Comparison of multiple distributions

A CASE STUDY

Fahad Deshmukh
UNIVERSITY OF POTSDAM | DESHMUKH@UNI-POTSDAM.DE

Table of Contents

1 Introduction

Swimming is a popular and exhilarating sport enjoyed by people of all ages around the world. It not only provides a great way to stay fit but also serves as a competitive arena for athletes to showcase their skills and abilities. One of the prestigious competitions in the realm of swimming is the European Aquatics Championships, a continent-wide sporting event that occurs biennially. Organized by LEN European Aquatics (LEN European Aquatics, 2022), this championship brings together athletes from various countries to compete in different types of swimming and diving events, vying for a total of 75 medals. It is an occasion where talents from across Europe come together to display their prowess in the water and claim victory for their nations. The most recent occurrence of this prestigious championship was observed in August 2022.

The primary focus of this study is to analyze the competition times of athletes across different swimming categories. The dataset utilized for this analysis consists of competition times in seconds, gathered from the women's 200m swimming semifinals for five distinct swimming styles. The main objective of this project is to compare the distribution of these competition times.

To achieve this goal, the study commences with a descriptive analysis, constructing histograms and boxplots to visualize the competition times in each category. Subsequently, hypothesis testing is employed to compare the average competition times both globally and in pairwise comparisons between categories.

Statistical methods used in this analysis include Global ANOVA, which compares the mean competition times across all categories, and two-sample t-tests, which compare means between each pair of categories. To address the challenges posed by simultaneous testing of multiple null hypotheses, correction procedures such as Bonferroni correction and Holm's step-down procedure are applied.

The results of this analysis reveal compelling insights. Notably, the categories of Freestyle and Backstroke exhibit the lowest and highest competition times, respectively, while the Butterfly category demonstrates the highest variance. The Global ANOVA results indicate that at least one category's mean time significantly differs from others, and the pairwise t-tests confirm that the mean times of the Butterfly and Backstroke categories differ significantly from each other.

The report is structured into several sections. Section 1 introduces the championship and its significance, while Section 2 presents the problem statement and details of the dataset. Section 3 elaborates on the statistical methods utilized in the analysis, providing their mathematical formulations. In Section 4, the statistical analysis is conducted, and the results are presented. Finally, Section 5 offers a concise summary of the study, recapping the key findings and contributions.

2 Problem statement

2.1 Data set and data quality

The dataset utilized for this analysis comprises competing times in various swimming categories of a 200 meters swimming contest. The data was obtained from the official results of the 2022 European Aquatic Championships' semifinals (LEN European Aquatics, 2022). It consists of 80 instances and 3 variables, with no missing values.

The three variables are as follows:

1. 'Name': This categorical variable represents the name of the participating athlete.
2. 'Time': This continuous variable shows the time taken in seconds by each athlete to complete the 200m race.
3. 'Category': This categorical variable represents the swimming style under which the athlete competed. The 'Category' variable includes five classes: Backstroke, breaststroke, butterfly, freestyle, and medley.

It has been observed that 7 athletes participated in two categories, with medley being the second category for most of them. However, this dual participation of athletes violates the assumption of mutual exclusivity of the sample, which is a precondition for the statistical tests employed in the analysis section of this study. To avoid this violation, such participants are excluded from the other category (non-medley) to preserve the information related to the category 'medley.' After removing these non-exclusive instances, the dataset now contains 73 observations.

This approach ensures that each category still has around 16 observations and prevents the deletion of 6 observations from a single category, which could have made it less informative and biased.

2.2 Objectives of the project

The objective of this study is to address two major research questions:

1. "Is there a time difference between all the categories?"

2. "Is there a pairwise time difference between the categories?"

These questions are investigated by comparing the distributions of the five categories. To achieve this, the mean competing times are collectively compared using a global ANOVA test, followed by pairwise comparisons of the time using a two-sample t-test. In both cases, the obtained p-values are examined to determine whether to reject or retain the null hypotheses. Descriptive analysis is employed as a preliminary tool to gain initial insights into the data.

3 Statistical Methods

The statistical methods used for this analysis are described in this section. For the implementation of these methods, assistance from the statistical software Rstudio, alongside the R programming language version 4.2.2, were used (R core team, 2022). For the visualizations and data manipulations, the libraries ggplot2 (Wickham, 2022) and dplyr(Wickham, 2022) were employed, respectively.

