TU Dortmund

Introductory Case Studies

Project 2: Comparison of multiple distributions

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Franziska Kappenberg

M. Sc. Marieke Stolte

Author: Mohammad Sakhawat Hossain

Matriculation No: 231838

Group number: 11

Group members: Muhammad Mahir Hasan Chowdhury, Aritra Paul, Sadia Mahjabin and Rachel John Christopher

December 9, 2022

Contents

1 Introduction

Sport is a powerful, direct universal language that unites peoples, cultures, and genders regardless of race, color and creed. When it comes to swimming, it builds endurance, muscle strength, and cardiovascular fitness. Swimming emphasizes the value of accuracy, precision and calls for close attention to detail. A total of 75 medal events were held at the 2022 European Aquatics Championships in Rome. The swimming, open water swimming (free entry), diving, artistic swimming, and high diving disciplines drew more than 55,000 spectators. During the European Championships in Rome, Italy emerged as a dominant force in swimming, with East Germany winning all of the women's swimming races (European Aquatics Championships, 2022).

The goal of this project is to look into how timing differs between the five categories: swimming, open water swimming, artistic swimming, diving, and high diving. We also want to see if there are any pairwise differences between the resulting times by considering all pairs of categories. At first, one-way analysis of variance (ANOVA) is used to conduct a global test, and then a pairwise t-test is applied to check differences in terms of timing between the pair of categories. Furthermore, to address the multiple comparison problem, the results of the pairwise t-tests are adjusted with the Bonferroni and Holm Bonferroni methods. In addition, necessary assumptions are also taken into consideration while conducting the tests.

In section 2, the data set and the project objectives are explained in detail. The data collection method, data quality, data preprocessing, data type as well as data size are explained here. Section 3 discusses the statistical methods such as ANOVA, pairwise t-tests, Q-Q plot, Bonferroni, and Holm-Bonferroni method that are used for data analysis in this project. We use these statistical approaches in section 4, and the results of different statistical tests are thoroughly interpreted and analyzed. Section 5 presents the result and an in-depth summary of the project, followed by a discussion of potential future research on this data set.

2 Problem Statement

2.1 Description of the data set

The data set is compiled and given by the instructors of the course, 'Introductory Case Studies' at TU Dortmund University in the winter session (2022/2023). On the website "European Aquatic Roma 2022," the original data set is accessible under "RESULTS" section (LEN European Aquatics, 2022) and processed by Microplus which is a sports related organization that provides excellent data processing services, by utilizing their technical strength and core competencies (MicroplusSrl, 2022).

The given data set 'SwimmingTimesFile.csv' contains the results of the women's 200-meter semi-finals, which have 80 observations and 3 variables. Under the nominal variable *Category* there are five types of swimming: backstroke, breaststroke, butterfly, freestyle and medley and each type contain 16 participants. The nominal variable *Name* consists of 80 participants, and the numerical variable *Time* has different times (in seconds) taken by swimmers to complete the race. There are some participants who took part in more than one category. We removed such 8 participants and therefore our final data set contain only 72 observations. In general, the quality of the data seems fine, as it originated from a professional web portal and doesn't contain any missing values.

2.2 Project objective

The main objective of this report is to apply the statistical methods to analyze the time difference (in seconds) among the five categories, to understand the underlying distribution, and to investigate if there is a significant relationship between the resulting times in paired categories. At first, quantile-quantile plot (Q-Q plot) is used to identify the underlying distribution in the data set. To see if there is at least one average time that differs across the five categories, we perform a one-way analysis of variance (ANOVA). Additionally, we use the pairwise t-test to see whether there is a pairwise difference in the categories in terms of time. For both tests to analyse the result we use p-value approach. Finally, we use the Bonferroni method and the Holm-Bonferroni method to deal with problems caused by multiple tests. We finish by comparing the results with and without the correction method such as Bonferroni and the Holm-Bonferroni correction.

3 Statistical Methods

In this section, several statistics methods are introduced which are later used in our analysis. The software R (R Core Team, 2021), ggpubr (Kassambara, 2020), ggplot (Wickham, 2016), and cowplot (Wilke, 2020) are used for all statistical test and visualisation.

