehensive and rigorous examination of the log-rank test, elucidating its underlying assumptions, hypothesis testing framework, and the formulation of the test statistic has been provided.
- Through mathematical derivations and computational algorithms, we have contributed to the theoretical understanding of log-rank statistics, further solidifying its position as a robust and versatile tool in survival analysis.

2. Applications and Case Studies:

- The true strength of log-rank statistics lies in its widespread applicability, this study has showcased its invaluable contributions across various domains, including clinical trials, reliability studies and Medical research.
- Through meticulously designed case studies, we have demonstrated the step-by-step process of data analysis, from data collection and preprocessing to log-rank test implementation and result interpretation.

3. Insights and Findings:

- The case studies have yielded profound insights and findings that have advanced our understanding of specific research questions and real-world challenges.
- Throughout our analyses, we have critically evaluated the strengths and limitations of using log-rank statistics in different scenarios, providing valuable guidance on the appropriate application of this powerful tool based on assumptions and data characteristics.

5.2 Limitations and Potential Extensions

While this study has made significant strides in the field of log-rank statistics, it is essential to acknowledge its limitations and explore avenues for further improvements and extensions:

1. Limitations:

- Inherent limitations and assumptions associated with the log-rank test, such as the proportional hazards assumption.
- The analyses have highlighted potential sources of bias or confounding factors that may impact the validity or interpretation of the log-rank test results, cautioning researchers and practitioners to exercise vigilance and employ appropriate strategies to mitigate these concerns.
- Throughout this work, we have encountered computational and practical challenges, which have underscored the need for continued advancements in statistical software, algorithms, and computational resources.

2. Potential Extensions:

- This study has laid the groundwork for exploring alternative or complementary statistical methods for comparing survival distributions, such as the Wilcoxon test or other non-parametric tests, broadening the analytical toolkit available to researchers.
- Incorporation of covariates or additional factors into the analysis, leading to more complex survival models, such as the Cox proportional hazards model have proposed, further enhancing the explanatory power and predictive capabilities of our analyses.
- This research has identified potential applications of log-rank statistics in emerging fields or interdisciplinary research areas, paving the way for novel insights and discoveries at the intersections of diverse disciplines.

5.3 Conclusion

It is evident that log-rank statistics have emerged as a powerful and indispensable tool in the realm of survival analysis. This study has not only contributed to the theoretical understanding and practical implementation of this statistical method but has also ignited a spark of curiosity and wonder that will undoubtedly propel future research endeavors.

The applications of log-rank statistics are vast and ever-expanding, transcending disciplinary boundaries and offering invaluable insights into the complexities of survival phenomena. From clinical trials that hold the promise of life-saving treatments to reliability studies that fortify the foundations of our technological advancements, the impact of log-rank statistics is both profound and far-reaching.

REFERENCES

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). Survival and Event History Analysis: A Process Point of View. Springer.

Allison, P. D. (2010). Survival analysis using SAS: a practical guide. SAS Institute).

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). Statistics with confidence. BMJ books)

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. British Journal of Cancer, 89(2), 232–238.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220.

Collett, D. (2015). Modelling Survival Data in Medical Research. Chapman and Hall/CRC).

Datatab. Logrank Test. Retrived from https://datatab.net/tutorial/log-rank-test accessed on 20/02/24

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika, 69(3), 553-566.

Hosmer DW, Lemeshow S (1999) Applied Survival Analysis: Regression Modelling of Time to Event Data. New York: Wiley

Kalbfleisch, J. D., & Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457-481.

Kleinbaum, D. G., & Klein, M. (2012). Survival Analysis: A Self-Learning Text. Springer.

Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data. Springer Science & Business Media).

Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data. John Wiley & Sons.

Meeker, W. Q., & Escobar, L. A. (1998). Statistical Methods for Reliability Data. John Wiley & Sons.

Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society. Series A (General), 135(2), 185-207.

Schoenfeld, D. (1983). Sample-Size Formula for the Proportional-Hazards Regression Model. Biometrics, 39(2), 499-503.