3.1 Statistical hypothesis testing

The process of verifying a certain assertion about the population parameter(s), based on the randomly collected sample data of size n, can be termed as hypothesis testing. If $f(x;\theta)$ represents the probability distribution of some population and $\theta \in \Theta, x$ (θ could also be a vector), then the hypothesis testing involves the verification of a specific statement about the unknown parameter θ (Dharmaraja Selvamuthu, 2018)

3.1.1 Components of statistical hypothesis testing

Null hypothesis and alternative hypothesis

Hypothesis testing presupposes two hypotheses, namely the null *(H_0)* and the alternative hypothesis *(H_A).* The statement to be tested about the unknown parameter θ is called null hypothesis and its complement statement is called the alternative hypothesis (Wasserman, 2013).

If the parameter space Θ is partitioned into two disjoint sets Θ_0 and Θ_1, then the null and alternate hypotheses can be mathematically represented as follows,

$$H_0 \colon \theta \in \Theta_0, \Theta_0 \subset \Theta,$$

$$H_A \colon \theta \in \Theta_1 = \Theta \setminus \Theta_{0.}$$

Test statistic and rejection region

The test statistic is a random variable, represented by 'X', whose value is determined based on the observed sample data. It is utilized to either provide evidence in favor of or contradict the null hypothesis. The specific formulation of the test statistic can vary depending on the type of hypothesis testing being employed (Dharmaraja Selvamuthu, 2018).

When testing the null hypothesis, researchers look for a subset of outcomes, denoted as 'R', which represents the rejection region. Let 'X' be the random variable with a range represented by '$\mathcal{X}$'. In this context, 'R ⊂ $\mathcal{X}$' denotes the rejection region. If the observed value of 'X' falls within the rejection region (X ∈ R), the null hypothesis is rejected in favor of an alternative hypothesis. On the other hand, if the observed value of 'X' does not fall within the rejection region, the test fails to provide sufficient evidence to reject the null hypothesis.

Typically, the rejection region is defined in the form:

$$R \ = \ \{x \colon T(x) \ > \ c\},$$

In this equation, 'T' represents the test statistic, and 'c’ is the critical value. The critical value sets the boundary for the rejection region (Wasserman, 2013). It is determined based on the desired significance level and the specific hypothesis test being conducted.

Types of errors

While testing hypotheses, two types of erroneous outcomes are possible. One possibility is that the H_0 is correct but it is rejected, this represents Type I error. Another possibility is that H_0 is incorrect but it is retained, leading to a Type II error. The other two cases where H_0 is correct and it is accepted, or gets rejected when it is incorrect, are unerring outcomes (Dharmaraja Selvamuthu, 2018).

Significance level

The significance level represents the probability of committing a Type-I error. It is denoted by α (Dharmaraja Selvamuthu, 2018).

P-value

Under the assumption of a correct H_0, the p-value for a test provides the probability of obtaining a result, that is at least as extreme as what is observed. H_0 is rejected if the p-value is less than α (Dharmaraja Selvamuthu, 2018).

3.2 One-way Analysis of variance (ANOVA)

One-way ANOVA is a statistical test used to compare the means of a single variable across multiple groups. The dependent variable is obtained by categorizing the independent categorical variables of the population, also known as factors or levels. To perform the ANOVA test, certain conditions must be met:

- The sample is obtained from the population through simple random sampling.
- The samples obtained are independent of each other.
- The variable being tested is normally distributed in each of the groups.
- The variable being tested has the same standard deviation for all the groups (Weiss, 2017).

Like any other hypothesis test, ANOVA presumes both null and alternative hypotheses:

$$H_0\text{: } Means\ of\ all\ the\ groups\ are\ equal. H_{\mathrm{a}}\text{: } Not\ all\ the\ means\ are\ equal.$$
$$H_{\mathrm{a}}\text{: } Not\ all\ the\ means\ are\ equal.$$

To perform ANOVA, first values for the following random variables must be calculated:

Group mean square (MSGR)

MSGR gives the variation among the group means. It is mathematically defined as,

$$MSGR = \frac{SSGR}{k-1}$$

Where, **SSGR** is the group sum of squares and it can be mathematically defined as,

$$\boldsymbol{SSGR} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2.$$

Here, k represents the number of groups under consideration, $\bar{x}_k$ represents the mean of k^{th} group and $\bar{x}$ represents the overall mean (Weiss, 2017).

Mean square error (MSE)

MSE gives the variation within the groups, mathematically it is defined as,

$$MSE = \frac{SSE}{n-k}$$

Where, **SSE** is sum of square errors and it can be computed using the following formula:

$$\boldsymbol{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2.$$

Here, n represents the total number of observations, n_k represents the number of observations for the k^{th} group, s_k represents the standard deviation for the k^{th} group (Weiss, 2017).

F-Statistic (*F*)

The ratio of variation among the group means to the variation within the groups is given by **F.** it represented as,

$$F = \frac{MSGR}{MSE}.$$

If the collected samples satisfy all the assumptions of ANOVA, then the *F*-ratio follows the Fisher–Snedecor's (UCLA, SOCR, 2022) distribution (Weiss, 2017).