3.1 Q-Q Plot

The quantile-quantile plot (Q-Q plot) is a graphical method for determining whether the distributional properties of the sample come from a theoretical distribution like the normal distribution. Consider sample data $y_1, y_2, ..., y_n$ are sorted in ascending order $y_{(1)}, y_{(2)}, ..., y_{(n)}$, they are referred to as observed quantiles. The probability points p_i, for $i = 1, 2, 3, ., n$ is calculated using the given formula:

$$pi = \begin{cases} \frac{i-\frac{3}{8}}{n+\frac{1}{4}} & \text{if } n \leq 10 \\ \frac{i-\frac{1}{2}}{n} & \text{if } n > 10. \end{cases}$$

The theoretical quantiles x_i, for their corresponding sorted sample quantile y_i are computed using probability points pi. We then have to find x_i, the theoretical quantile for $i = 1, 2, .., n$ such that $P(X \leq x_i) = pi$ where $X \sim N(0, 1)$ i.e X is normally distributed with mean value 0 and variance 1. For plotting the graph, ordered pair (x_i, y_i) are used. On the same graph a reference line is superimposed. In normal probability Q-Q plot, the sample mean is the y-intercept and the standard deviation is the slope of reference line. For non standardized data reference, line is of the form $y = \mu + z$ where z is the transformation computed using $(y - \bar{y})/s$, μ represents mean and s represents standard deviation of proposed theoretical normal distribution. For standardized data, the reference line takes the form $y = x$. Once the plot is ready and the reference line is superimposed on to the plot, it is possible to visually inspect if the data points follow the theoretical distribution. If the points follow the reference line, then it is highly likely that they are normally distributed. Deviations from the reference line imply a deviation from normality and have a noticeable concave trend. The presence of such concave trends leads to skewness in the plot(Hay-Jahans, 2019, p. 147-152).

3.2 Null hypothesis(H_0) and alternative hypothesis (H_1)

Statistical hypothesis tests are the foundation of many statistical analysis methods and understanding the fundamentals of hypothesis testing is critical. The data is interpreted by assuming a specific outcome and then using statistical methods, we either reject or confirm the assumption. This assumption about the outcome is referred to as a hypothesis, and the statistical hypothesis tests are used to test these hypothesis. To accurately reflect the question that the tester wishes to answer, the hypothesis test must be carefully designed. There are two mutually exclusive hypothesis in a statistical hypothesis testing. First one is null hypothesis H_0 which aim to test and maintain if there is no strong evidence against it. The other one is alternative hypothesis H_1 which is a statement that directly contradicts the null hypothesis. A hypothesis test can be two-sided or one-sided. If the test is $H_0 : \theta = \theta_0 \quad versus \quad H_1 : \theta \neq \theta_0$ then it's called two-sided test and one-sided test is of the form $H_0 : \theta \leq \theta_0 \quad versus \quad H_1 : \theta \geq \theta_0$. Here, θ is population mean and θ_0 is a specific number. The null hypothesis states that the population parameter θ is equal to a specific value of θ_0 and the alternative hypothesis states that the population parameter θ is not equal to a specific value of θ_0 (Banerjee et al., 2009).

3.3 Significance level (α) and p-value

The significance level is denoted by α and defined as the probability of rejecting the null hypothesis when it is actually true. Researchers or statistician define the significance level before conducting the test. The significance level is .05 means there is a 5% chance of taking alternative hypothesis if null hypothesis is true. In our report we will use 5% significance level (Wasserman, 2010).

One of the techniques to draw conclusion whether the null hypothesis should be rejected or not is p-value method. If the assumption for the null hypothesis is true then p-value can be represented as the probability of getting the test results at least as extreme as the result actually observed. Conclusion about a particular hypothesis depends on p-value and significance level. If our p-value is less than the significance level ($p < \alpha$) then we reject the null hypothesis and we can say that the result is statistically significant. If the p-value is greater than the significance level ($p > \alpha$) then we don't reject the null hypothesis and hence the result is assumed statistically insignificant (Du Prel et al., 2009).

