Multiple Comparison (Time to event data)

Aaron Akyea Mensah

Department of Mathematical Sciences
Montana State University

May 9th, 2024

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Aaron Akyea Mensah

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Prof. John Borkowski
Writing Project Advisor

Date

Dr. Katharine Banner
Writing Projects Coordinator

Abstract

In recent times, researchers are increasingly interested in comparing multiple treatments across multiple groups, often necessitating the performance of multiple hypothesis tests simultaneously. However, this approach can lead to multiplicity, resulting in an inflated overall type I error rate.

This project addresses this challenge by examining two non-parametric tests for comparing two or more survival curves: the logrank test and the Peto-Peto extension of the Wilcoxon Rank Sum test. The logrank test assigns more weight to later survival times, while the Peto-Peto extension assigns more weight to earlier survival times.

Additionally, we investigate five distinct p-value adjustment methods applied to these two tests to help control the overall type I error rates in multiple hypothesis testing. These methods include the Bonferroni, Holm, Hochberg, Hommel, and Benjamini-Hochberg procedures. Each method exhibits varying degrees of conservatism or liberalism, influencing their respective effectiveness in controlling type I error rates. Our investigation aims to understand how these adjustment methods interact with the non-parametric test in controlling the overall type I error rate while maintaining statistical power.

To this effect, we perform a numerical example to illustrate the application of the 10 distinct combinations control the type I error rates. In the final phase of our study, we conduct a simulation study generating random survival data from the exponential distribution to test the power of these combinations. Our goal is to investigate which methods exhibit significant power in accurately detecting differences in survival curves, considering different mean combinations and sample sizes.

Contents

1 Introduction

1.1 Background

Survival analysis is a statistical model used to analyze time-to-event data, where the ”event” refers to a specific occurrence of interest Klein and Moeschberger (2003). This statistical methodology finds applications in various fields and goes by different names, such as reliability analysis in engineering, duration analysis in economics, and event history analysis in sociology. A common characteristic of this type of data is that it often involves censoring Kaplan and Meier (1958). Censoring takes place when we possess partial information about individuals in the study, but we lack the precise time when the event of interest occurred Schober and Vetter (2018). There are three fundamental types of censoring;

- Right censoring occurs when an individual exits the study before its conclusion or has not yet experienced the event of interest by the end of the study Klein and Moeschberger (2003). In this case, if an individual departs the study at time t we anticipate that the event of interest may have occurred at time t or at some point in the indefinite future Klein and Moeschberger (2003).

- Left Censoring: In this case, individuals have already experienced the event of interest before the study begins Klein and Moeschberger (2003). The event precedes the observation period.

- Interval Censoring: Interval censoring is observed when the event of interest happens within a specific time interval. We know it occurred within this range, but we lack precise information about when it occurred Klein and Moeschberger (2003).

In survival analysis, it is preferable for censoring to be non-informative, indicating that participants leave the study for reasons unrelated to the study itself. Informative censoring on the other hand takes place when participants are lost to follow-up due to reasons directly related to the study Klein and Moeschberger (2003). There are also other types of censoring; Type 1 censoring, which involves individuals dropping out of a study in a random manner Klein and Moeschberger (2003). In this form of censoring, the duration of the study is predetermined and fixed by the researcher Klein and Moeschberger (2003). It is worth noting that, in such studies, the number of uncensored observations is a random variable. The second type, often referred to as Type 2 censoring, takes place when the study has no predefined endpoint. In this scenario, the study continues until a specific number of the events of interest have been observed Klein and Moeschberger (2003).

In survival analysis, the time variable describes the duration in years, months, weeks, or days from the commencement of follow-up until the occurrence of a specific event Klein and Moeschberger (2003) or the conclusion of the study. This time variable is referred to as the survival time (t). The event of interest, which we aim to observe, is usually termed "failure" Klein and Moeschberger (2003) In the context of modeling survival data using the semi-parametric approach, we typically do not assume that the data follows a known distribution. Instead, we employ methods that allow us to model such data. In this project, we adopt the approach introduced by Kaplan and Meier(1958) Kaplan and Meier (1958).

To set the stage for their methods, we begin by defining some key variables. Let T represent the time until an event of interest occurs. Additionally, we define a variable d to represent the number of individuals who experience the event of interest at a particular time point. Lastly, we introduce a variable n to represent the number of individuals at risk of experiencing the event of interest at a specific time point. In this context, two important functions come into play. First, there's the survival function, denoted as $s(t)$, which describes the probability that an individual will survive beyond a certain time t Klein and Moeschberger (2003). Mathematically given as $S(t) = Pr(T > t)$. This function is characterized as a probability function that equals one at time zero and approaches zero as time extends to infinity. It is a monotone, non-decreasing function, from the beginning of the study to its conclusion Klein and Moeschberger (2003).

The hazard function, denoted as $h(x)$, is also a crucial function. It represents the probability of the event of interest occurring in the next instant, given that the event has not already occurred before that time Klein and Moeschberger (2003). The hazard function is a non-negative, monotone, non-decreasing function, and its mathematical expression is given as:$h(x) = \frac{P(t \leq T < t+\delta | t \leq T)}{\delta}$. In this formula, t represents the current time, and δ is a small increment in time.