Rejecting the null hypothesis

Rejecting the null hypothesis in a one-way ANOVA can be accomplished using two common approaches: the critical value approach and the p-value approach.

Critical Value Approach:

In the critical value approach, we first determine the critical value denoted as $\boldsymbol{F\alpha}$. This critical value is obtained by comparing the specified significance level (α) to the F-distribution table with $(k-1, n-k)$ degrees of freedom, where 'k' is the number of groups being compared, and $'n'$ is the total number of observations. The critical value $\boldsymbol{F\alpha}$ represents the cutoff point beyond which we reject the null hypothesis (H_0).

Next, we calculate the F-statistic for the data collected from the experiment. If the calculated value of F falls under the rejection region, i.e., if it is greater than F_α, then we reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_a).

P-Value Approach

In the p-value approach, we calculate the $p-value$ associated with the F-statistic. The $p-value$ represents the probability of obtaining an F-statistic as extreme as the one calculated, assuming the null hypothesis (H_0) is true. The p-value is obtained from the F-distribution table with $(k-1, n-k)$ degrees of freedom.

If the $p-valu$e is less than the specified significance level (α), we reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_a). A small $p-value$ indicates that the observed differences among the group means are unlikely to be due to random chance, providing evidence in support of the alternative hypothesis (Weiss, 2017).

3.3 Paired t-test

The paired t-test is a statistical test used to compare the means of two related groups. It is commonly used when the same participants are measured under two different conditions or at two different time points. This test helps to determine whether there is a significant difference between the two sets of measurements, and it is particularly useful for pre-post intervention studies or within-subject experimental designs.

The paired t-test involves two hypotheses:

1. Null Hypothesis (H_0): There is no significant difference between the two sets of measurements or conditions.

2. Alternative Hypothesis (H_a): There is a significant difference between the two sets of measurements or conditions.

In mathematical notation, the hypotheses can be expressed as follows:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

Test statistic:

The test statistic for the paired t-test is based on the differences between the paired observations and the standard error of these differences. The formula for the test statistic is:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Where:

- $\bar{d}$ is the sample mean of the paired differences.
- s_d is the sample standard deviation of the paired differences.
- n is the number of paired observations.

Rejecting the null hypothesis

To determine whether to reject the null hypothesis, we compare the calculated t-value to a critical value from the t-distribution with degrees of freedom equal to the sample size minus 1. The critical value corresponds to the desired level of significance (usually 0.05 or 0.01) and indicates the threshold beyond which we consider the result statistically significant.

If the calculated t-value exceeds the critical value, we reject the null hypothesis, suggesting that there is a significant difference between the two sets of measurements. On the other hand, if the calculated t-value does not exceed the critical value, we fail to reject the null hypothesis, indicating that there is no significant difference between the measurements.

3.4 Q-Q Plot (Quantile to Quantile Plot)

A normal probability Q-Q plot is used as a graphical method of comparing the distribution of the collected sample with the standard normal distribution. The theoretical quantiles of the normal distribution are plotted on the horizontal axis, whereas sample quantiles are plotted on the vertical Y-axis.

For constructing a Q-Q plot, the sample is sorted in ascending order and these quantiles are denoted by $\{y_{(1)}, y_{(2)}, \dots, y_{(n)}\}$. The probability points (p_i) are computed in the succeeding step, with application of the following order statistic:

$$p_i = \begin{cases} (i - 3/8)/(n + 1/4) & \text{if } n \leq 10, \\ (i - 1/2)/n & \text{if } n > 10. \end{cases}$$

Here, $i = 1,2, \dots, n$ represents the number of instances.

The theoretical quantiles (x_i), associated with respective $y_{(i)}$ are calculated using points p_i. The points $x_{i,}$ are obtained in such a way, that they satisfy the equality,

$P(X \leq x_i) = p_i$, where $X \sim N(0,1)$.

If all the plotted points closely follow a linear trend such that a reference line with slope=1 is plotted over the points, and a nearly perfect alignment of the line and points is observed, then the Q-Q plot suggests that the underlying variable is approximately following a normal distribution. The reference line with slope=1 represents the ideal alignment with the standard normal distribution. Deviations from this line indicate departures from normality. (Hay-Jahans, 2019).

3.5 Multiple testing problem and the Family-wise error rate

If multiple null hypotheses $(H_{01}, \dots, H_{0m})$ are tested simultaneously, the probability of making a false statistical inference can considerably increase (Shi-Yi Chen, 2017).

