TU Dortmund

Introductory Case Studies

Project II: Comparison Of Multiple Distributions

Lecturers:
Prof. Dr. Jörg Rahnenführer
Dr. Franziska Kappenberg
M. Sc. Marieke Stolte

Author: Bushra Tariq Kiyani

Group number: 5

Group members: Prerana Rajeev Chandratre, Sathish Ravindranth, Shivam Shukla, Janani Veeraraghavan

December 9, 2022

Contents

1 Introduction 1

2 Problem statement 2

2.1 Data set and data quality . . . 2

2.2 Project objectives . . . 3

3 Statistical methods 3

3.1 Hypothesis Testing . . . 3

3.2 Global Testing . . . 5

3.2.1 One-Way ANOVA Test . . . 5

3.3 Pairwise Testing . . . 7

3.3.1 T-Test . . . 7

3.3.2 Multiple T-Tests . . . 8

3.4 Multiple Test Adjustment Methods . . . 8

3.4.1 The Bonferroni Correction . . . 9

3.4.2 Bonferroni–Holm Method . . . 9

3.5 Graphical Tools . . . 9

3.5.1 Q-Q Plot . . . 10

4 Statistical analysis 10

4.1 Verifying the Assumptions . . . 12

4.2 One-Way ANOVA: Testing Significance of finishing time differences between the five categories . . . 14

4.3 Multiple t-tests . . . 15

4.4 Bonferroni's correction . . . 16

4.5 Bonferroni–Holm Method . . . 16

5 Summary 17

Bibliography 19

1 Introduction

The European Aquatics Championships is the aquatics championship for Europe, organized by a European governing body for aquatic sports, LEN (French: Ligue Européenne de Natation). LEN was founded in 1927. The Championships are currently held every two years. They included five water sports: swimming (long course/50m pool), diving, artistic swimming, open water swimming, and high diving for 2022. Before 1999, water polo was also included but later moved to the European Water Polo Championships (LEN European Aquatics, 1927). In 2022 the European Aquatics Championships took place from 11 to 21 August in Rome, Italy, and consisted of 75 medal events in all disciplines (The EAC Rome, 2022).

This project aims to compare the multiple distributions of a small sample taken from the data set (The EAC Semi-Final Results, 2022). The data set contains the women's 200m swimming semi-final results (time in seconds) in the five categories of swimming: backstroke, breaststroke, butterfly, freestyle, and medley. We want to check time variations between the five categories. To do this, we first look at the frequency distributions of finishing time using histograms. We observe the mean and median of all distributions to compare different distributions. Box plots are used to study the variations within and between all categories. After that, we check the distribution of the data using Q-Q Plots. Then to test whether times differ between the categories, we perform ANOVA. Finally, to observe the pairwise differences in times of all categories, we conduct two-sample tests for all pairs, and then compare the test results after adjusting with the Bonferroni and Bonferroni-Holm methods.

The second section describes the structure and quality of the data set in more detail. Additionally, we state the goals of the project in the second section. The third section explains the hypothesis testing, one-way ANOVA, Bonferroni, and Bonferroni-Holm methods. It also describes the graphical representations used in statistical analysis (histograms, boxplots, Q-Q plots). The fourth section focuses on the application of these methods and the interpretation of the graphs and results. Finally, the fifth section summarizes the most important findings and discusses possible further analyses of the data set.

2 Problem statement

2.1 Data set and data quality

This report deals with the analysis of a small sample of the data set of the women's semi-final results in the European Aquatics Championships, taken from the website (The EAC Semi-Final Results, 2022). The website is an official database of the results of all five categories: swimming, diving, artistic swimming, open water swimming, and high diving. It contains data of 557 athletes (both men and women), who participated in the event from 42 member countries of LEN. It contains start lists, results, medalists by events, a list of broken records, trophies, medal standings and complete results of each major five categories. Data is gathered by the LEN officials so we expect high accuracy.

We have a small sample of this data set, which includes only women's semi-final results in the swimming category. In swimming, there are further five sub-categories: backstroke, breaststroke, butterfly, freestyle, and medley. The data set consists of 80 observations and 3 variables. Description of the variables according to the (The EAC Semi-Final Results, 2022) is given below:

Variable	Type	Description
Category	Nominal	Name of the swimming category. The data set contains 5 categories.
Name	Nominal	The name and surname of the player. In this, both the surname and the first name are listed and separated by capital letters.
Time	Numeric	It shows the finishing time of each player in seconds.

Table 1: Data Description

We do not observe missing values in the data set but we do note that some swimmers Anastasya Gorbenko, Katie Shanahan, Kristyna Horska, Laura Lahtinen, Mireia Belmonte Garcia, Katinka Hosszu, and Marrit Steenbergen competed in more than one category. Which may violate the assumption of independence and introduce bias that affects the test and overall results. So we need to find a way to remove this dependency by only putting these players in one category. As the data set is small, we'll not remove the data of these athletes from all categories.