In medical research among other fields, researchers are mostly interested in comparing the effectiveness of an intervention across multiple groups Gamel and Vogel (1997). They want to know if there are significant differences between any pairs of survival distributions. Several tests have been proposed in literature to aid in achieving this comparison. The most popular among them is the log-rank test which assigns equal weights to survival times in its computation Schober and Vetter (2021). This project compares the results from the log-rank test to another test known as the Peto-Peto test which assigns higher weights to early survival times Peto and Peto (1972). These two tests help us to better understand survival curves as the Kaplan Meier curves just gives us a rough visual comparison of two survival distributions without telling us if they are significantly different or not Kaplan and Meier (1958). These tests help us to compare multiple survival distributions to know which pairs are significantly different. This allow us to conduct multiple hypothesis testing for multiple survival curves simultaneously. However, when conducting multiple hypothesis tests simultaneously, the likelihood of committing a Type I error increases as the number of comparisons increases Lydersen (2021). The family-wise error rate and the false discovery rate approaches 1 as the number of hypothesis increases but ideally we want to bound the probability of committing any type 1 error by a pre-defined significance level Bretz, Hothorn, and Westfall (2016).

Multiple testing procedures provides means of overcoming this issue of multiplicity by adjusting p-values for effects estimates Lydersen (2021). For this project, we will consider the Bonferroni method Bretz et al. (2016), Holm procedure Holm (1979), Hochberg procedure Hochberg (1988), Hommel procedure and, the Benjamini and Hochberg procedure Benjamini and Hochberg (1995) in mitigating issues associated with multiplicity in multiple hypothesis testing.

Statistical power which is the probability of correctly rejecting all false null hypotheses, and is significantly influenced by the use of multiple testing procedures in multiple hypothesis testing *Estimating Statistical Power When Using Multiple Testing Procedures* (2017). This factor directly affects the probability of detecting an effect when it exists. There exists a trade-off between reducing the likelihood of detecting a true effect through adjustment and increasing the risk of a false positive when adjustments are not made Maxwell, Kelley, and Rausch (2008)Zhang and Gou (2016).

1.2 Estimators Of The Survival and Cumulative Hazard Functions

A graphical representation of the probability of survival over a designated time period known as survival curves, are constructed based on observed survival times and censored data. The utilization of survival analysis techniques, such as the Kaplan-Meier method, facilitates the comparison of survival curves, offering statistical measures to understand the significance of observed differences.

1.2.1 The Kaplan Meier Method (KM)-Product-Limit Estimator

The KM method is a very popular method used to model survival data Rich et al. (2010). It is used in calculating survival probabilities. The KM method calculates the probability that an event will occur beyond a certain time given the data observed Machin, Cheung, and Parmar (2006). This probability can be represented graphically as a survival curve, showing the proportion of individuals surviving at each time point Machin et al. (2006).The curve begins at 1 and declines over time. The size of each step is determined by both the number of events and the number of individuals at risk D'Arrigo et al. (2021). The method involves calculating probabilities based on the observed survival times, considering the event occurrence or censoring at each time interval Altman (1992). These probabilities are then multiplied to get the overall survival probability. The KM method assumes that censoring is non-informative, meaning the probability of being censored is independent of the probability of experiencing the event of interest D'Arrigo et al. (2021). This is an important assumption for accurate estimation.

Mathematically, the Kaplan-Meier estimator is given as:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Where;

$\hat{S}(t)$ is the estimated survival probability at time t.

t represents a specific time point.

t_i are the distinct observed event times.

d_i is the number of events of interes that occurred at time t_i.

n_i is the number of individuals "at risk" at time t_i

Kaplan-Meier curves serve as valuable tools to assess differences between two or more groups in survival analysis. These curves illustrate how survival probabilities evolve over time; the farther away the curve is from the vertical axis, the higher the survival probability of that group. When comparing curves, the greater the deviation from the vertical axis relative to other curves, the higher the probability of survival. However, it is essential to perform statistical tests as an additional measure to ensure that observed differences are not merely due to chance.

Several non-parametric tests are commonly employed for this purpose, including the logrank test Schober and Vetter (2021), Gehan Wilcoxon test Hazra and Gogtay (2017), Tarone-Ware test Tarone (1981), Peto-Peto test Peto and Peto (1972), and the Fleming-Harrington test Fleming and Harrington (1991), among others. For this project, we focus on two tests: the logrank test and the Peto-Peto test. This tests are commonly used and have different approaches to determining significant difference between survival probabilities.

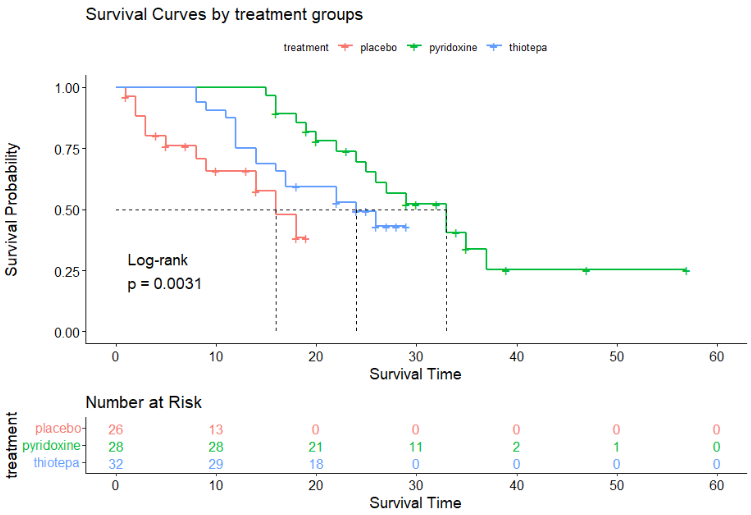


Figure 1: Kaplan-Meier curve for three treatment groups in a bladder cancer treatment.Utkarshx27 (2022)

By employing these tests, we can better understand the significance of observed differences between survival curves. This enhances the reliability of our findings and ensures appropriate conclusions in survival analysis studies.