Family-wise error rate $(FWER)$ generalizes the probability of committing a Type-I error in a test setting of m null hypotheses. $FWER$ gives the probability of committing at least one Type-I error (Tibshirani, 2013).

Consider a test setting where m null hypotheses are being tested and V represents the number of Type-I errors, then $FWER$ is by,

$$\text{FWER } = \Pr(V \geq 1)$$

If the null hypotheses whose $p - values$ are less than their respective α levels are being rejected, then $FWER$ looks like,

$$\text{FWER}\ (\alpha) = 1 - \Pr\ (V = 0)$$

$$= 1 - \Pr\left(\bigcap_{j=1}^{m} \{\text{ do not falsely reject } H_{0j}\}\right)$$

Assuming that, the m tests are mutually exclusive, then the $FWER$ is given by,

$$\text{FWER}\ (\alpha) = 1 - \prod_{j=1}^{m} (1-\alpha) = 1 - (1-\alpha)^m.$$

3.6 The Bonferroni correction method

The Bonferroni method tries to control the $FWER$ at a level α. Consider a test setting where m null hypotheses are being tested. Let A_j denote an event, where Type-I error has occurred while testing the J^{th} null hypothesis. Then, the $FWER$ is given by,

$$\text{FWER}\ (\alpha)\ \leq \sum_{j=1}^{m} \Pr\left(A_j\right)$$

The Bonferroni correction method controls the $FWER$, by setting a threshold of rejecting the null hypothesis to α/m, such that $\Pr(A_j) \leq \alpha/m$. Which implies,

$$\text{FWER}\ (\alpha) \leq m \times \frac{\alpha}{m} = \alpha$$

All the null hypotheses whose $p - values$ are below the set thresholds, gets rejected. Hence, decreasing the chances of committing a Type-I error. This can also be viewed from another perspective, where a $p - value$ is multiplied by the number of tests m, which leads to the decrease in the number of false positives (Tibshirani, 2013).

3.7 Holm's Step-Down Procedure

Holm's method is a modification of the Bonferroni's correction method, it controls the $FWER$, by rejecting more null hypotheses than the previous method, thus decreasing the chances of committing a Type-II error. The operation of this method is demonstrated in the following sequential steps:

1. An α level is specified, at which the $FWER$ is to be controlled.
2. $p - values$ $(p_1, \dots, p_m)$ corresponding to the m null hypotheses are calculated and sorted in ascending order such that, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
3. Define, $L = min\left\{j: p_{(j)} > \frac{\alpha}{m+1-j}\right\}$.
4. All the null hypotheses H_{0j} for which $p_j < p_{(L)}$, are rejected.

Holm's method does not assume the mutual exclusivity of the m tests, hence the threshold used for rejecting a null hypothesis is calculated using all the $m, p - values$. Which also makes this method uniformly more powerful (refer) than the Bonferroni's method, and it will reject at least as many null hypotheses as the Bonferroni's (Tibshirani, 2013).

4 Statistical analyses

The results obtained after the application of the statistical methods introduced in the previous section are presented in this section.

4.1 Descriptive analysis

The dataset contains information about the time taken by athletes (in seconds) to complete a 200-meter swimming race. It comprises 80 observations and 3 variables: swimming style, name, and time. The swimming styles are classified into 5 categories: backstroke, breaststroke, butterfly, freestyle, and medley. Seven observations were omitted from the dataset as they represented athletes who participated in more than one swimming style. This step ensures that their dual presence does not violate the assumption of mutual exclusivity, which is necessary for certain statistical methods to be implemented in the subsequent part of this analysis.

The overall mean time taken to complete the race, irrespective of the swimming style, is 132.5 seconds, with a median of 132.6 seconds. Looking at individual swimming styles, the mean times are as follows: 119.35 seconds for freestyle, 131.38 seconds for backstroke, 131.84 seconds for butterfly, 133.94 seconds for medley, and 146.4 seconds for breaststroke. It is evident from these means that the freestyle and breaststroke categories have the lowest and highest completion times, respectively, and they are the categories that deviate the farthest from the global mean in the negative and positive directions, respectively.

Figure 1 displays histograms showing the frequency distribution of time for all the swimming styles. For most categories, the data appears to be clustered around the mean, except for the butterfly category. The interesting observation is that the freestyle category is represented by the lowest completion times and is slightly far from the overall mean. In contrast, for the other categories, the bar representing the lowest time has a frequency of 1, and their modes are near their respective means.

The analysis provides insights into the mean completion times for different swimming styles and their relation to the overall mean. However, it is important to note that further statistical tests and analyses may be necessary to draw more conclusive and precise inferences from the data.