3.4 Statistical tests and assumptions

To determine whether there is a significant difference between means of populations, some statistical tests are used. In this report two different tests, namely pairwise t-test and one-way ANOVA are used as inferential statistics. When we have multiple pair of categories, pairwise t-test examines two groups or categories but one-way ANOVA can be used to analyze mean time of more than two groups. When the null hypothesis is already rejected according to one-way ANOVA, the pairwise t-test can give more detailed information on the sources of difference in population means. To proceed with the inference about the means of population certain assumptions have to be held for both tests (Black et al., 2018, p. 409).

- Populations should be normally distributed.
- Populations should have equal variances.
- Observations in the populations should be independent with each other.

3.4.1 Analysis of variance (ANOVA)

ANOVA is a statistical test which is used to determine the mean difference between more than two data samples or groups. ANOVA can be one way or two way test. In this report we will use one way ANOVA as we have only one quantitative independent variable. For two independent variable we use two-way ANOVA test. In ANOVA if we have k samples then the following hypothesis are tested:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \text{At least one of the means is different from the others.}$$

When one of the sample mean is different than others, then we reject the null hypothesis and we can say there is significant difference among the groups or data samples. To conduct the test at first we have to measure between and within group variances. Variances between the groups (SSB) can be calculated by subtracting variances within groups (SSW) from total variances (SST)

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{i,j} - \bar{X})^2 \text{ and } SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{i,j} - \bar{x_j})^2.$$

Here, $\bar{X}$ is the grand mean which can be defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{i,j}.$$

x_{ij} is the individual data points of all groups, k is the number of groups or data samples and $\bar{x_j}$ is the mean value of group j. n is the total number of sample size. Next, mean square between group (MSB) and mean square within group (MSW) is calculated by using the following formula:

$$MSW = \frac{SSW}{\sum_{j=1}^{k}(n_j - 1)} = \frac{SSW}{n-k} = \frac{SSW}{df_w} \text{ and } MSB = \frac{SSB}{k-1} = \frac{SSB}{df_b}$$

Here, degree of freedom is referred to as the maximum number of logically independent values that have the freedom to vary. The degree of freedom between groups $df_b = k-1$, where k is the number of groups and degree of freedom within groups $df_w = n-k$, where k is the number of groups and n is the number of sample sizes (Tukey, 1949). Finally F-statistics F^* is the ratio of the two variances:

$$F^* = \frac{MSB}{MSW}$$

To check whether the null hypothesis should be rejected or not, p-value method can be used. In ANOVA p-value corresponds to the probability $P(F \geq F^*|H_0)$ where F is random variable comes from $F_{(df_w, df_b)}$ distribution in significance level α. If the p-value is less than the significance level ($p < \alpha$), the null hypothesis is rejected. In contrary, if the If the p-value is greater than the significance level ($p > \alpha$), then we don't reject the null hypothesis (Black et al., 2018, p. 406-411).

3.4.2 Pairwise T-test

By comparing the mean values of the groups in the preceding part, the ANOVA test merely indicated whether it rejected the null hypothesis or not; no information was provided as to which two groups had similar mean values and which did not. To find out whether there is a significant difference in the mean values between pairwise groups or categories pairwise t-test should be done. Similar to the ANOVA test, this test's assumptions are already provided at the beginning of this section. For all pairs i, j

where $i, j = 1, ..., k$ the following is the hypothesis for this test:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j \text{ with } i \neq j$$

In order to obtain t-statistic we need to compute a pooled sample standard deviation (S_p) which is a weighted average of standard deviation taken from two or more than two independent groups or data samples. The formula of pooled SD can be defined as follows:

$$S_p^2 = \frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2 + .. + (n_k - 1)\ s_k^2}{n_1 + n_2 + .. + n_k - k}$$

If $n_1, n_2.....n_k$ are the sample sizes of group k and $s_1, s_2, ...s_k$ are the standard deviation of k groups then assuming the pooled SD same ($\sigma_1 = \sigma_2$) we can write the pairwise-test formula as:

$$t^* = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{(n_1-1)\ s_1^2+(n_2-1)\ s_2^2+..+(n_k-1)\ s_k^2}{n_1+n_2+..+n_k-k}}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\bar{x_1} - \bar{x_2}$ is the mean difference between two groups.