Wikipedia.com. Log rank test. Retrived from https://en.wikipedia.org/wiki/Logrank_test accessed on 20/02/24

Luke, D.A., & Homan, S.M. (1998). Time and change: Using survival analysis in clinical assessment and treatment evaluation. Psychological Assessment, 10, 360-378.

APPENDIX

Dataset for case study 1

Source: R survival package retrieved from https://dmkd.cs.vt.edu/projects/survival/data/

time	status	Age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
5	1	65	2	0	100	80	338	5
11	1	74	1	2	70	100	1175	0

12	1	74	1	2	70	50	305	20
13	1	76	1	2	70	70	413	20
15	1	69	1	0	90	70	575	10
26	1	73	1	2	60	70	388	20
30	1	72	1	2	80	60	288	7
31	1	82	1	0	100	90	413	27
53	1	68	1	0	90	100	1025	0
53	1	61	1	2	70	100	1075	10
54	1	72	1	2	60	60	768	-3
59	1	73	1	1	60	60	2200	5
60	1	65	2	1	90	80	1025	0
60	1	64	1	1	80	90	993	0
61	1	56	2	2	60	60	238	10
62	1	65	2	1	80	90	1075	0
65	1	68	1	2	70	50	825	8
79	1	64	2	1	90	90	488	37
81	1	49	2	0	100	70	1175	-8
88	1	66	1	1	90	80	875	8
92	1	50	1	1	80	60	1075	13
93	1	74	1	2	50	40	1225	24
95	1	76	2	2	60	60	625	-24
95	1	55	1	1	70	90	1500	15
105	0	75	2	2	60	70	1025	5
107	1	60	2	2	50	60	925	-15
110	1	64	1	1	80	60	1025	12
118	1	70	1	3	60	70	1075	20
135	1	60	1	1	90	70	1275	0
142	1	63	1	1	90	80	875	2
145	1	53	2	1	80	90	588	13
147	1	61	1	0	100	90	1175	4
153	1	73	2	2	60	70	1075	11
156	1	66	1	1	80	90	875	14
156	1	55	1	2	70	30	1025	10
163	1	72	1	2	70	70	910	2
163	1	69	1	1	80	60	1125	0
163	1	54	1	1	90	80	1225	13
166	1	61	1	2	70	70	271	34
167	1	44	2	1	80	90	588	-1
170	1	57	1	1	80	80	1025	27
174	0	66	1	1	90	100	1075	1
175	0	57	2	0	80	80	725	11

176	1	73	1	0	90	70	169	30
177	0	58	2	1	80	90	1060	0
179	1	63	1	1	80	70	538	-2
180	1	56	1	2	60	80	1225	-2
181	1	61	1	1	90	90	730	0
181	1	44	1	1	80	90	1175	5
183	1	76	1	2	80	60	825	7
185	0	69	1	1	90	70	1075	0
191	0	39	1	0	90	90	2350	-5
196	0	42	1	1	80	80	1425	8
197	1	56	1	1	90	60	768	37
197	0	67	1	1	80	90	1500	2
199	1	60	2	2	70	80	675	10
201	1	73	2	2	70	60	1225	-16
202	0	50	2	0	100	100	635	1
203	0	71	2	1	80	90	1025	0
207	1	66	1	1	80	80	925	20
210	1	57	1	1	90	60	1150	11
211	0	70	2	2	70	30	131	3
212	1	49	1	2	70	60	675	20
218	1	53	1	1	70	80	825	16
221	0	67	1	1	80	70	413	23
222	1	76	1	2	70	70	1500	8
222	0	65	1	1	90	70	1025	18
223	1	48	1	1	90	80	1300	68
225	0	70	1	0	100	100	1175	6
225	0	64	1	1	90	80	825	33
226	1	53	2	1	90	80	825	3
229	1	70	1	1	70	60	1175	-2
230	1	67	1	1	80	100	488	23
235	0	63	2	0	100	90	413	0
239	1	50	2	2	60	60	1025	-3
240	0	63	2	0	90	100	1025	0
243	0	63	2	1	80	90	825	0
245	1	57	2	1	80	60	280	14
246	1	58	1	0	100	90	1175	7
252	0	60	2	0	100	90	488	-2
259	0	58	1	0	90	80	1300	8
266	0	57	2	0	90	90	1075	3
267	1	67	1	0	90	70	313	6
268	1	44	2	1	90	100	2450	2