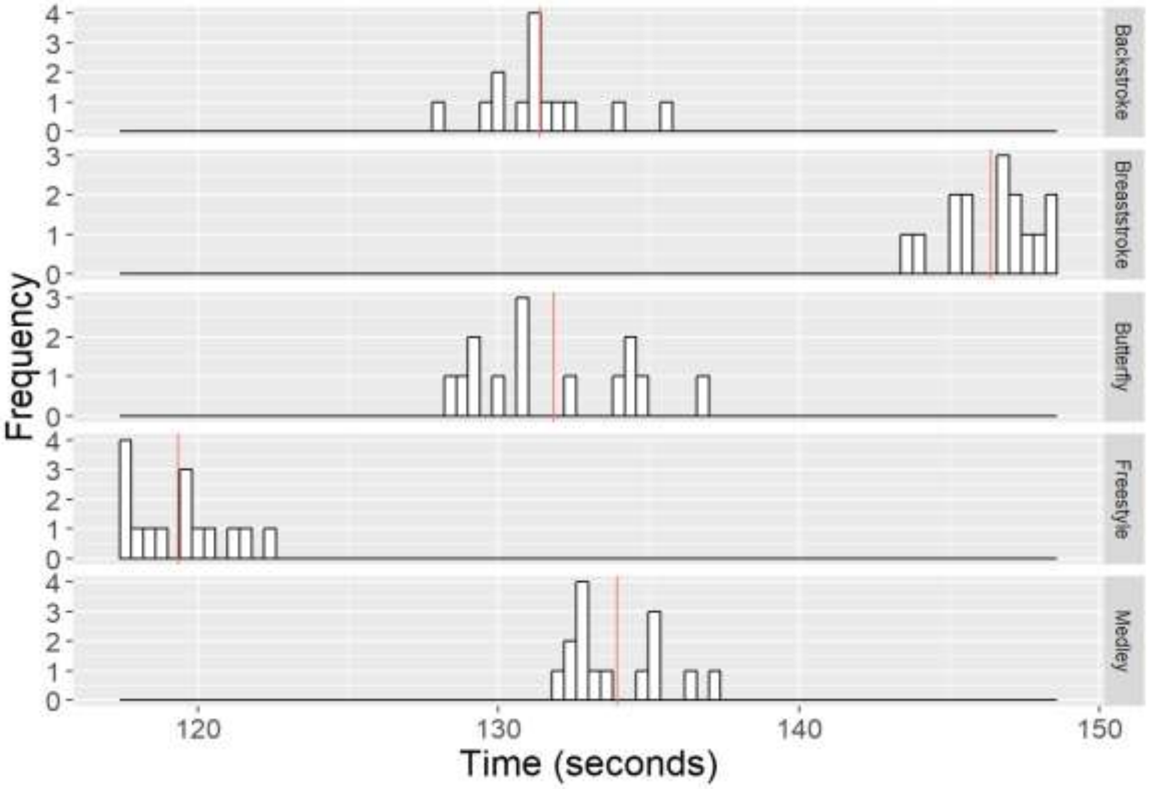
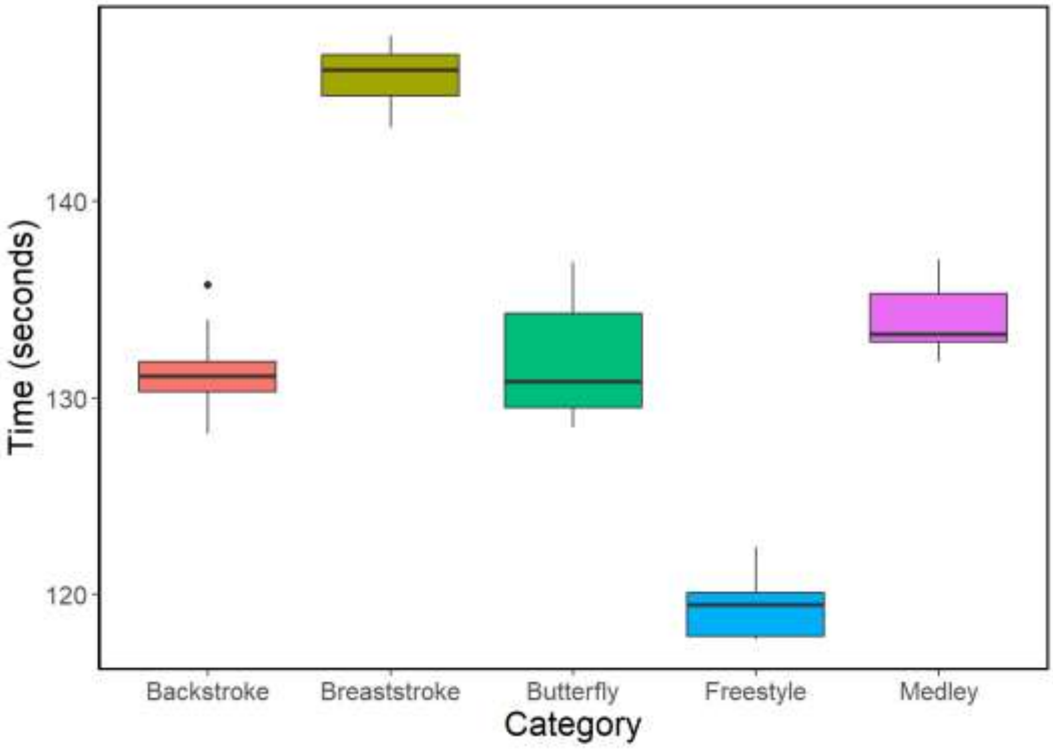