2.2 Project objectives

In this project, we analyze whether the variation of times between the five sub-categories of swimming is significant. We start by observing the frequency distribution of times in all categories using histograms. The result is interpreted by looking at the mean and spread of the distributions. Then, for the analysis of the variability in the mean finishing time of all categories, we verify that the three assumptions that are needed before applying ANOVA and t-test are satisfied. For fulfilling the independence assumption we put each swimmer in exactly one category. Box plots are used to check the homogeneity of the variances in each category and Q-Q plots are used to check normally distributed data assumption. Then the ANOVA test is performed to compare the mean time differences between categories.

To check for pairwise differences between the finishing times, we perform pairwise t-tests. The results are interpreted by comparing the p-value at a significance level of 0.05. Then to address the multiple testing problem, we adjust the results using Bonferroni, and Bonferroni-Holm adjustment methods. Results are interpreted and compared with and without adjusting for multiple testing.

3 Statistical methods

This section discusses statistical methods used to analyze the variability and significant differences between the times of different categories. Hypothesis tests, one-way ANOVA test, Bonferroni correction, and Bonferroni-Holm methods are presented. The Q-Q plot (used to assess the normality of data) is also discussed.

3.1 Hypothesis Testing

A statistical hypothesis is a statement about the nature or distribution of one or more random variables. Hypothesis testing is an inferential statistical method of analyzing whether a hypothesis can be accepted or rejected. It helps us to determine the characteristics of a population by analyzing a sample data set. We formulate two types of hypotheses. The first is null hypothesis H_0, the other one is alternative hypothesis H_1. The null and the alternative hypothesis are set up before performing the hypothesis testing (Mood et al., 1973, p. 402).

Null Hypothesis H_0: The null hypothesis is a concise mathematical statement which indicates that there is no difference between certain characteristics of data. It can be simple or composite. A simple hypothesis states the value of the parameter θ uniquely such that the hypothesis $H_0 : \theta = \theta_0$ is simple. A composite hypothesis states that the parameter can have multiple values. In the testing process, H_0 is analyzed using sample data and the analysis determines whether H_0 can be accepted or rejected (Mood et al., 1973, p. 405).

Alternative Hypothesis H_1: It is an alternative to the null hypothesis and shows that there is statistical significance between two outcomes. It is denoted as H_1 or H_A. If H_0 is rejected, then the alternative hypothesis H_1 is accepted. The decision whether to accept or reject a hypothesis is taken by comparing the level of significance α and the p-value (Mood et al., 1973, p. 405).

Test Statistic: A test statistic (e.g. t-value/F-value) is calculated from a statistical test. It summarizes the observed data into a single number and shows how closely observed data match the distribution expected under the null hypothesis. It is used to calculate the p-value, which helps in deciding whether to reject the null hypothesis (Mood et al., 1973, p. 419).

P-Value and Level of Significance α: In hypothesis testing, the p-value indicates whether the results obtained after conducting a statistical test are statistically significant or not. It also tells the probability of making an error while rejecting or not rejecting a null hypothesis. Its value is always between $[0, 1]$. The significance level α is a measure of the strength of evidence required to reject the null hypothesis. Typically, α is chosen to be $5\%(0.05)$ or $1\%(0.01)$. A lower value of α indicates that stronger evidence is needed to reject the null hypothesis. The p-value is compared with the significance level α. We reject the null hypothesis, If the p-value is smaller than the significance level α (Mood et al., 1973, p. 402-403).

Types of Error: There are two types of error in hypothesis testing: Type I errors (False Positive) and Type II errors (False Negative), both are related to the wrong conclusion about the null hypothesis. A type I error occurs if we accept the null hypothesis when it is false. A type II error occurs if we reject the null hypothesis when it is actually true. The size of a Type I error is defined to be the probability that a Type I error is made; which is α, and similarly the size of a Type II error is the probability that a Type II error is made; which is $\beta = 1 - \alpha$ (Mood et al., 1973, p. 405).

3.2 Global Testing

3.2.1 One-Way ANOVA Test

The Analysis of variance(ANOVA) is a statistical technique used to check if the means of two or more groups are significantly different from each other or not. It analyzes the levels of variance between and within the groups. There are many variations of ANOVA but the main two types are one-way ANOVA and two-way ANOVA. In the one-way ANOVA, we compare the effect of one independent variable on the dependent variable. (Rasch et al., 2020, p. 108)

Grand mean: In ANOVA calculations there are two types of the mean: separate sample means $(\mu_1, \mu_2, ..., \mu_k)$ and the grand mean (μ). The grand mean is the mean of all sample means. We'll denote empirical grand mean by $\bar{x_G}$ and sample means as $\bar{x_1}, ..., \bar{x_k}$, then the grand mean is defined as (Rasch et al., 2020, p. 108):

$$\bar{x_G} := \frac{1}{k}\sum_{i=1}^{k}\bar{x_i}$$

Hypothesis: In ANOVA the null and the alternate hypothesis are:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k(\text{means of all groups are equal})$$

$$H_1 : \text{at least one } \mu_i \text{ is different from the other means}$$

The null hypothesis is valid when the sample means don't have significant differences while the alternate hypothesis is accepted when at least one of the sample means is different from the rest (Rasch et al., 2020, p. 114).

