TU Dortmund

Introductory Case Studies

Project 2: Comparison of multiple distributions

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

Author: Raj Pawar

Group number: 17

Group members: Amritha Sukhdev Singh Agarwal, Sagar Basnet, Muhammad Fahad, Siddhartha Karki

December 8, 2022

Contents

1 Introduction

People, ethnicities, and genders come together via sports. The United Nations plans to campaign for the Sustainable Development Goals (Census, 2020) with this strong, straightforward language. The European Aquatics Championships Roma 2022 were planned to actively contribute to the accomplishment of some of the Sustainable Development Goals through the implementation of practical and tangible initiatives that can benefit the community and the environment. The championships consisted of 75 medal events in categories swimming, open water swimming, artistic swimming, diving, and high diving. (European Aquatics, 2022)

The goal of this report is to analyse the data-set which consists results of the women's 200 meters semifinals in the following 5 categories Breaststroke, Backstroke, Butterfly, Freestyle, and Medley. Firstly, we perform the descriptive analysis on the data and after that we perform a statistical test to check if the time differs between the swimming categories. In order to do this, we first conduct a global test over all five swimming categories, then a multiple two-sample test across all pairs of swimming style categories to determine pairwise differences. Additionally, we modify the results of the multiple two-sample tests to account for the issue with multiple testing before re-examining the pairwise differences between swimming categories.

In addition to this introduction, this report comprises of four other sections. The data-set, data quality, and variable properties are all thoroughly explained in Section 2. Additionally, the project's goals are briefly explained. The statistical techniques used in this project are described in Section 3. We conduct a global test after that the idea of statistical hypothesis testing is introduced. We discuss over how to conduct multiple two-sample tests over all pairs of swimming categories. Additionally, the issue of multiple testing is covered, along with a solution strategy. After using these statistical techniques on the provided data-set, interpretations and conclusions are the key points of Section 4. The resulting graphs and tables are also presented. Section 5 ends with a summary of the findings and the most important information from the provided data set.

2 Problem statement

2.1 Source and Quality of Data

The popular European Aquatics Championships 2022, which take place in Rome, provided the data for this project (Federazione Italiana Nuoto, 2022). The data-set "SwimmingTimes.csv" consists of swimming timings for 80 women swimmers from 5 swimming categories collected in year 2022. There are 3 variables available in this data-set named Category, Name, and Time. Category is a categorical variable which describes the 5 swimming styles named Backstroke, Breaststroke, Butterfly, Freestyle, and Medley. Name variables consists of full names of women's who participated in the European Aquatics Championship. Time variable describes that how much time in seconds does each participant took to complete the 200 meter swimming. We don't have any null values present so overall quality of the data is good. So, we can use all the observations for our analysis.

2.2 Objectives

The objective of this report is to conduct a descriptive analysis and statistical tests on the data-set. After that we perform a global test with ANOVA method. Pairwise comparison is conducted using a t-test to check whether pairs of swimming categories shows a difference in timings. The p-values of the t-tests are further corrected using the Bonferroni and Holm-Bonferroni correction procedure to eliminate inaccuracies brought on by multiple testing.

3 Statistical methods

In this section, we explain the various statistical methods and tests used for the analysis of the given data. The Python programming language (Python Software Foundation, 2020), version 3.9.0 was used for analysis and compilation. All the used packages are described in the code file.

3.1 Hypothesis Testing

An statement or supposition on the population characteristics of one or more random variables is known as a statistical hypothesis. The capital letter H stands for the statistical hypothesis. We normally evaluate two types of hypotheses when testing an assumption. The null hypothesis H_0 is the hypothesis which need to be tested and other one is alternate hypothesis H_1. We can reject the null hypothesis if the sample contains sufficient data to reject the statement that there is no effect in the population. Otherwise, we fail to reject the null hypothesis. The complement of the null hypothesis is the alternative hypothesis. Both are mutually exclusive, meaning that only one can be true at a time.

3.1.1 Errors in statistical testing - Type I and Type II

When conducting statistical testing, errors can occur. This is possible because the sample chosen for the test might not accurately reflect the population, and therefore the conclusion drawn will not accurately reflect the population as a whole. This could result in inaccurate results.