Figure 1: Frequency distribution for the time in various categories, red line represents mean.

The boxplots in Figure 2 illustrate that the completion times are generally homogenous within the categories. However, it is important to highlight that the butterfly category exhibits the highest spread, as evidenced by the interquartile range (IQR) of 4.8. This means that the completion times in the butterfly category are more dispersed compared to the other categories.

Additionally, the backstroke category includes an outlier, Federica Toma, whose completion time is 135.74 seconds. This time is 4.64 seconds more than the median completion time of the backstroke category. An outlier is a data point that significantly deviates from the rest of the data, and in this case, Federica Toma's completion time stands out from the typical completion times in the backstroke category.



4.2 Collective testing for the time difference between the categories

To test whether there exists a time difference between the five swimming styles, the sample is subjected to a global ANOVA test. The result from this test is considered reliable only if its presumptions are met. The first presumption, other than the random sampling, is that the variable subjected to the test must be normally distributed, and to check this precondition Q-Q plots have been made. From the plots in **Figure 3**, it is seen that, few categories show a slight deviation from the refence line, but these deviations are not strong enough to question their distribution's normality. Hence, it is apparent that all the categories approximately follow the normal distribution.

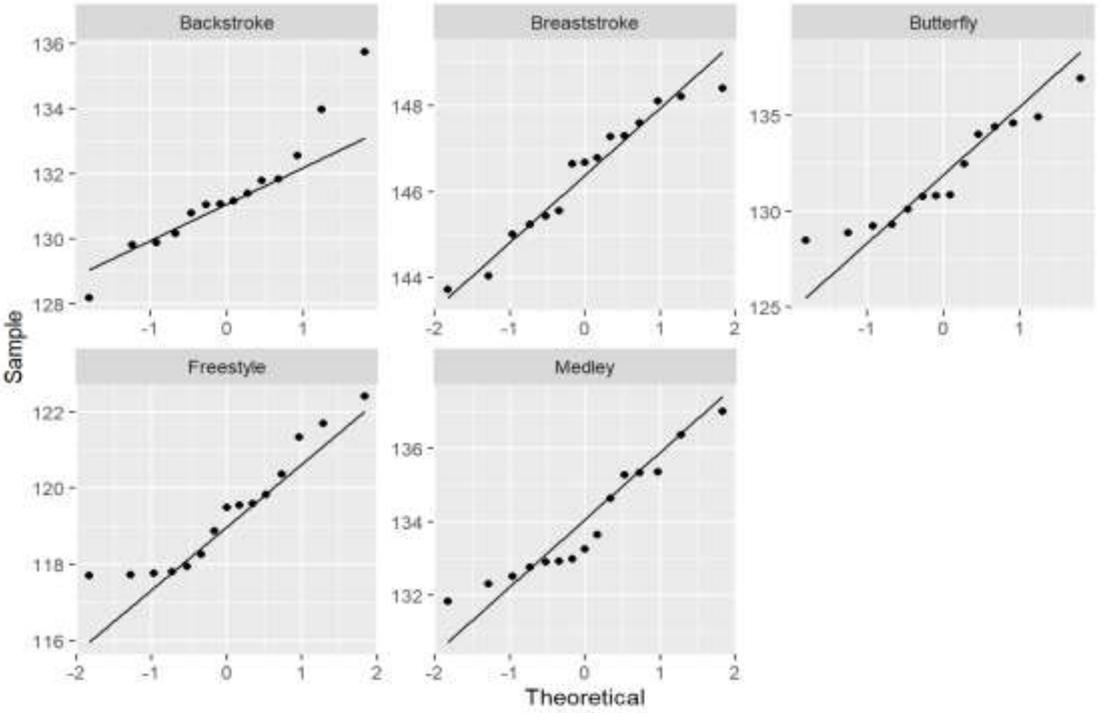


Figure 2:Normal Q-Q plots for all the categories.