Variability Between Groups: It is the variation between the distributions of the individual groups. To calculate between-group variability the difference between the individual sample means and the grand mean is calculated. If the samples deviate greatly from each other, the difference between the individual mean and the overall mean would therefore also be significant. For sample sizes $n_1, ..., n_k$, the sum of squares for between-groups variability $\text{SS}_{\text{Between}}$, is calculated as (Rasch et al., 2020, p. 108-112):

$$\text{SS}_{\text{Between}} = \sum_{i=1}^{k} n_i(\bar{x_i} - \bar{x_G})^2$$

To calculate between-group mean squared deviation we divide the $\mathrm{SS}_{\mathrm{Between}}$ by degrees of freedom (df_{B}): $\mathrm{K} - 1$ where K is a number of sample means. Between-group mean squared deviation is denoted as $\mathrm{MS}_{\mathrm{Between}}$ and calculated as (Rasch et al., 2020, p. 108-112):

$$\mathrm{MS}_{\mathrm{Between}} = \frac{\mathrm{SS}_{\mathrm{Between}}}{\mathrm{K} - 1}$$

Variability Within Groups: As the variability of each sample is increased, their distributions overlap and they become part of a large population. Because values within each group are not the same this causes this variation. So the variance between individual points in each sample is calculated separately and referred to as within-group variability $\mathrm{SS}_{\mathrm{Within}}$ (Rasch et al., 2020, p. 108-112).

$$\mathrm{SS}_{\mathrm{Between}} = \sum_{i=1}^{K} \sum_{j=1}^{n_j} (x_{ij} - \bar{x_i})^2$$

With degrees of freedom (df_{W}): N - K (N is a total number of observations) within-group mean squared deviation $\mathrm{MS}_{\mathrm{Within}}$ is calculated as (Rasch et al., 2020, p. 108-112):

$$\mathrm{MS}_{\mathrm{Within}} = \frac{\mathrm{SS}_{\mathrm{Within}}}{\mathrm{N} - \mathrm{K}}$$

Total variability can be calculated by adding both between and within groups variances.

F-Statistic: F-Ratio (F) is the statistic which measures if the means of different groups are significantly different or not. A lower value of F-Ratio indicates that means are similar and we cannot reject the null hypothesis (H_0) (Rasch et al., 2020, p. 108).

$$\mathrm{F} = \frac{\mathrm{MS}_{\mathrm{Between}}}{\mathrm{MS}_{\mathrm{Within}}}$$

ANOVA Assumptions: There are three primary assumptions about the data. These assumptions must be met before applying ANOVA: Normality: The responses for each sample group are taken from a normal population distribution. Equal variances: The variances of the populations that the samples come from are equal. Independence: All samples are drawn independently of each other (Rasch et al., 2020, p. 108-109).

Problems with One-Way ANOVA: If the overall p-value from the ANOVA table is less than a certain level of significance, then we have enough evidence that at least one group mean is different from the other. However, this does not tell us which groups differ from each other. It just tells us that not all group means are equal (Rasch et al., 2020, p. 108-114).

3.3 Pairwise Testing

As ANOVA does not provide any information about the pairwise differences among the groups. Multiple pairwise t-tests between each of the pairs of interest can be performed to provide this type of detailed information (Christopher, 2019, p. 251).

3.3.1 T-Test

A t-test is a statistical test which is used to compare the means of two groups. It is used in hypothesis testing to determine whether two groups are different from each other or not. It makes the same assumptions about the data as ANOVA. It assumes that data is independent, taken from a normal distribution and has an equal variance (Christopher, 2019, p. 251).

One-tailed or Two-tailed t-test: If we want to know whether the two populations are different from one another, we perform a two-tailed t-test. While to know whether one population mean is greater than or less than the other, we perform a one-tailed t-test (Christopher, 2019, p. 251-253).

Hypothesis: In the t-test the null and alternate hypotheses are (Christopher, 2019, p. 255).:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Performing a t-test: The t-test estimates the difference between two group means using the ratio of the difference in group means to the pooled standard error of both groups (Christopher, 2019, p. 273-274):

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

In this formula, t is the t statistic/t value, x_1 and x_2 are the means of the two groups being compared, s^2 is the pooled standard error of the two groups, and n_1 and n_2 are the numbers of observations in each group (Christopher, 2019, p. 273-274).

A large t value indicates that the difference between the group means is greater than the standard error, indicating a significant difference between the groups. We compare the calculated t value against the values in a critical value chart (e.g., Student's t table) to determine whether the t value is greater if so, we can reject the null hypothesis and conclude that the two groups are different (Christopher, 2019, p. 273-274).

3.3.2 Multiple T-Tests

T-test compares means of two groups, so to compare means of more than two groups we can perform multiple t-tests to check all pairwise differences. But there is one problem with this approach. As each one of those independent tests has an α level and α level is a type I error rate. Which means 95% confidence. But the error compounds with each test. This alpha (α) inflation is caused by performing repeated statistical tests on the same data. Which will lead to a higher probability of making a Type I error and lower confidence (Christopher, 2019, p. 268).

Family-Wise Error Rate (FWER/FWE): Family-wise error rate is the probability of making at least one Type I error. If m independent comparisons are performed and α is a significance level for an individual test, the family-wise error rate (FWER), is given by:

$$\tilde{\alpha} = 1 - (1 - \alpha)^m$$

So in multiple t-testing, the probability of getting a significant result simply due to chance keeps increasing (Matsunaga, 2007).