After having the t-statistic, in order to check whether we should reject null hypothesis or not, p-value approach can be used. In two-tailed pairwise t-test p-value corresponds to the probability $2P(T \geq | t^* | | H_0)$ where T is a random variable comes from $T_{n_1+n_2+...n_k-k}$ distribution in significance level α. If the p-value is less than significance level , we have sufficient evidence to reject the null hypothesis. However, if the p-value is greater than significance level, it indicates we have insufficient evidence to reject the null hypothesis. If we reject the null hypothesis then we conclude that there is a significance difference between two population mean on the significance level α (Hay-Jahans, 2019, p. 261).

3.5 Bonferroni method

Whenever we reject the null hypothesis wrongly then this type of instance is called type I error or false positive. On the other hand, whenever we don't reject the null hypothesis wrongly then this type of instance is called type II error or false negative (El-gohary, 2019). When several dependent or independent statistical test are performed at a time the false positive rate increases substantially and therefore the Bonferroni is used with

a view to reduce the instance of false positive rate. Let $H_1, ..., H_m$ is a family of hypothesis and there corresponding p-values $p_1, ..., p_m$. Let m_0 is the number of true null hypothesis where m is the total number of null hypothesis then family wise error rate (FWER) is the probability of making at least one type I error i.e probability of rejecting at least one true H_i. We can calculate the family wise error rate as follows:

$$FWER = 1 - (1 - \alpha)^m$$

where m is number of tests and α is significance level.

To avoid the type I error, we can test each hypothesis by using alternate significance level i.e, α/m. Therefore we can use the adjusted α and compare it with the p-value. Adjusted α value can be calculated as follows:

$$\alpha_{adjusted} = \alpha/m$$

Although this method significantly reduces type I error but increases chance to the vulnerability of type II error (Hay-Jahans, 2019, p. 274).

3.6 Holm-Bonferroni method

Holm-Bonferroni procedure is a sequential approach whose goal is to increase the power of the statistical tests while keeping the familywise error rate under control. At first the tests are performed in order to obtain their p-values, then the steps are as follows:

- all p-values are sorted from smallest to largest. Let's say, m is the number of the p-values and α is our significance level.
- No subsequent p-values are significant if the first p-value is larger than or equal to α/m, which ends the operation. If not, we continue.
- the first p-value is considered significant and then, the second p-value is compared to $\alpha/(m-1)$. If the second p-value is larger than or equal to $\alpha/(m-1)$, the procedure is stopped and no further p-values are significant. Otherwise, we go on until the i-th ordered p-value is such that:

$$p_i \geq \alpha/(m+1-i)$$

For the smallest p-value, both the Holm's and the Bonferroni methods will yield the same adjusted p-value. However, the other adjusted p-values will be smaller for Holm's method compared to the Bonferroni method. Therefore, Bonferroni is more likely not to reject a false null hypothesis in comparison to Holm's method, and thus Holm's method is statistically more powerful than the classical Bonferroni method (Fu et al., 2014).

4 Statistical Analysis

In this section, the statistical methods discussed in the earlier sections are applied to our sample data extract to investigate our hypothesis about the underlying distribution of the data. In order to keep the dataset balanced, eight observations that appeared in more than one category have been removed. Six contestants who took part in one of the four categories and the medley have been eliminated. Additionally, two swimmers who competed in both the breaststroke and butterfly have been dropped. Therefore, backstroke and breaststroke each have 14 observations, butterfly has 13, freestyle has 15, and the medley category has 16 observations, which will be used in our analysis.

4.1 Descriptive analysis and checking assumptions

In this subsection, we check the assumptions and describe the data as well as the distribution for each category. Summary statistics for all 5 categories are represented in the table 1. According to this table, the minimum (144 seconds) and maximum (148 seconds) times taken by the swimmers are higher in the breaststroke category, which also indicates that the swimmers belonging to this category take the highest average time (146 seconds) to complete the race. In contrast, contestants in the freestyle event take the shortest amount of time which is 119 seconds on average.