1.3 Objectives

- To investigate two non-parametric tests for comparing multiple survival curves.
- Investigate the performance of different p-value adjustment methods.
- Run a simulation study to compare the impact of five p-value adjustment techniques on type 1 error rates and statistical power when analyzing survival data with multiple comparisons.

2 Methodology

2.1 Non-Parametric

2.1.1 Logrank Test

The logrank test is a statistical method used to assess whether significant differences exist between the survival curves of different groups Peto and Peto (1972)Schober and Vetter (2018). It is widely regarded as one of the most commonly employed tests for comparing survival curves Schober and Vetter (2021). In this test, the null hypothesis assumes that there are no differences in survival distributions of between the groups being compared. It operates under the assumption of proportional hazards, where weights are equally distributed for different survival times. It is important to note that the logrank test evaluates the entire curve rather than focusing on the survival probability at a specific time point Schober and Vetter (2021). While the test determines statistical significance, it does not estimate an effect size. For estimating effect sizes, other tests such as the Cox proportional hazards model are typically used. The logrank test statistic is approximately distributed as a Chi-square test statistic with degrees of freedom equal to the number of groups being compared minus one Schober and Vetter (2021).

2.1.2 Peto and Peto Test

The Peto-Peto test is an extension of the Wilcoxon sum rankPeto and Peto (1972). This test enables the Wilcoxon rank sum test to adapt and handle the uncertainty associated with censored data. The test suggests a scoring system that accounts for the uncertainty introduced by censoring. It assigns higher scores or weights to early survival times.

Weights decrease as the survival time increases. The Peto-Peto test statistic is also approximately distributed as a Chi-squared test statistic with degrees of freedom equal to the number of groups being compared minus onePeto and Peto (1972).

Let $\chi^2_{(n-1)}$ be the Chi-squared test statistic with $k-1$ degrees of freedom;

$$\chi^2_{(n-1)} = \sum_{j=1}^{k}\sum_{i=1}^{n} w_i \frac{(O_{ij}-E_{ij})^2}{Var_j}$$

$j = 1, ..., k$ groups where

- O_{ij} is the observed number of events in Group j at time i.
- E_{ij} is the expected number of events in Group j at time i.
- Var_j is the variance of difference between the observed and expected number of events for group i

$O_{ij} = m_{ij}$ and $E_{ij} = \frac{n_{i1}}{n_{i1}+n_{i2}} \times (m_{i1}+m_{i2})$

For comparing two groups;

$$Var_i = \sum_{i=1}^{2} \frac{n_{i1}n_{i2}(m_{i1}+m_{i2})(n_{i1}+n_{i2}-m_{i1}-m_{i2})}{(n_{i1}+n_{i2})^2(n_{i1}+n_{i2}-1)}$$

Where;

- m_{i1} is the number of failures in group 1 at time i
- m_{i2} is the number of failures in group 2 at time i
- n_{i1} is the number at risk in group 1 at time i

- n_{i2} is the number at risk in group 2 at time i

To calculate the test statistic for the two methods, The log-rank test assign a weight $w_i = 1$ and the Peto-Peto test assigns a weight $w_i = \prod(1 - \frac{d_i}{n_i+1})$.

2.2 Type of errors

There are two common error rates associated with any hypothesis test problem. We have the type 1 error rate and the type 2 error rates. While the former represents false positives, the latter represent false negatives. Assume there are m null hypothesis to be tested. The table below summarizes the type 1 and type 2 errors associated with any test.

summary of the type 1 and type 2 errors associated with any hypothesis test			
Hypothesis	Not Rejected	Rejected	Total
True	U	V	m_o
False	T	S	$m - m_o$
Total	W	R	m

Table 1: A tables summarizing the type 1 and type 2 errors in multiple hypothesis testing.

- R denotes the number of rejected hypothesis
- V denotes the number of type 1 errors.
- R, W and m are observable random variable while S, T, U and V are unobservable random variables.

In univariate hypothesis testing ($m = 1$), a test is chosen such that the type 1 error rate is maintained at a pre-defined significance level. Extension of this idea into multiple hypothesis problems are possible. Some of there error rates are defined therein;

2.2.1 Family-wise Error Rates(FWER)

This is the probability of committing at least one type one error Bretz et al. (2016). Mathematically, it is defined as;

$$FWER = P(V > 0)$$

Where V is defined as in Table 1.

As the number of hypotheses increases, we can extend the FWER to allow for the probability of committing at least k type 1 errors. This is called the generalized Family-Wise error rate, and given as;

$$gFWER = P(V > k)$$

.

2.2.2 False Discovery Rate(FDR)

The FDR is the expected proportion of falsely rejected hypotheses among the rejected hypotheses Bretz et al. (2016). Mathematically, it is given as:

$$FDR = E(Q)$$

where $Q = \frac{V}{R}$ for $R > 0$ and 0 otherwise.

where V and R are as defined in Table 1.

2.2.3 Per-Comparison Error Rate(PCER)

The PCER is the expected proportion of type 1 errors among m decisions Bretz et al. (2016). Mathematically, it is defined as:

$$PCER = \frac{E(V)}{m}$$

where V and m are defined as in Table 1.

In general, a multiple comparison procedure that controls the FWER also controls the FDR and the PCER but not vice versa Bretz et al. (2016). That is;

$$PCER \leq FDR \leq FWER$$

In contrast, FWER controlling procedures are more conservative as compared to FDR controlling procedures in the sense that they lead to smaller number of rejected hypotheses.