3.4 Multiple Test Adjustment Methods

Methods which deal with multiple testing adjust α in some way so that the probability of observing at least one significant result due to chance remains below the desired significance level. These tests are referred to as "multiple comparison" tests. Commonly used multiple comparison analysis statistics include the following tests: Tukey, Bonferroni, Bonferroni-Holm and Dunnett. We are using Bonferroni and Bonferroni-Holm and comparing both with multiple t-tests in this project (Christopher, 2019, p. 268).

3.4.1 The Bonferroni Correction

The assumptions for this procedure are the same as ANOVA. That is random variables are independent and normally distributed with equal variance. The Bonferroni correction fixes the significance level at $\frac{\alpha}{m}$. The Bonferroni correction is slightly conservative. If p_i is the p-value of hypothesis H_i, the Bonferroni correction rejects the null hypothesis if

$$p_i \leq \frac{\alpha}{m}$$

thereby controlling the FWER at $\leq \alpha$. Alternatively, we can compare each $m{*}pi$ against joint significance level α. If the p-value becomes greater than 1, then its value is adjusted to 1 (Christopher, 2019, p. 274).

3.4.2 Bonferroni–Holm Method

The Holm-Bonferroni method (or Holm's Sequential Bonferroni Procedure) is another way to deal with familywise error rates (FWER) for multiple hypothesis tests. It is a modification of Bonferroni Correction but provides a uniformly more powerful test than Bonferroni.

If we have m p-values which are sorted into order lowest-to-highest $P_1, ..., P_m$ and corresponding hypotheses $H_1, ..., H_m$ and we want the FWER to be no higher than a pre-specified significance level α. The Bonferroni–Holm method is as follows: Start with the hypothesis H_k with the lowest p-value and check for each p$_k$-value, whether

$$p_k < \frac{\alpha}{m + k + 1}$$

holds. If yes, then reject H_k and continue to examine the larger P values, otherwise exit (Aickin and Gensler, 1996).

3.5 Graphical Tools

Different data visualization methods are helpful to analyze data assumptions. These graphical tools are chosen according to the type of data (Crowder et al., 2020, p. 63).

3.5.1 Q-Q Plot

A Q-Q plot (quantile-quantile plot) is a probability plot and a graphical tool that is used to compare two probability distributions. A Q-Q plot is a scatterplot which is created by plotting quantiles of the distributions to be compared, against one another. We should see the points forming a roughly straight line ($y = x$) if both sets of quantiles came from the same distribution.

Quantiles are often referred to as percentiles, these are points in our data sorted in ascending order, below which a certain proportion of the data fall. A Q-Q plot is just a visual check, not proof, it allows us to see at-a-glance if our assumption is plausible or not (Christopher, 2019, p. 146).

Normal Probability Q-Q Plot: Normal probability Q-Q plots give us a way for comparing the distribution of a sample against standard normal distribution. Normal probability Q-Q plot takes our sample data, sort it in ascending order, and then plots them versus quantiles calculated from a theoretical standard normal distribution. If the points seem to fall about a straight line we can assume that the sample data came from a normally distributed population (Christopher, 2019, p. 147).

4 Statistical analysis

In this section, a detailed description of the application of statistical test methods mentioned in the previous section on our data set is discussed. The R software (R Core Team, 2022) Version: 4.2.1 with additional packages **ggpubr** (Kassambara, 2022a), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **rstatix** (Kassambara, 2022b), **gridExtra**(Auguie, 2017) is used for statistical testing, to calculate all statistical measures and to visualize graphical representations.

First, we perform a descriptive analysis of the data set. Table 2 summarizes the descriptive statistical information about the data. We see the highest mean (146.314) and median (146.660) values of finishing times are in the Breaststroke category while Freestyle has the lowest mean (119.358) and median (119.500). There is very little difference in the mean (131.380) of Backstroke and mean (131.656) of Butterfly. Butterfly has the highest standard deviation (2.61). The overall minimum finishing time (117.7) lies in Freestyle while the overall maximum finishing time (148.4) is in Breaststroke.

	Category	median	mean	sd	variance	minimum	maximum	IQR
1	Backstroke	131.115	131.380	1.85	3.43	128.18	135.74	1.52
2	Breaststroke	146.660	146.314	1.51	2.27	143.73	148.40	2.00
3	Butterfly	130.835	131.656	2.61	6.82	128.48	136.92	4.42
4	Freestyle	119.500	119.358	1.56	2.43	117.70	122.42	2.23
5	Medley	133.455	134.040	1.59	2.52	131.84	137.02	2.47

Table 2: Summary Statistics

Frequency Distribution of finishing time: Figure 1 shows frequency distribution of finishing time in all categories. If we look at the overall frequency distribution of finishing time it’s almost normal.