Category	Variable	No. of participants	Min	Max	Mean	Median	IQR	SD
Backstroke	time	14	128	136	131	131	1.52	1.85
Breaststroke	time	14	144	148	146	147	2.00	1.51
Butterfly	time	13	128	137	132	131	5.1	2.72
Freestyle	time	15	118	122	119	120	2.24	1.56
Medley	time	16	132	137	134	133	2.48	1.59

Table 1: Summary statistic of different categories (time taken in seconds).

The boxplot (Figure 1) shows that the time taken by participants in the backstroke category is symmetrical, as the mean and median values (131 seconds) are the same. Additionally, one exceptional competitor in the backstroke category finished the race in 136 seconds, which has an effect on overall variance of this category. The other four groups' means and medians only differ by one second. From the boxplot, we also observe that the backstroke (1.52) has a lower interquartile range (IQR), but the IQR for the butterfly (5.1 seconds) is more than double that of the backstroke. The breaststroke (1.51) and butterfly (2.72) categories have lower and greater standard deviations (SD), whereas freestyle and medley have SDs that are almost identical.

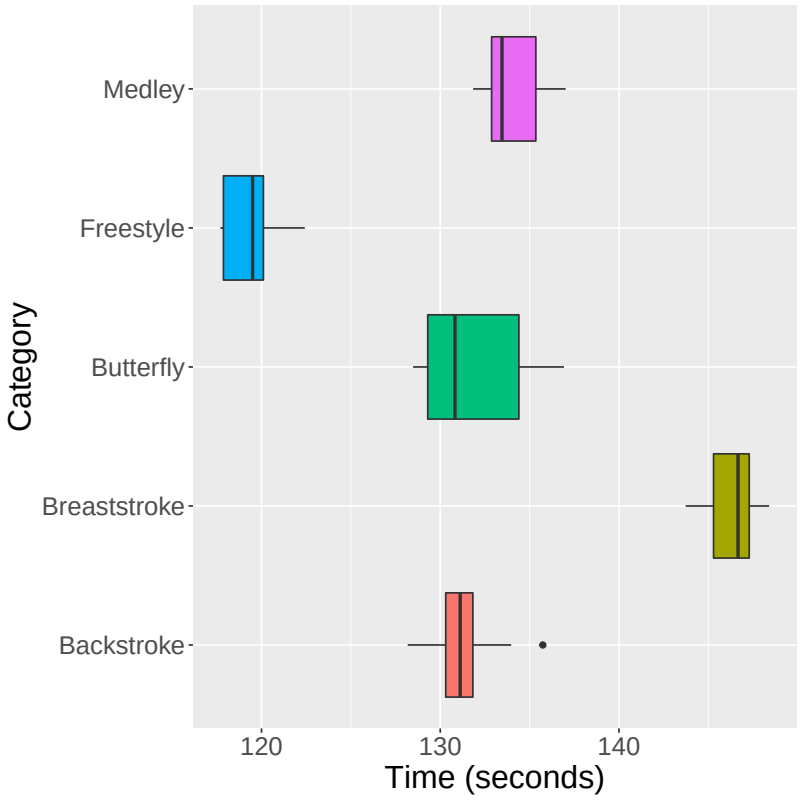

Figure 1: Box plot of all categories with respect to time.

In order to visualize whether the samples are normally distributed or not, Q-Q plots are used. One individual Q-Q plot is shown for each category (Figure 2). Breaststroke seems to follow a normal distribution since its quantile is approximately on the reference line. Similarly, almost all the data points for the freestyle and medley categories lie on the reference line. However, one data point in the backstroke category is far from the

reference line, indicating one swimmer took an exceptionally long time to complete the race.