2.3 P-value Adjustment Methods

When conducting multiple hypothesis tests simultaneously, the likelihood of committing a Type I error increases. The family-wise error rate approaches 1 as the number of hypotheses increases. The family-wise error rate is the probability of committing at least one type 1 error $P(V > 0)$, where V is the number of type 1 errors. Therefore, various techniques exist to manage and control the occurrence of such errors. Ideally we want to bound the probability of committing any type 1 error by some α. Several post-hoc tests procedures for pairwise comparison exist. These test includes; Bonferroni, Holm (1979),

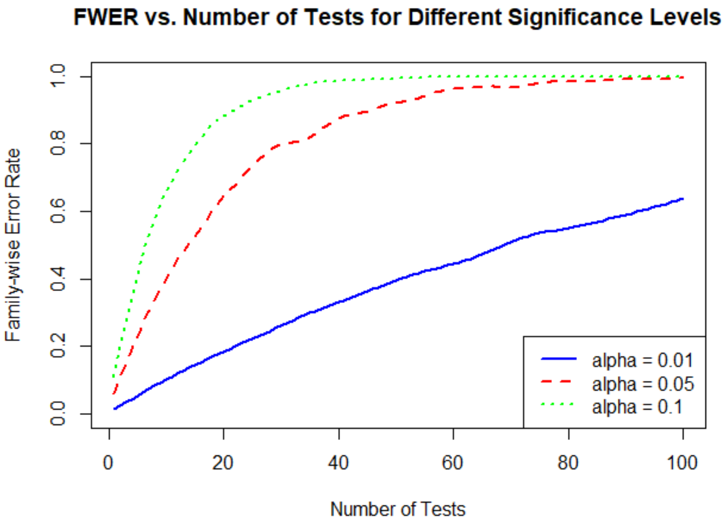


Figure 2: The family-wise error rate against the number of tests across different significance levels. From this plot we can see that, the family-wise error rate increases with increasing number of hypotheses and increasing significance level.

Hochberg (1988), Hommel (1988), and Benjamini and Hochberg (1995).

2.3.1 The Bonferroni Method

The Bonferroni correction is a method used to adjust the significance level α of individual hypothesis tests when multiple tests are conducted simultaneously. It helps control the overall Type I error rate. When we wish to perform all possible pairs of comparisons, there are $\binom{k}{2}$ such comparisons, where k is the number of groups to be compared. The Bonferroni correction assumes the null hypothesis is true for all tests in comparison.

Let α be the family-wise alpha level (overall level of significance). The Bonferroni correction rejects the null hypothesis for the ith pairwise comparison if;

$$np_i \leq \alpha$$

for all i,

where p_i is the p-value of i^{th} pairwise comparison and n is the total number of comparisons.

2.3.2 Holm Procedure

The Holm procedure is a more powerful improvement of the Bonferroni approach. It basically consists of repeatedly applying the Bonferroni inequality while testing the hypotheses in a data-dependent order Holm (1979). Let $P_{(1)} \leq ... \leq P_{(m)}$ denote the ordered unadjusted p-values associated with null hypotheses $H_{(1)}, ..., H_{(m)}$. Then, the $H_{(i)}$ is rejected if $P_{(i)} \leq \frac{\alpha}{m-j-1}$. That is, $H_{(i)}$ is rejected if $P_{(i)} \leq \frac{\alpha}{m-i+1}$ and all hypotheses $H_{(j)}$ preceding $H_{(i)}$ are also rejected.

The Holm's procedure can be described by the following sequentially rejective test procedure.

The method: Start testing the null hypothesis associated with the smallest p-value, and if $p_{(1)} > \frac{\alpha}{m}$, then the procedure stops, and no hypothesis is rejected. Otherwise, $H_{(1)}$ is rejected, and the procedure continues testing $H_{(2)}$ at a larger significance level $\frac{\alpha}{m-1}$. These steps are repeated until either the first non-rejection occurs or all null hypotheses are rejected.

2.3.3 Hochberg Procedure

The Hochberg procedure can be seen as a reversed Holm procedure. It uses the same critical values but in a reversed test sequence Hochberg (1988). $H_{(i)}$ is rejected if there is a

$j = 1, ..., m$ such that

$$P_{(i)} \leq \frac{\alpha}{m - j + 1}.$$

Alternatively, the Hochberg procedure can be described by the following sequentially rejective test procedure. Start testing the null hypothesis $H_(m)$ associated with the largest p-value $p_{(m)}$. If $p_{(m)} \leq \alpha$, the procedure stops and all hypothesis $H_{(1)}, ..., H_{(m)}$ are rejected. Otherwise, $H_{(m)}$ is retained and the procedure continues testing $H_{(m-1)}$ at a smaller significance level $\frac{\alpha}{2}$. If $p_{(m-1)} \leq \frac{\alpha}{2}$, the procedure stops and all hypothesis $H_{(1)}, ..., H_{(m-1)}$ are rejected. This iterative step continues until either the first rejection occurs or all null hypothesis $H_{(1)}, ..., H_{(m)}$ are retained. By construction, the Hochberg procedure is more powerful than the Holm procedure.