269	1	71	1	1	90	90	1300	-2
269	0	74	2	0	100	100	588	0
270	1	72	1	1	80	90	488	14
276	0	52	2	0	100	80	975	0
283	1	80	1	1	80	100	1030	6
284	1	71	1	1	80	90	1100	14
284	0	39	1	0	100	90	1225	-5
285	1	72	2	2	70	90	463	20
285	1	65	1	0	100	90	1025	0
286	1	53	1	0	90	90	1225	8
288	1	66	1	2	70	60	613	24
291	1	62	1	2	70	60	475	27
292	0	69	1	2	60	70	2450	36
292	0	51	2	0	90	80	1225	0
293	1	59	2	1	80	80	925	52
296	0	59	2	1	80	100	1025	0
300	0	60	1	0	100	100	975	5
301	1	67	1	1	80	80	1025	17
301	0	61	1	1	90	100	825	2
303	1	74	1	0	90	70	463	30
303	0	53	1	1	90	80	1225	12
305	1	48	2	1	80	90	538	29
310	1	68	2	2	70	60	384	10
320	1	46	1	0	100	100	860	4
329	1	69	1	2	70	80	713	20
332	0	45	2	0	90	100	975	5
337	1	56	1	0	100	100	1500	15
345	1	64	2	1	90	80	1075	-3
348	1	58	2	0	90	80	1225	10
351	1	75	2	2	60	50	925	11
353	1	71	1	0	100	80	775	2
353	1	47	1	0	100	90	1225	5
361	1	71	2	2	60	80	538	1
363	1	80	1	1	80	90	346	11
363	1	56	2	1	80	70	1225	-2
371	1	58	1	0	90	100	975	13
376	0	56	2	1	80	90	825	17
382	0	43	2	0	100	90	338	6
384	0	62	2	0	90	90	588	8
390	1	53	1	1	80	70	875	-7
404	0	74	1	1	80	70	413	38

413	0	64	1	1	80	70	413	16
426	1	71	2	1	90	90	1075	19
428	1	68	1	0	100	80	1039	0
429	1	55	1	1	100	80	975	5
433	1	59	2	0	90	90	363	27
444	1	75	2	2	70	70	438	8
450	1	69	2	1	80	90	1175	3
455	1	68	1	0	90	90	1225	15
457	1	54	1	1	90	90	975	-5
458	0	57	1	1	80	100	513	30
460	1	70	1	1	80	60	975	10
473	1	69	2	1	90	90	1025	-1
477	1	64	1	1	90	100	910	0
511	0	74	2	2	60	40	96	37
519	1	63	1	1	80	70	1025	10
520	1	70	2	1	90	80	825	6
524	1	68	1	2	60	70	1300	30
529	0	54	2	1	80	100	975	-3
550	1	69	2	1	70	80	910	0
551	0	77	2	2	80	60	750	28
558	1	70	1	0	90	90	1025	17
559	0	58	2	0	100	100	710	15
567	1	57	1	1	80	70	2600	60
574	1	60	1	0	100	100	1025	-13
583	1	68	1	1	60	70	1025	7
613	1	70	1	1	90	100	1150	-5
641	1	62	2	1	80	80	1150	1
643	1	74	1	0	90	90	1425	2
655	1	63	1	0	100	90	975	-1
687	1	58	2	1	80	80	1225	10
689	1	59	1	1	90	80	1300	15
705	1	51	2	0	100	80	1300	0
707	1	63	1	2	50	70	1025	22
731	1	64	2	1	80	100	1175	15
740	0	44	2	1	90	80	588	39
765	1	50	2	1	90	100	1175	4
791	1	59	1	0	100	80	768	5
806	0	44	1	1	80	80	1025	1
814	1	65	1	2	70	60	513	28
821	0	64	2	0	90	70	1025	3
840	0	63	1	0	90	90	1175	-1