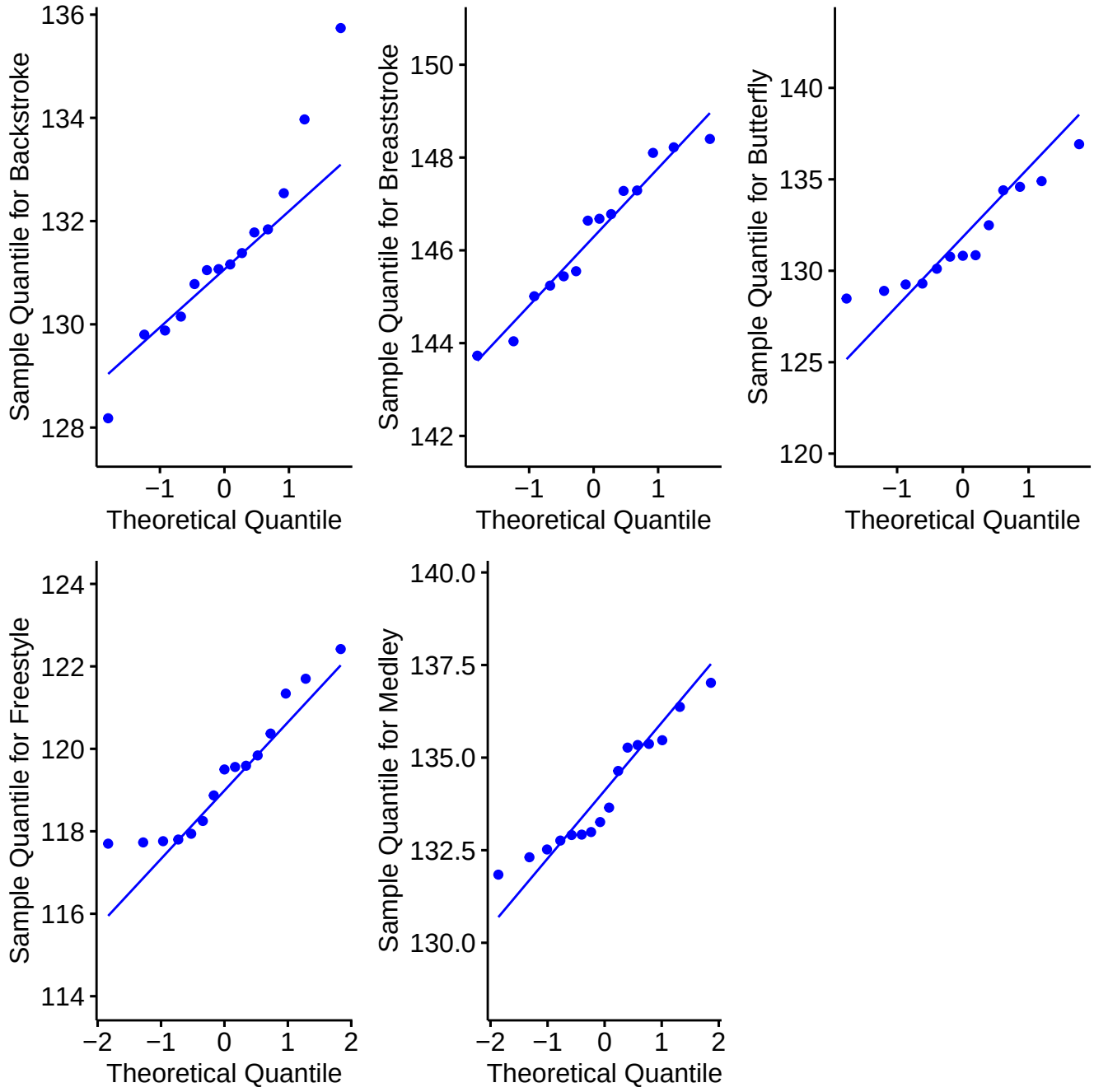

Figure 2: QQ plot for all categories.

From the box plot and Q-Q plot, we observe some disparity in the assumptions of variance homogeneity and normality of the data distribution. It might be reason that the sample size we are considering here are very small. However, we proceed further by assuming that the variance homogeneity and normality of the data distribution exist in our sample data set. Furthermore, the observations in the given data set are unrelated to one another, so the occurrence of one property has no effect on another. It implies that we can assume that the independence assumptions are likewise true.

4.2 Global test

In this section, we will use a global test called ANOVA to investigate whether the time differs between the five categories or not. In the method sections, the underlying assumptions for this test are stated and checked in the preceding section. The hypothesis for this test can be assumed as follows:

- Null Hypothesis, H_0 : For all five categories, the average time (in seconds) is the same.
- Alternate Hypothesis, H_1: The average time taken (in seconds) differs in five categories.