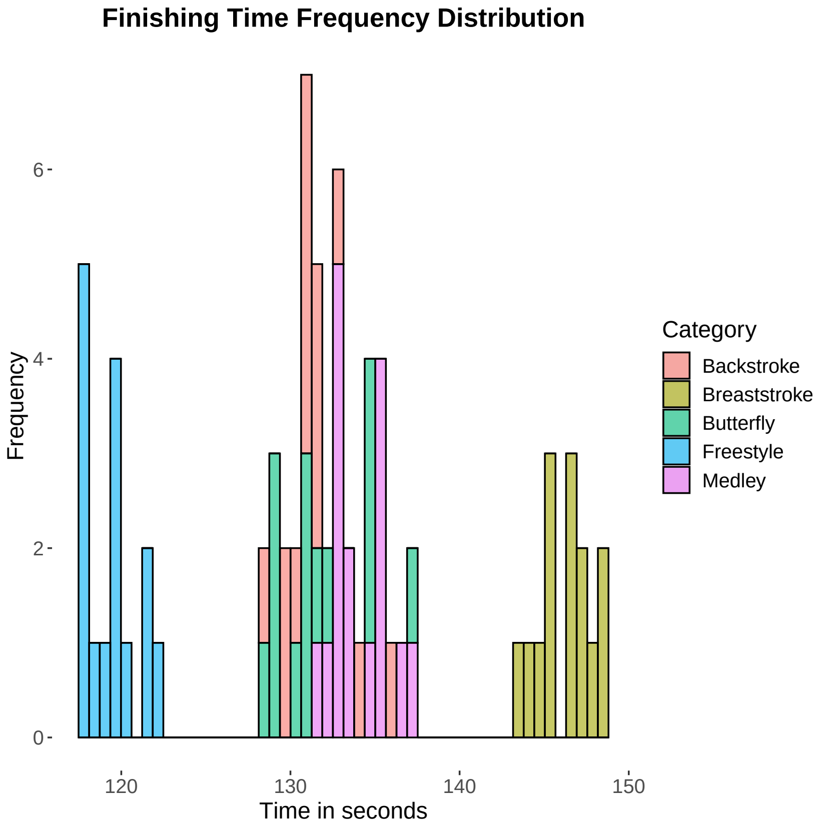

Figure 1: Frequency Distribution of finishing Time in all Categories

It can be seen that Freestyle has overall less time than all the other categories while Breaststroke has the highest finishing time range. The finishing time of the other three categories lies in middle. Frequency from 132-135 seconds is the highest. In Freestyle, we see the highest frequency 5 at 117 seconds. Overall highest frequency 6 is observed.

4.1 Verifying the Assumptions

As we discussed in the previous section, in order to perform ANOVA and t-tests, we need to verify certain assumptions, only when those assumptions are met can we perform these tests. In this section, we'll discuss if our data is fulfilling these assumptions.

Normality: We want to check whether the sample data is normally distributed or not. Figure 2 shows Normal probability Q-Q plots of all five categories. We observe that data points do not follow the reference line well, mostly on the edges. Nevertheless, we know our sample data is drawn from a normally distributed population. So, we assume that our sample data is normally distributed. Because of the small sample size, it might show little deviation from normality.

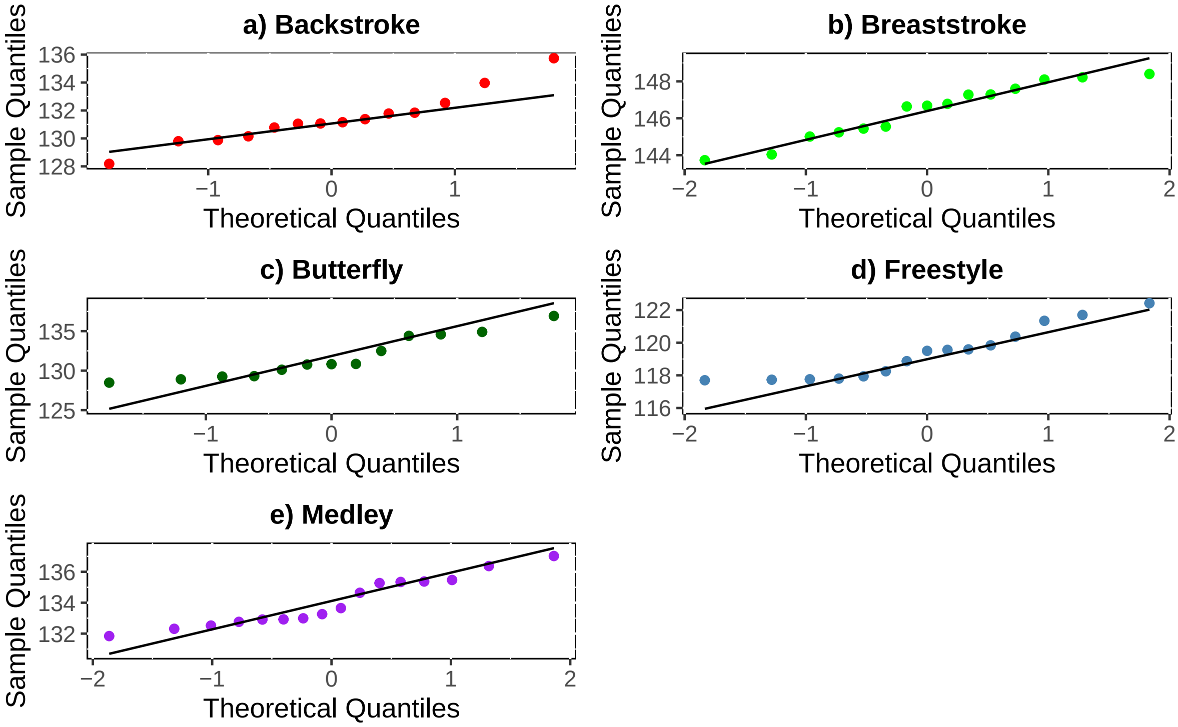

Figure 2: Normal Q-Q Plots for Swimming categories (a) Backstroke, (b) Breaststroke, (c) Butterfly, (d) Freestyle, (e) Medley