2.3.4 Hommel Procedure

The Hommel procedure, introduced by Hommel (1988), is an advanced method for controlling the family-wise error rate in multiple hypothesis testing scenarios. It applies the Simes test to each intersection hypothesis of a closed testing procedure (reject H_0 if $p_{(k)} \leq \frac{k\alpha}{n}$ for at least one k) Hommel (1988). The Hommel procedure adjusts p-values using the Simes test, considering the joint distribution of all p-values to determine the appropriate significance thresholds. It evaluates the entire set of p-values jointly and determines rejection based on the relationships among them.The procedure rejects the null hypothesis if any of the following events occur: $p_3 \leq \alpha$ or $p_2 \leq \frac{2\alpha}{3}$ and $p_1 \leq \frac{\alpha}{2}$ or $p_1 \frac{\alpha}{3}$ is true. Thus if $\frac{\alpha}{2} < p_2 < \frac{2\alpha}{3}$ and $p_1 \leq \frac{\alpha}{2}$, the Hommel procedure rejects the null hypothesis. The decision for the individual hypothesis can be performed a simpler way; compute

$j = \max i \in 1, ..., n : p_{n-1+k} > \frac{k\alpha}{i}$ for $k = 1, ..., i$. If the maximum does not exist, reject all $H_{i(i=1,...,n)}$, otherwise reject all H_i with $p_i \leq \frac{\alpha}{j}$.

2.3.5 Benjamini and Hochberg (BH) Procedure

The BH procedure is a method for controlling the false discovery rate (FDR), which is the expected proportion of false rejections among all rejected hypotheses Bretz et al. (2016). It is not strictly a step-down process like the Bonferroni correction, which controls the family-wise error rate (FWER). In the BH procedure, $H_{(i)}$ (the i-th hypothesis sorted by its p-value) is rejected if $P_{(i)} \leq \frac{i\alpha}{m}$. This means that if the i-th p-value is less than or equal to a critical value based on its rank (i) and the FDR threshold (α), then $H_{(i)}$ is rejected. The procedure begins by sorting the p-values in ascending order, denoted as $p_{(1)}, p_{(2)}, ..., p_{(m)}$. Then, the largest p-value, $p_{(m)}$, is compared with α. If $p_{(m)} \leq \alpha$, all hypotheses are rejected. If $p_{(m)} > \alpha$, the procedure proceeds to the next smallest p-value, $p_{(m-1)}$. This continues until a p-value, say $p_{(k)}$, is encountered such that $p_{(k)} \leq \frac{k\alpha}{m}$, at which point all hypotheses $H_{(1)}, H_{(2)}, ..., H_{(k)}$ are rejected, while hypotheses $H_{(k+1)}, H_{(k+2)}, ..., H_{(m)}$ are not rejected. The BH procedure is indeed a sequentially rejective test procedure in the sense that it goes through the sorted p-values sequentially until it finds the cutoff point where the condition for rejection is no longer met. The BH procedure is an improvement over the Hochberg method, offering more power under certain conditions while controlling the FDR Benjamini and Hochberg (1995).

3 Numerical Example and Simulation Study

3.1 Numerical Example

In this section, we demonstrate how the two non-parametric tests discussed above are used to make comparisons between the survival distributions across different treatment groups. The example used to demonstrate how these two tests work is on a data from a study which involved determining the time until recurrence of bladder cancer after receiving one of three treatmentsUtkarshx27 (2022). The data set is briefly described below and the number of censored and non censored observations are described in Table 2. Variable description:

- treatment: treatment received (placebo, pyridoxine or thiotepa)
- status: 0=censored , 1=recurrence.
- time: censoring or recurrence time (in months).

Number of events censored or non-censored in each treatment group		
Treatment	Censored	NonCensored
Placebo	15	11
Pyridoxine	12	16
Thiotepa	15	17
Total	42	44

Table 2: The number of bladder cancer participants censored for each treatment group .

To begin, we state our hypotheses:

$H_0 : S_p(t) = S_{py}(t) = S_{th}(t)$ vs $H_a : S_i(t) \neq S_j(t)$ for atleast one $i \neq j$

Where $S_p(t), S_{py}(t), S_{th}(t)$ are the time until bladder cancer recurrence after receiving one of the three treatments (Placebo, Pyridoxine and Thiotepa, respectively)

Next, we run the two tests (Log-rank and Peto-Peto) on the data with treatment as the only predictor. The p-values of the pairwise comparisons are adjusted for multiplicity to control the overall type one error rate. The adjusted p-values based on the different adjustment methods and the type of non-parametric test applied is summarized in Tables 3 and 4.

Log-rank Test			
Adjustment Methods	Pyridoxine vrs Placebo	Pyridoxine vrs Thiotepa	Placebo vrs Thiotepa
None	0.0003	0.1346	0.0860
Bonferroni	0.001	0.4037	0.2558
Holm	0.001	0.1346	0.1720
Hochberg	0.001	0.1346	0.1720
Hommel	0.001	0.1346	0.1720
BH	0.001	0.1346	0.1290

Table 3: The adjusted pvalues for the different pairwise group comparisons using the log-rank test across the different adjustment methods.