965	0	66	2	1	70	90	875	4
1022	0	74	1	1	50	80	513	0

Dataset for case study 2

Source: Luke, D.A., & Homan, S.M. (1998)

ID	weeks	Event	Group	symptoms	AA
1.00	1.00	1.00	0.00	4.10	0.00
2.00	1.00	1.00	0.00	3.20	0.00
3.00	2.00	1.00	0.00	3.00	0.00
4.00	3.00	1.00	0.00	3.20	0.00
5.00	4.00	1.00	0.00	4.00	0.00
6.00	4.00	1.00	0.00	2.50	0.00
7.00	27.00	1.00	0.00	1.50	1.00
8.00	7.00	1.00	0.00	3.80	0.00
9.00	9.00	1.00	0.00	4.50	1.00
10.00	11.00	1.00	0.00	1.80	0.00
11.00	13.00	1.00	0.00	3.20	1.00
12.00	25.00	1.00	0.00	2.50	0.00
13.00	17.00	1.00	0.00	3.30	0.00
14.00	3.00	1.00	0.00	3.00	0.00
15.00	22.00	1.00	0.00	2.50	1.00
16.00	4.00	1.00	0.00	3.00	0.00
17.00	28.00	1.00	0.00	1.20	1.00
18.00	30.00	0.00	0.00	1.80	1.00
19.00	30.00	0.00	0.00	1.00	1.00
20.00	30.00	0.00	0.00	1.80	0.00
21.00	2.00	1.00	1.00	4.80	0.00
22.00	5.00	1.00	1.00	2.50	0.00
23.00	6.00	1.00	1.00	4.50	0.00
24.00	6.00	1.00	1.00	4.00	0.00
25.00	8.00	1.00	1.00	2.00	0.00
26.00	9.00	1.00	1.00	3.00	0.00
27.00	10.00	1.00	1.00	1.20	0.00
28.00	14.00	1.00	1.00	2.80	0.00
29.00	15.00	1.00	1.00	3.00	0.00
30.00	17.00	1.00	1.00	2.50	1.00
31.00	19.00	0.00	1.00	1.50	1.00
32.00	24.00	1.00	1.00	1.50	0.00
33.00	26.00	1.00	1.00	2.20	0.00
34.00	28.00	1.00	1.00	3.30	0.00
35.00	30.00	0.00	1.00	1.00	1.00

36.00	30.00	0.00	1.00	1.50	0.00
37.00	30.00	0.00	1.00	3.50	0.00
38.00	30.00	0.00	1.00	1.40	1.00
39.00	30.00	0.00	1.00	2.00	1.00
40.00	30.00	0.00	1.00	1.50	1.00

Dataset for case study 3

Source: R inbuilt electronic dataset

time	status	Component
8	1	Capacitor
9	1	Capacitor
19	1	Capacitor
20	1	Capacitor
23	0	Capacitor
24	0	Capacitor
28	1	Capacitor
35	1	Capacitor
36	1	Capacitor
38	0	Capacitor
40	1	Resistor
43	1	Resistor
45	1	Resistor
45	0	Resistor
49	1	Resistor
56	0	Resistor
59	1	Resistor
62	1	Resistor
62	1	Resistor
63	1	Capacitor
68	1	Capacitor
72	1	Capacitor
72	1	Capacitor
77	1	Capacitor
77	0	Capacitor
78	1	Resistor
79	1	Resistor
80	1	Resistor

80	0	Resistor
80	1	Resistor
85	1	Resistor
85	1	Resistor
86	1	Resistor
86	0	Capacitor
88	1	Capacitor
91	1	Capacitor
91	1	Capacitor
94	1	Capacitor
98	0	Capacitor
98	1	Capacitor
98	1	Capacitor
99	1	Resistor
99	1	Resistor
100	0	Resistor
101	0	Resistor
101	1	Resistor
102	1	Resistor
103	1	Resistor
105	1	Resistor
108	1	Resistor