We can see the degree of freedom, sum of squares, mean square, F-statistic, and p-value from the table 2. By using the p-value method, we can decide whether the null hypothesis will be rejected or not. Since the p-value is less than the presumed significance level (.05), we reject the null hypothesis. Rejecting the null hypothesis means there is a statistically significant difference among the five categories in terms of time. However, to find out which category caused the rejection, we will use a pairwise t-test in the next section.

Variable	DF	Sum square	Mean square	F Value	P-value
Category	4	5326	1331.6	380.3	$2*10^{-16}$
Residuals	67	235	3.5	-	-

Table 2: Output of the ANOVA test.

4.3 Pairwise comparison and adjusting p-values

Pairwise t-test are performed for ten pairs of categories to determine whether there are pairwise time differences between the two categories. Since it is assumed that all groups come from populations with a common standard deviation, by using pool.sd = TRUE statement in pairwise t-test function a common variance estimator for all groups is obtained. The underlying assumptions are already checked, and the hypothesis can be defined as follows:

- Null Hypothesis, H_0 : The mean value of time does not differ between any two categories.
- Alternate Hypothesis, H_1: Between the two categories, there is a difference in the mean value of time.

The results of 10 pairwise t-tests are demonstrated in the table (Table 3). We observe that all the pairs except butterfly and backstroke have smaller p-values, which are less than the significance level (.05). Therefore, we can reject the null hypothesis for these 9 pairs and thus conclude that there are significant differences between the categories in terms of average time. On the other hand, the p-value for butterfly and backstroke is .68, which is greater than the significance level, so we cannot reject the null hypothesis, thus there is no difference in average time between these two categories. This is no surprise, as we have seen (from Table 1) that the mean, median, minimum, and maximum values are almost the same for these two categories.

Categories	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$< 2*10^{-16}$	-	-	-
Butterfly	0.68	$< 2*10^{-16}$	-	-
Freestyle	$< 2*10^{-16}$	$< 2*10^{-16}$	$< 2*10^{-16}$	-
Medley	.00024	$< 2*10^{-16}$	.001	$< 2*10^{-16}$

Table 3: p-value of pairwise comparison without adjustment method.

Since we conduct the 10 statistical analyses on the same sample data, the family-wise error rate (FWER) increases, and FWER can be calculated as follows:

$$FWER = 1 - (1 - \alpha)^m = 1 - (1 - .05)^{10} = .40$$

So, the chance of erroneously rejecting the null hypothesis at least once among the family of analyses is equal to 40%. As mentioned earlier, to control the inflated family-wise error

rate in multiple testing, we should make an adjustment. In this report, Bonferroni and Holm-Bonferroni are used to control the FWER. From the table 4, we see all the adjusted p-values with the Bonferroni method. Here, the p-value for butterfly and backstroke is 1, while this value is 0.68 in the t-test without adjustment procedure. All other p-values are less than the significance level, which is also the case in the previous table (Table 3). However, the p-value for the medley and butterfly categories (.01) has increased compared to the t-test without adjustment but is still less than the significance level.

Categories	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$2.54*10^{-30}$	-	-	-
Butterfly	1.00	$2.43*10^{-29}$	-	-
Freestyle	$2.14*10^{-25}$	$1.29*10^{-46}$	$1.65*10^{-25}$	-
Medley	$2.38*10^{-03}$	$2.92*10^{-26}$	.01	$3.60*10^{-31}$

Table 4: p-value of pairwise comparison with Bonferroni adjustment method.

Similarly, if we use the Holm-Bonferroni correction (Table 5), the p-value for the butterfly and backstroke is slightly lower (.68) compare to p-value in the Bonferroni method but equal to p-value obtained by using pairwise t-test. This adjusted p-value (.68) still greater than the significance level. All other p-values are less than the significance level. Therefore, we can conclude that regardless of the correction method, we can't reject the null hypothesis for the butterfly and backstroke categories and hence there is no mean difference in time between these two categories. On the contrary, for all other 9 pairs, we can reject the null hypothesis as p-values are less than α and thus there is a significant difference between these categories.