Independence: In the data set (The EAC Semi-Final Results, 2022) six swimmers appeared in more than one category. Which can violate the assumption of independence. We observe that five out of six swimmers Anastasya Gorbenko, Katie Shanahan, Kristyna Horska, Mireia Belmonte Garcia, Katinka Hosszu, and Marrit Steenbergen participated in the Medley, so we keep these swimmers in the Medley category and remove from the other categories. One swimmer Laura Lahtinen competed in the Butterfly and Breaststroke, as we already removed two swimmers Mireia Belmonte Garcia and Katinka Hosszu from the Butterfly and only one swimmer Kristyna Horska from the Breaststroke; we'll keep Laura Lahtinen in the Butterfly (in which we have fewer swimmers).

Variance Homogeneity: The third assumption we need to verify is constant variance across all distributions. To verify variance homogeneity we use boxplots. Figure 3 shows box plots of finishing time in all categories. We compare box widths in a boxplot. We see box widths of all boxes are almost similar except for the Butterfly category. But it's also not that much large. So overall there's not much difference in boxplot widths. Hence we assume that variance is constant.

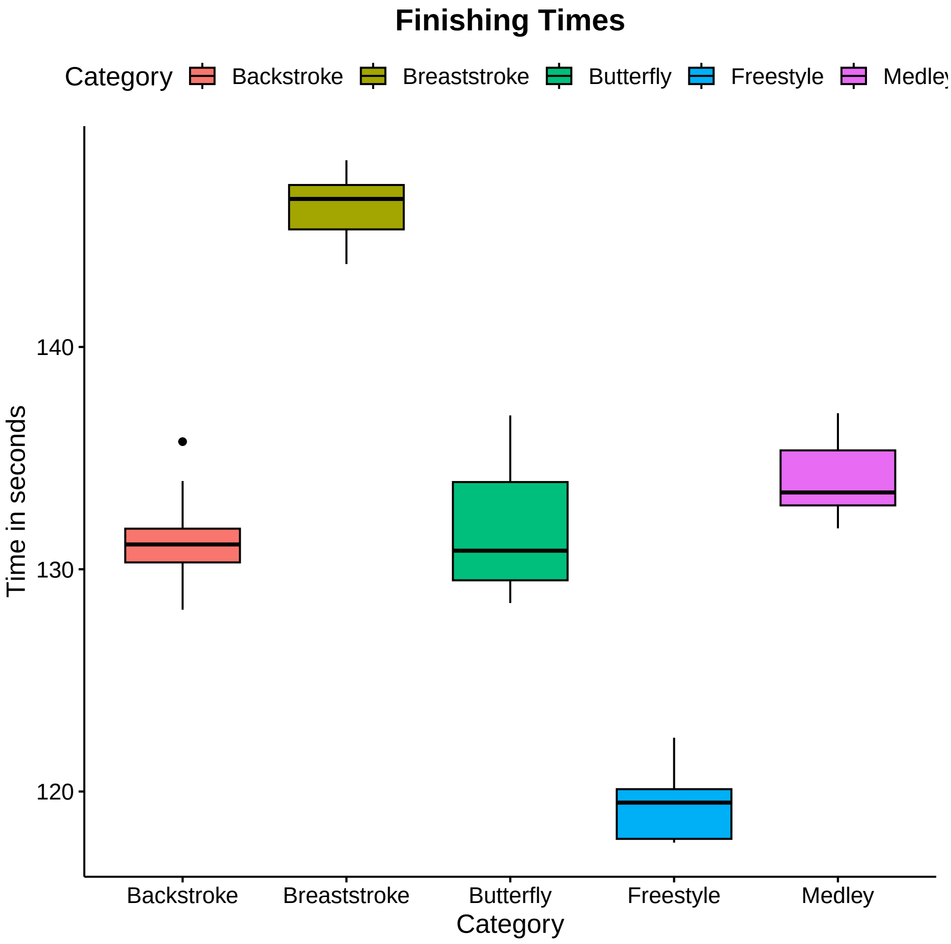

Figure 3: Box plots of finishing time in all categories

4.2 One-Way ANOVA: Testing Significance of finishing time differences between the five categories

As the three assumptions needed to perform ANOVA are fulfilled, now to test the significance of differences in the finishing time of all categories we'll conduct an ANOVA test. We take μ as the mean finishing time of the categories and specify significance level $\alpha = 0.05$. We formulate our null and alternate hypotheses as follows:

$$H_0 : \mu_{backstroke} = \mu_{breaststroke} = \mu_{butterfly} = \mu_{freestyle} = \mu_{medley}$$

$$H_1 : \text{at least one } \mu \text{ is different from the other means}$$

We conducted ANOVA in R, Table 3 shows the resulting output.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Category	4	5327	1331.8	385.9	<2e-16 ***
Residuals	68	235	3.5		