Peto-Peto Test			
Adjustment Methods	Pyridoxine vrs Placebo	Pyridoxine vrs Thiotepa	Placebo vrs Thiotepa
None	0.0003	0.0559	0.0627
Bonferroni	0.0008	0.1676	0.1881
Holm	0.0008	0.1118	0.0627
Hochberg	0.0008	0.0627	0.0627
Hommel	0.0008	0.0627	0.0627
BH	0.0008	0.0838	0.0627

Table 4: The adjusted pvalues for the different pairwise group comparisons using the Peto-Peto test across the different adjustment methods

The results after the p-value adjustments show that the survival distributions of Pyridoxine group is significantly different from the placebo group across all adjustment methods and tests. When we look at the Pyridoxine and Thiotepa pair, we can observe

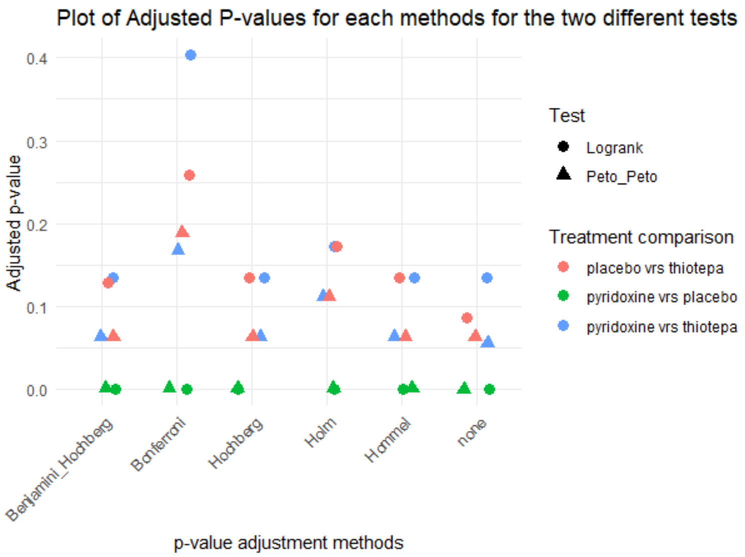


Figure 3: The adjusted p-values for the pairwise comparison for the different adjustment methods across the two tests.

that these two curves are not significantly different across all the adjustment methods for the two tests. We can see that the Log-rank test shows higher p-values than the Peto-Peto tests which may be as a result of how each test assigns scores. The last pairwise comparison is between the Placebo and Thiotepa groups. We can see from the tables that, the p-values provide weak to no evidence against the null hypothesis of no difference in survival distributions. Again, the p-values after adjustment from the Peto-Peto test are relatively smaller as compared to those from the Log-rank. This again maybe be due to how these tests assign weights. In general, the Log-rank test and the Peto-Peto test yielded similar results in terms of detecting significant or non-significant differences between the curves being compared, suggesting that either test can be used for comparing survival curves in this context. In general, the p-value adjustment methods help to control the overflow of type one errors in multiple hypothesis and it is evident from the tables that the adjusted p-values are higher than the unadjusted p-values and hence the probability of false positives are reduced. The Kaplan Meier curves in Figure 1 also help to visualize the relationships between the survival distributions of the groups.

3.2 Simulation Studies

In this section of the project, we conducted a comprehensive simulation study to compare the power of five different p-value adjustment methods across two commonly used survival tests: the log-rank test and the Peto-Peto test. The aim was to evaluate the performance of these methods under various scenarios, considering different effect sizes and sample sizes. For each group in the simulation, we modeled the survival distribution using random samples the Exponential distribution with a rate parameter λ. The survival distribution for the i^{th} group was represented as;

$$S_i(t) = \frac{1}{\lambda_i} \exp\left(-\frac{t}{\lambda_i}\right)$$

where λ represents the rate parameter. We considered multiple scenarios by varying the rate parameters as follows: $\lambda = (1, 1, 1), (1, 2, 3), (1, 3, 5), (1, 4, 8)$ .These scenarios reflected different effect sizes in the survival functions of the respective groups. The simulations were performed for a range of sample sizes per group: n=10, 25, 50, 100, and 200 with equal allocation across groups. Additionally, we randomly assigned event occurrences, where approximately 80 percent of observations experienced the event, while the remaining 20 percent were censored. For each combination of rate parameters and sample sizes, test statistics were generated based on the respective survival tests under the alternative hypothesis. The empirical significance level was 0.05 . P-values were extracted from each test and subsequently adjusted for multiplicity using the specified adjustment methods discussed earlier in the project. The power of each adjustment method was estimated as the proportion rejecting null under the alternative hypothesis for the two tests. This process was iterated 200 times for each scenario, and the average proportions of rejected nulls were reported as the estimated power. The simulated power under various scenarios is summarized in the subsequent tables, providing insights into the comparative performance of the p-value adjustment methods across different effect sizes,sample sizes and test.

4 Results

The provided tables contain estimated power values for both the Log-rank test and the Peto-Peto test using different adjustment methods across various sample sizes and mean combinations. A comparison of power values across all sample sizes, mean combinations, and adjustment methods reveals that the Log-rank test consistently exhibits higher power than the Peto-Peto test in all instances. These power values allow us to assess how effectively the two tests detect differences between survival distributions of groups, if such differences exist.

In general, we expect an increase in statistical power as sample sizes grow larger. This suggests that larger sample sizes are associated with higher statistical power of tests. Additionally, when examining individual adjustment methods, there is a continuous trend of increasing power as sample sizes increase.

Observing the power values in the provided tables, a significant increase is evident as the distance between the means of the groups increases. Notably, when all group means are set at 1, the null hypothesis is true in this instance and statistical power is minimal and represents the probability of a type 1 error. However, as the mean differences between groups increase, so does the power of the tests. A difference of 3 or more between group means consistently results in an expected power close to 1 across all tests and adjustment methods. This observation is an indication that, the two tests considered in this project efficiently detect differences in survival distributions if they exist. However, it is also evident that some methods may prove more efficient when differences in means are not substantial.

As expected, the Benjamini-Hochberg (BH) adjustment method demonstrates the highest power values across all tests and sample sizes, followed closely by the Hommel procedure, with the Hochberg and Holm procedures producing similar power values in most cases. This similarity can be attributed to their mode of application of these methods. Conversely, the Bonferroni procedure consistently exhibits the lowest power among all adjustment methods across the two tests. This can be attributed to the fact that the Bonferroni method is conservative in nature and hence has lower statistical power.