Categories	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$2.03*10^{-30}$	-	-	-
Butterfly	0.68	$1.70*10^{-29}$	-	-
Freestyle	$8.57*10^{-26}$	$1.29*10^{-46}$	$8.24*10^{-26}$	-
Medley	$7.13*10^{-04}$	$1.75*10^{-26}$	$2.39*10^{-03}$	$3.24*10^{-31}$

Table 5: p-value of pairwise comparison with Holm-Bonferroni adjustment method.

5 Summary

The data set analyzed in this report was put together by the instructors of the course Introductory Case Studies at TU Dortmund University during the winter term 2022/23 and originates from the web portal "European Aquatic Roma 2022." It includes 80 observations, of which 8 were duplicates that were removed, so we used 72 observations in our analysis. The data set contains the nominal variable *Category* which consists of five types of swimming events: backstroke, breaststroke, butterfly, freestyle, and medley. After removing the duplicate values, the first two types contain 13 participants each, butterfly has 13, freestyle has 15, and medley contains all 16 observations.

The purpose of this report is to analyze the time difference among the categories and the pairwise time difference between two categories. For the first task, we used an ANOVA test to see whether two categories were varying in terms of average time. We found out that there is a significant difference between the categories. We used the p-value approach to reject the null hypothesis where significance level (α) was assumed .05. In the second task, we investigated the pairwise difference between the categories. Here, we used a pairwise t-test, which gives us p-values that are much smaller than the significance level (.05). Therefore, we rejected the null hypothesis, and all nine pairs of events differ except one, which is butterfly and backstroke. Since this task requires performing multiple statistical tests, the risk of a family-wise error rate increases. To address this issue, we introduced two correction methods such as Bonferroni and Holm-Bonferroni, where each method gives the same result though the p-value is smaller in Holm-Bonferroni compared to the Bonferroni method. However, the final outcome is the same, i.e., all nine pairs of events are different in terms of timing except one pair of events called the butterfly and backstroke.

In this report, a very small data set (80 observations) was used, so the variances among the observations couldn't be understood precisely. The data set also includes swimmers who participated in multiple categories. Such duplicates create bias in small data sets. In further analysis, it might be of interest to include a large data set with other important variables like body size (leg or arm size). Furthermore, the age and height of the participants are important, so these variables can also be included in the data set.

Bibliography

Amitav Banerjee, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18(2):127, 2009.

Ken Black, John Asafu-Adjaye, Paul F Burke, Nazim Khan, Gerard King, Nelson Perera, Andrew Papadimos, Carl Sherwood, and Saleh Ahmed Wasimi. *Business analytics and statistics*. John Wiley & Sons Australia, Limited, 2018.

Jean-Baptist Du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(19):335, 2009.

Tarek M El-gohary. Hypothesis testing, type i and type ii errors: Expert discussion with didactic clinical scenarios. *International Journal of Health and Rehabilitation Sciences (IJHRS)*, 8(3):132, 2019.

European Aquatics Championships. European Aquatics Roma, 2022. URL `https://www.roma2022.eu/en/`. [Visited on 23-11-2022].

Guifang Fu, Garrett Saunders, and John Stevens. Holm multiple correction for large-scale gene-shape association mapping. In *BMC genetics*, volume 15, pages 1–8. BioMed Central, 2014.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL `https://CRAN.R-project.org/package=ggpubr`. R package version 0.4.0.

LEN European Aquatics, 2022. URL `https://roma2022.microplustimingservices.com`. [Visited on 23-11-2022].

MicroplusSrl. Microplus, 2022. URL `https://www.microplus.it/`. [Visited on 23-11-2022].

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

John W Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer, New York, 2010. ISBN 9781441923226 1441923225. URL http://www.amazon.de/All-Statistics-Statistical-Inference-Springer/dp/1441923225/ref=sr$_{12}$?$ie = UTF8qid = 1356099149sr = 8 - 2$.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. URL https://CRAN.R-project.org/package=cowplot. R package version 1.1.1.