Table 1 displays the standard deviation and variance of time in all five categories. It is observed that all categories, except for butterfly, have almost the same standard deviations. Although the butterfly category exhibits slightly more variability compared to the other categories, it is not significant enough to be considered heterogeneous. Thus, the condition of similar variances among the groups is met.

Initially, the condition of mutual exclusivity among the groups was not met due to a few athletes participating in two categories. To overcome this issue, those athletes were omitted from one category to ensure they become unique within the dataset. As a result, this condition is also satisfied now.

Table 1: Standard deviation and variance for all the categories.

Category	Standard deviation	Variance
Backstroke	1.85	3.42
Breaststroke	1.49	2.22
Butterfly	2.68	7.20

Freestyle	1.55	2.42
Medley	1.59	2.54

The hypotheses for this test are,

$$H_0: \mu_{\text{Backstroke}} = \mu_{\text{Freestyle}} = \mu_{\text{Butterfly}} = \mu_{\text{Breaststroke}} = \mu_{\text{Medley}}.$$

$$H_A: Not\ all\ the\ means\ are\ same.$$

Testing them through the global ANOVA at a significance level of $\alpha = 0.05$, produces results shown in **table 2**. The $p - value$ obtained through the test is exponentially smaller than the decided value of α. Hence the null hypothesis H_0, can be rejected in favor of the alternate hypothesis H_A.

Table 2: Global ANOVA summary.

	Degree of freedom	Sum of squares	Mean squares	$F\ Value$	$Pr(> F)$
Category	4	5543	1385.7	394.2	$\mathbf{< 2e - 16}$
Residuals	68	239	3.5		

Therefore, the test concludes that there exists at least one category whose mean is dissimilar from the others.

4.3 Testing for pairwise time difference between the categories

From the results obtained through ANOVA testing, it is evident that at least one category has a mean that is considerably dissimilar from the other categories. To further investigate this dissimilarity, but now from a pairwise perspective, the two-sample t-test has been implemented. The two-sample t-test, is also based on the similar presumptions as global ANOVA, and those conditions have already been checked during the previous test, hence it can be concluded that all the t-test's conditions remain satisfied.

4.3.1 The paired t-test without correction.

The sample being tested has 5 categories, and to investigate if there exists a pair-wise dissimilarity of mean between the categories, every category was compared with the other. The hypotheses for every pair tested are,

$$H_0 : \mu_1 - \mu_2 = 0.$$

$$H_A : \mu_1 - \mu_2 \neq 0.$$

Here, μ_1, μ_2 represents the means of the two paired categories subjected to a test. The test is carried out at a significance level $\alpha = 0.05$, produces the results tabulated in **table 3**. From the table it is apparent that, Butterfly-Backstroke is the only pair whose $p - value$ is greater than the presumed value of α. The $p - value$ for rest of the pairs is less than the presumed value of α.

Table 3: Results of the pairwise t-test

	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$< 2e - 16$	-	-	-
Butterfly	$\mathbf{0.51649}$	$< 2e - 16$	-	-
Freestyle	$< 2e - 16$	$< 2e - 16$	$< 2e - 16$	-
Medley	0.00046	$< 2e - 16$	0.00358	$< 2e - 16$

Hence, for all the pairs except for the pair Butterfly-Backstroke, the test rejects the null hypothesis. In the case of the Butterfly-Backstroke pair, the null hypothesis is not rejected in favour of the alternative hypothesis. This implies that there is no considerable difference between the mean times of the two categories in this pair.

4.3.2 The two-sample t-test with corrections.

Considering the multiple testing problem that arises with the simultaneous testing of multiple hypotheses, Bonferroni and Holm's correction procedures were applied. These procedures decrease the chances of making incorrect statistical inferences in their individual ways (see section 3.6).

With the Bonferroni correction

The Bonferroni method reduces the chances of committing a Type-I error by increasing the $p - values$, in relation to the number of tests performed. There are 10 null hypotheses that have been simultaneously tested in the previous section, for these many hypotheses, the calculated family-wise error rate is around 40% ($since, FWER = 1 - (1 - \alpha)^m$), which is considerably high. To account for this high $FWER$, the Bonferroni correction multiplies the $p - values$ for every test by m (number of tests). Test results obtained after the correction can be seen the **table 4**. The obtained $p - values$ for all the pair-wise tests are infinitesimally small, whereas for the Butterfly-Backstroke pair the $p - value$ is 1.0 which greater than the decided $\alpha - value$, hence the test fails to reject the null hypothesis for this pair, Seconding the results obtained from the preceding step.

Table 4: Results of the pairwise t-test with Bonferroni correction.