4.0.1 Tables of Statistical Power

Comparison of Adjustment Methods between Log-Rank Test and Peto-Peto Test

Group Mean Combination 1, 1, 1

2*Sample Size	BH		Bonferroni		Holm		Hochberg		Hommel	
	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto
10	0.0117	0.0083	0.0083	0.0083	0.0100	0.0083	0.0100	0.0083	0.0100	0.0083
25	0.0117	0.0150	0.0117	0.0133	0.0117	0.0133	0.0117	0.0133	0.0117	0.0133
50	0.0233	0.0200	0.0217	0.0150	0.0217	0.0150	0.0217	0.0150	0.0217	0.0167
100	0.025	0.0250	0.0217	0.0167	0.0233	0.0183	0.0233	0.0183	0.0233	0.0200
200	0.035	0.0283	0.0283	0.0233	0.0317	0.0233	0.0317	0.0233	0.0333	0.0250

Group Mean Combination 1, 2, 3

2*Sample Size	BH		Bonferroni		Holm		Hochberg		Hommel	
	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto
10	0.1867	0.1683	0.1617	0.1383	0.17	0.1517	0.1733	0.155	0.175	0.1583
25	0.5	0.44	0.435	0.3667	0.47	0.3983	0.4767	0.415	0.4767	0.4167
50	0.725	0.655	0.645	0.5733	0.7167	0.6317	0.72	0.6383	0.72	0.6383
100	0.8983	0.8667	0.8383	0.7783	0.8983	0.8633	0.8983	0.8667	0.8983	0.8667
200	0.9817	0.965	0.9583	0.9367	0.9817	0.9817	0.9817	0.965	0.9817	0.965

Group Mean Combination 1, 3, 5

2*Sample Size	BH		Bonferroni		Holm		Hochberg		Hommel	
	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto
10	0.4183	0.3817	0.3533	0.3167	0.3967	0.3533	0.4	0.3533	0.4017	0.3583
25	0.7183	0.6733	0.665	0.5933	0.7117	0.645	0.7117	0.645	0.7117	0.65
50	0.885	0.8383	0.8167	0.7783	0.8383	0.8383	0.8383	0.8383	0.8383	0.8383
100	0.9617	0.935	0.915	0.8833	0.9617	0.935	0.9617	0.935	0.9617	0.935
200	0.9983	0.9967	0.9983	0.9867	0.9983	0.9967	0.9983	0.9967	0.9983	0.9967

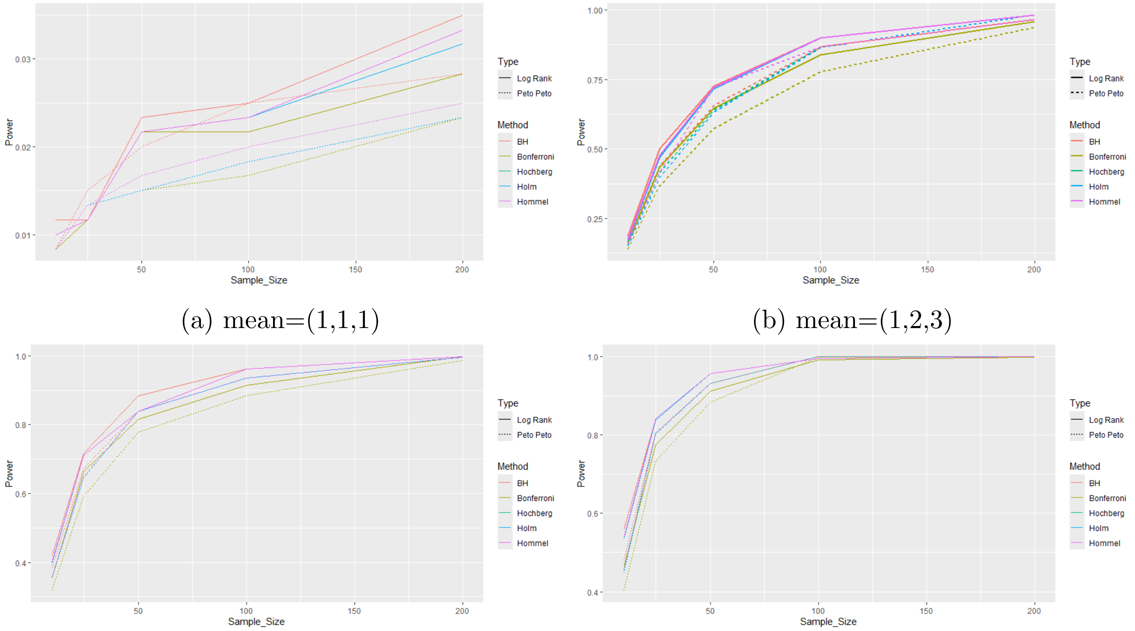
Group Mean Combination 1, 4, 8

2*Sample Size	BH		Bonferroni		Holm		Hochberg		Hommel	
	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto	Log-Rank	Peto-Peto
10	0.5583	0.4783	0.4633	0.4017	0.535	0.4467	0.5383	0.4533	0.5383	0.455
25	0.8417	0.805	0.775	0.7333	0.8383	0.8017	0.8417	0.805	0.8417	0.805
50	0.9567	0.9317	0.9117	0.8833	0.9567	0.9317	0.9567	0.9317	0.9567	0.9317
100	0.995	1	0.9917	0.9983	0.995	1	0.995	1	0.995	1
200	1	1	0.9983	0.9983	1	1	1	1	1	1

Table 5: Tables of Statistical Power for the two tests.