Table 3: ANOVA Output

In the table, the Category is showing values between groups. We see the degree of freedom (DF) is $(K-1=4)$, the sum of the square between categories (Sum Sq) is 5327 and the Mean square between categories is $\left(\frac{5327}{4} = 1331.8\right)$. Residuals show the values within groups. DF is $(N-K=68)$, the sum of the square within is 235 while the Mean square is $\left(\frac{235}{68} = 3.5\right)$. The F value is then calculated. We see p-value (Pr(>F)) is less than the significance level $\alpha = 0.05$, so we'll reject the null hypothesis that the mean finishing time across all the categories is equal.

If the total p-value from the ANOVA table is less than a certain level of significance, then we have sufficient evidence that at least one of the means of the groups is different from the others.

We have concluded that the mean finishing time is different across given categories, to look at the pairwise significance of this difference we perform multiple pairwise t-tests.

4.3 Multiple t-tests

In this section, we'll discuss multiple t-tests and results. We have already fulfilled the assumptions required to perform a t-test. We make all possible pairs of our categories as:

"Backstroke_Breaststroke", "Backstroke_Butterfly", "Backstroke_Freestyle",

"Backstroke_Medley", "Breaststrok_Butterfly", "Breaststroke_Freestyle",

"Breaststroke_Medley", "Butterfly_Freestyle", "Butterfly_Medley", "Freestyle_Medley"

For the first pair we formulate hypotheses as follows: T-test1:

$$H_0 : \mu_{Backstroke} = \mu_{Breaststroke}$$

$$H_1 : \mu_{Backstroke} \neq \mu_{Breaststroke}$$

Likewise, we formulate hypotheses for all pairs and specify significance level $\alpha = 0.05$. After running the t-test separately on each pair in R, we get the results given in Table 4.

	Categories pair	p-values	Reject Yes/No
1	Backstroke_Breaststroke	< 0.01	Yes
2	Backstroke_Butterfly	0.75	No
3	Backstroke_Freestyle	< 0.01	Yes
4	Backstroke_Medley	< 0.01	Yes
5	Breaststroke_Butterfly	< 0.01	Yes
6	Breaststroke_Freestyle	< 0.01	Yes
7	Breaststroke_Medley	< 0.01	Yes
8	Butterfly_Freestyle	< 0.01	Yes
9	Butterfly_Medley	< 0.01	Yes
10	Freestyle_Medley	< 0.01	Yes

Table 4: Multiple T-Test Results

We see for Backstroke_Butterfly the p-value is 0.75 which is greater than $\alpha = 0.05$, so in this case, the null hypothesis is not rejected. It means that there is not a significant difference in the mean finishing time of Backstroke and Butterfly. If we look at the mean finishing time values for Backstroke = 131.38 and for Butterfly = 131.66. We see there is not much difference in their mean values. (Note time is in seconds). For all other pairs, the p-value is less than the α. So all other null hypotheses for these pairs are rejected. This shows that there is a significant difference in the mean finishing time between these pairs.

4.4 Bonferroni's correction

To deal with the family-wise error rate we use Bonferroni's correction method. P-values from multiple t-tests are adjusted using Bonferroni's method in R and the result is displayed in Table 5.

	Categories pair	Adjusted p-values	Reject Yes/No
1	Backstroke_Breaststroke	< 0.01	Yes
2	Backstroke_Butterfly	1.00	No
3	Backstroke_Freestyle	< 0.01	Yes
4	Backstroke_Medley	< 0.01	Yes
5	Breaststroke_Butterfly	< 0.01	Yes
6	Breaststroke_Freestyle	< 0.01	Yes
7	Breaststroke_Medley	< 0.01	Yes
8	Butterfly_Freestyle	< 0.01	Yes
9	Butterfly_Medley	0.05	Yes
10	Freestyle_Medley	< 0.01	Yes

Table 5: Bonferroni's correction results

We observe that the p-value for Backstroke_Butterfly (1.00) is increased as compared to the p-value (0.75) in multiple t-tests. The p-value for Butterfly_Medley is changed from < 0.01 to 0.05. Similarly, we observe an overall increase in the p-values of other pairs although in this data set this increase is not significant and the final result of rejecting the null hypothesis is the same as multiple t-tests. We'll not reject the null hypothesis for Backstroke_Butterfly. All other null hypotheses are rejected.

4.5 Bonferroni–Holm Method

The Bonferroni–Holm method is another method to deal with the family-wise error rate. It is less conservative than Bonferroni's correction. Table 6 shows the results after applying the Bonferroni-Holm method. We see that the p-value for Backstroke_Butterfly (0.75) is the same as in multiple t-tests but the p-value for Butterfly_Medley is changed from < 0.01 to 0.01.