	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$< 2e - 16$	-	-	-
Butterfly	**1.00**	$< 2e - 16$	-	-
Freestyle	$< 2e - 16$	$< 2e - 16$	$< 2e - 16$	-
Medley	0.0046	$< 2e - 16$	0.0358	$< 2e - 16$

.

Testing after the application of this correction procedure also implies that the means of Butterfly and Backstroke categories are same.

With the Holm's correction

The Holm's step-down procedure is employed to control the family-wise error rate and reduce the likelihood of making incorrect statistical inferences, though its implementation is somewhat more intricate compared to the Bonferroni correction (refer to section 3.7). The results obtained after applying this correction procedure are presented in table 5. From the p-values apparent in the table, it is reasonable to reject the null hypotheses for all pairs except for the Butterfly-Backstroke pair. The p-values for all the other pairs are extremely small, significantly smaller than the α-value of 0.05. Consequently, we have sufficient evidence to reject the null hypotheses for those pairs. However, in the case of the Butterfly-Backstroke pair, the p-value exceeds the level of significance, leading to a failure to reject the null hypothesis for this specific comparison.

Table 5:Results of the pairwise t-test with Holm's correction.

	Backstroke	Breaststroke	Butterfly	Freestyle
Breaststroke	$< 2e - 16$	-	-	-
Butterfly	**0.5165**	$< 2e - 16$	-	-
Freestyle	$< 2e - 16$	$< 2e - 16$	$< 2e - 16$	-
Medley	0.0014	$< 2e - 16$	0.0072	$< 2e - 16$

This step also seconds the results obtained from the previous two t-tests, implying that means of the Butterfly-Backstroke pair are same.

5 Summary

The aim of this study was to compare the competition times of athletes in a 200-meter swimming contest across five different categories. The dataset includes three variables: swimming categories, athletes' names, and their competing times. Initially, the dataset contained 80 observations, but 7 were omitted due to a violation of the assumption of mutual exclusivity, which is required for conducting the global ANOVA and two paired t-tests.

Descriptive analysis was conducted as a preliminary step to gain a brief understanding of the data. The analysis revealed that athletes participating in the freestyle race had the fastest competition times, while the slowest performances were observed in the breaststroke category. The relatively slower performances in the breaststroke category were justifiable, considering the propulsion and resistance generated by the strokes. Additionally, the butterfly category exhibited the highest variance, but overall, all categories were approximately homogeneous within themselves.

Using the global ANOVA test, it was found that the mean times were not equal across all categories, indicating a significant difference in competition times. Subsequently, paired t-tests were performed for pairwise mean comparisons, and it was observed that the mean times of the Butterfly and Backstroke categories were the same. To address potential errors from multiple testing, the groups were re-tested using the Bonferroni correction and Holm’s step-down procedure in conjunction with the paired t-tests. Interestingly, the results from both corrections were consistent with the original t-test results, suggesting that the mean times of the paired categories Backstroke and Butterfly are indeed the same. In summary, the hypothesis testing indicates differences in mean competing times among various categories, and it also shows a similarity in means between the Butterfly and Backstroke categories.

To enhance the analysis, the inclusion of other informative factors as variables could provide more insights, and increasing the dataset size would likely lead to more reliable results.

References

Boslaugh, S. (2012). Statistics in a Nutshell, 2nd Edition (pp. 152-159). O'Reilly.

Dharmaraja Selvamuthu, D. D. (2018). Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control (p. 147). Springer.

Hay-Jahans, C. (2019). R Companion to Elementary Applied Statistics (p. 147). Chapman & Hall.

LEN European Aquatics. (2022, December 09). European Aquatics Championship. From https://www.roma2022.eu/en/

LEN European Aquatics. (2022, December 09). LEN European aquatics 2022 | Roma. From https://roma2022.microplustimingservices.com/indexRoma2022_web.php

R Core Team. (2022, December 09). R: The R project for statistical computing. From https://www.r-project.org/

Shi-Yi Chen, Z. F. (2017). A general introduction to adjustment for multiple comparisons. Journal of Thoracic Disease, Volume 9, no.6.

Tibshirani, J. W. (2013). An Introduction to Statistical Learning (pp. 562-566). Springer.

UCLA, SOCR. (2022, December 09). F-Distribution tables. From http://www.socr.ucla.edu/Applets.dir/F_Table.html

Wasserman, L. (2013). All of Statistics (p. 149). Springer.

Weiss, N. A. (2017). Introductory Statistics (p. 718). Pearson.

Wickham, H. (2022, December 09). A Grammar of Data Manipulation. From https://dplyr.tidyverse.org/

Wickham, H. (2022, December 09). Create elegant data visualization using the grammar of graphics. From https://ggplot2.tidyverse.org/