4.0.2 Visualizations of Statistical power

Visual Comparison of Statistical Power of the two Tests Across Adjustment Methods for Different Sample Sizes and Mean Combinations.



(a) mean=(1,1,1)

(b) mean=(1,2,3)

(c) mean=(1,3,5)

(d) mean=(1,4,8)

Figure 4: Plots of statistical power of the two tests across different sample sizes, adjustment methods and mean combinations. These plots show increasing statistical power as sample size increases. In all cases, the Log-rank tests shows higher statistical power over the Peto-Peto test. The BH adjustment method shows the highest power, followed by the Hommel procedure,then the Hochberg method, followed by Holms procedure and the Bonferroni with the lowest statistical power

5 Conclusion and Recommendation

5.1 Conclusion

The analysis and the results from both the numerical example and the simulation study offer valuable insights into the usefulness of the two non-parametric tests discussed in this project. Additionally, they shed light on the significance of p-value adjustment methods in controlling the overall type I error rate in multiple hypothesis testing scenarios. Moreover, the effects of sample size on statistical power and the impact of differences between group means on the ability of tests to detect significant effects, if they exist, have been established.

From the numerical example earlier, it became evident that the different choices of p-value adjustment methods have distinct ways of handling multiplicity. Some methods, such as the Bonferroni, are conservative, while others, like the BH procedure and the Hommel procedure, are more liberal. This characteristic of adjustment methods makes them suitable for their respective uses, highlighting the importance of researchers clearly defining their end goals and expectations before deciding on a particular adjustment method.

When comparing multiple survival curves using the two tests discussed in this project, it is crucial to understand the population or the characteristics of the data being analyzed. Researchers can opt for either the popular Log-rank test, which assigns equal weights to each observation, or the Peto-Peto test, which assigns higher weight to earlier observations. Generally, the Log-rank test and the Peto-Peto test yielded similar results in terms of detecting significant or non-significant differences between the curves being compared, suggesting that either test can be used for comparing survival curves in this context.

The various sample sizes used in this project have demonstrated the importance of having enough sample size to detect significant differences if they exist. Across all tests and adjustment methods, it has been shown that an increased number of observations is associated with an increase in statistical power.

In conclusion, there is a trade-off between statistical power and controlling the overall type I error rate. It is imperative for researchers to clearly define research goals and make informed decisions to achieve the best results. Adequate sample size is crucial for detecting significant differences if they exist

5.2 Recommendations

- Explore additional non-parametric tests: Future studies should consider investigating a broader range of non-parametric tests for comparing multiple survival curves beyond the Log-rank test and the Peto-Peto test. Exploring alternative tests could provide insights into how suitable these methods are for handling different types of survival data.

- Application to Real-world Data: It is necessary to apply the methods discussed in this project to real-world datasets that encompasses diverse populations and conditions. By analyzing real-world data, researchers can assess the performance of thees methods in different contexts, enhancing our understanding of their effectiveness and limitations. This real-world application will provide valuable insights into how these methods perform under various scenarios and inform best practices for survival analysis in practical settings.

References

Altman, D. G. (1992). Analysis of survival times.

(In: Practical statistics for Medical research; pp. 365–93)

Benjamini, Y., & Hochberg, Y. (1995, 01). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289-300.

Bretz, F., Hothorn, T., & Westfall, P. (2016). *Multiple comparisons using r.* CRC Press.

D'Arrigo, G., Leonardis, D., Abd ElHafeez, S., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to analyse time-to-event data: The kaplan-meier survival curve. *Oxidative Medicine and Cellular Longevity*, *2021*, 2290120.

Estimating statistical power when using multiple testing procedures. (2017, 11).

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis.* Wiley.

Gamel, J., & Vogel, R. (1997). Comparison of parametric and non-parametric survival methods using simulated clinical data. *Statistics in Medicine*, *16*, 1629–1643.

Hazra, A., & Gogtay, N. (2017, May-Jun). Biostatistics series module 9: Survival analysis. *Indian Journal of Dermatology*, *62*(3), 251-257.

Hochberg, Y. (1988, 12). A sharper bonferroni procedure for multiple tests of significance. , *75*, 800-800.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, *75*(2), 383–386.

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

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). New York: Springer.

Lydersen, S. (2021, 09). Adjustment of p-values for multiple hypotheses. *Tidsskrift for Den norske legeforening*.

Machin, D., Cheung, Y. B., & Parmar, M. (2006). *Survival analysis: A practical approach.* Wiley.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008, 01). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537-563.

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*, *135*, 185–20.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., & Wang, E. W. (2010, 09). A practical guide to understanding kaplan-meier curves. *Otolaryngology–Head and Neck Surgery*, *143*, 331-336.

Schober, P., & Vetter, T. R. (2018, Sep). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and analgesia*, *127*(3), 792-798.

Schober, P., & Vetter, T. R. (2021). Kaplan-meier curves, log-rank tests, and cox regression for time-to-event data. *Anesthesia & Analgesia*, *132*(4), 969–970.

Tarone, R. (1981). On the distribution of the maximum of the logrank statistic and the modified wilcoxon statistic. *Biometrics*, *37*, 79–85.

Utkarshx27. (2022). *Bladder cancer recurrences.* Kaggle.

Zhang, F., & Gou, J. (2016, 10). A p-value model for theoretical power analysis and its applications in multiple testing procedures. *BMC Medical Research Methodology*, *16*.