Final results after applying Bonferroni–Holm Method regarding rejecting null hypotheses stay the same as for the above methods. We don't reject the null hypothesis for Backstroke_Butterfly and all other null hypotheses are rejected. This shows that the finishing time of Backstroke_Butterfly is not significantly different while for all other pairs it's significantly different from each other.

	Categories pair	Adjusted p-values	Reject Yes/No
1	Backstroke_Breaststroke	< 0.01	Yes
2	Backstroke_Butterfly	0.75	No
3	Backstroke_Freestyle	< 0.01	Yes
4	Backstroke_Medley	< 0.01	Yes
5	Breaststroke_Butterfly	< 0.01	Yes
6	Breaststroke_Freestyle	< 0.01	Yes
7	Breaststroke_Medley	< 0.01	Yes
8	Butterfly_Freestyle	< 0.01	Yes
9	Butterfly_Medley	0.01	Yes
10	Freestyle_Medley	< 0.01	Yes

Table 6: Bonferroni–Holm Method results

5 Summary

In this project, a comparison of multiple distributions on the data set containing the women's 200m swimming semi-final results (time in seconds) in the five categories of swimming: backstroke, breaststroke, butterfly, freestyle, and medley was performed. The data set was taken from the European Aquatics Championships website (The EAC Semi-Final Results, 2022) and we were interested in analyzing the finishing times of swimmers in all categories. The aim was to find out whether there was a significant difference in the finishing times across all categories or not. We performed a global test (ANOVA) and then multiple t-tests for pairwise differences were performed. Then two multiple-test adjustment methods Bonferroni correction and Bonferroni–Holm were used and compared to each other and to multiple t-test results.

We looked at the summary statistics of the data set which showed that Breaststroke had the highest finishing time while the lowest finishing time was in Freestyle. The Butterfly had the highest standard deviation. The frequency distribution graph also showed overall normal distribution. It also showed that the overall finishing time of Freestyle was less than all other categories and in Breaststroke overall finishing time was the highest.

Before starting the test, we checked the assumptions that needed to be satisfied before ANOVA and t-test. We confirmed data independence by keeping all swimmers in just one category. Normality was checked by Q-Q plots and we ensured homogeneity of variance by analyzing boxplots of finishing time. The ANOVA result showed a p-value less than $\alpha = 0.05$ which meant there was a significant difference between the mean finishing times of all categories but it didn't tell about pairwise differences in finishing times of all categories.

To know pairwise differences we performed multiple pairwise t-tests which resulted in all p-values less than α except Backstroke_Butterfly whose p-value was 0.75. So only for Backstroke_Butterfly, we did not reject the null hypothesis. Multiple tests suffer from the family-wise error rate and to cope with that we used adjustment methods. After applying Bonferroni's Correction and Bonferroni–Holm Method p-values increased, the increment was more in Bonferroni's Correction than the Bonferroni–Holm Method. In both methods, we rejected the null hypothesis only for Backstroke_Butterfly. So in all other pairs testing results showed that the difference in finishing time was significant.

Bibliography

M Aickin and H Gensler. Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American Journal of Public Health*, 86(5):726–728, 1996. doi: 10.2105/AJPH.86.5.726. URL `https://doi.org/10.2105/AJPH.86.5.726`. PMID: 8629727.

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.

Hay-Jahans Christopher. *Introduction to the theory of statistics.* Taylor Francis Group, 2019. 1st Edition.

Stephen Crowder, Collin Delker, Eric Forrest, and Nevin Martin. *Introduction to Statistics and Probability*, pages 59–80. Springer International Publishing, Cham, 2020. ISBN 978-3-030-53329-8. URL `https://doi.org/10.1007/978-3-030-53329-8`$_4$.

Alboukadel Kassambara. *ggpubr: ggplot2' Based Publication Ready Plots*, 2022a. URL `https://rpkgs.datanovia.com/ggpubr/`. R package version 3.4.

Alboukadel Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2022b. URL `https://rpkgs.datanovia.com/rstatix/`. R package version 3.3.

LEN European Aquatics. French: Ligue Européenne de Natation (LEN), English: European Swimming League, 1927. URL `https://www.len.eu/`. Accessed: 18-11-2022.

Masaki Matsunaga. Familywise error in multiple comparisons: Disentangling a knot through a critique of o'keefe's arguments against alpha adjustment. *Communication Methods and Measures*, 1(4):243–265, 2007. doi: 10.1080/19312450701641409. URL `https://doi.org/10.1080/19312450701641409`.

Alexander McFarlane Mood, Franklin A Graybill, and Duane C Boes. *Introduction to the theory of statistics.* McGraw-Hill, 1973. 3rd Edition.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Dieter Rasch, Rob Verdooren, and Jürgen Pilz. *Applied Statistics.* WILEY, 2020. 1st Edition.

The EAC Rome. The European Aquatics Championships (EAC) Rome, 2022. URL https://www.roma2022.eu/en/. Accessed: 18-11-2022.

The EAC Semi-Final Results. The European Aquatics Championships (EAC) Rome, Semi-Final Results, 2022. URL https://roma2022.microplustimingservices.com/indexRoma2022$_w$*eb.php.Accessed* : $18 - 11 - 2022$.

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

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.