When an actual true null hypothesis is rejected, this is Type I error and it's also called a false positive. The chosen significance level (α) provides the probability of making a Type I error. When a null hypothesis that is actually false is accepted, this type of error is called a Type II error, which is also known as a false negative. Type II error probability is represented by β. Type I error can be reduced by lowering the value of α and Type II error can be reduced by increasing the sample size. (Mood et al., 1973, pg.402-405)

3.1.2 Test statistic

A random variable is one whose possible values are the numerical results of an uncertain event. A random variable that can be used to disprove the null hypothesis is a test statistic. To conduct statistical tests, the test statistic often makes use of the central tendency and variance of the observed data. (Taboga, Marco, 2021)

3.1.3 p-value

The p-value helps us to decide weather to reject or fail to reject the null hypothesis H_0. It demonstrates how the chosen sample data disprove the null hypothesis and demonstrates the statistical significance of the measured values. There is a much higher chance to reject the null hypothesis and adopt the alternative hypothesis if the p-value is lower than the decided significance level (α). The test statistic and degrees of freedom, which is the difference between the number of observations and the number of independent variables, are used to calculate the p-values. (Heumann et al., 2016, pg.215)

3.1.4 Significance level

The significance level, also represented by α, is a predetermined value for the statistical test. The significance level is the probability of rejecting the null hypothesis, given that the null hypothesis was assumed to be true. The significance level is compared to the p-value to decide if the null hypothesis should be rejected. If the p-value is less than the significance level, the null hypothesis can be rejected and the alternative hypothesis can be regarded as statistically significant. The most typical significance threshold is 0.05 or 5%, though it can be adjusted even lower depending on how significant the test is. To be considered statistically significant under the null hypothesis, the results must have a 5% or lower chance of occurring, according to a 0.05 significance level. (Heumann et al., 2016, pg.213)

3.2 Analysis of variance (ANOVA)

ANOVA stands for "Analysis Of Variance", it is a parametric test technique that allows us to evaluate the parameter hypothesis without raising the Type I error. The null hypothesis H_0 in an ANOVA allows us to determine if the means of all individual groups are equal to one another, while the alternative hypothesis H_1 assumes that at least one group has a different mean from the other groups. Mean is the parameter utilized in this hypothesis test.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = ... = \mu_k$$

$$H_1 : \text{At least one of the group has different mean}$$

where, μ_k represent the mean of k^{th} group and k is the number of group.

There are 3 assumptions of ANOVA. First assumption is the samples taken from the population are independent within and between the groups. Second is the samples are obtained from the normally distributed population. And third one is distribution of samples have equal variance between the different groups. To test the null hypothesis, ANOVA uses the F-statistics. The variation between the group mean and the variance within each group is what we refer to as the F-statistics. It is described as:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M_G)^2}{K-1}}{\sum_{i=1}^{k} \frac{\sum_{i=1}^{n_j} (x_{ij} - M_j)^2}{n_j - 1}}$$

where, k is the number of groups, n_j is the size of j^{th} group, M_j is the mean of j^{th} group, M_G is the total mean of all observation together, and x_{ij} is the i^{th} observation in group j. The p-value for the test statistic is provided by the ANOVA. If there is a statistically significant difference between the means of two or more independent groups, the p-value can be used to determine this. (M.H. Herzog and Clarke., 2019, pg.69) (Black, 2009, pg.407)

3.3 Multiple testing problem

If we need to examine data from more than two groups, we must run additional two sample tests. Every group combination needs to be tested. The issue here is that as the number of comparisons increases, the probability of a Type I error rises as well. This means, the probability of rejecting null hypothesis and opting for alternative hypothesis is increasing with each comparison, even if the null hypothesis is true.

If the null hypothesis is true, the significance level, for example $\alpha = 0.05$, limits the chance of Type I error in a two-sample test. That indicates that in circumstances when $1 - \alpha = 0.95$, we do not cause Type I error when the null hypothesis is true. So, the probability of not making any Type I error for n comparison is $(1 - \alpha)^n$. Now, the probability of making atleast one Type I error for n independent comparison is $1 - (1 - \alpha)^n$. We can infer from these equations that as the number of comparisons rises, so does the chance of Type I error. (M.H. Herzog and Clarke., 2019, pg.63). There are multiple methods to fix the error generated while multiple testing. Two of them are discussed below.

3.3.1 Bonferroni correction

The possibility of committing a Type I error rises when several tests are performed on the same dependent variable, increasing the probability that a significant result would occur by accident. The Bonferroni correction method is used to correct the Type I error. Bonferroni correction method is also called as Bonferroni type adjustment method. The FEWR (Family-Wise Error Rate) is the chance of making one or more Type I errors while conducting multiple hypothesis tests. FWER can be calculated as,

$$\alpha_{FW} = 1 - (1 - \alpha)^k,$$

Here, α denotes the significance level, k denotes the number of hypotheses on the same dependent variable and FWER is denoted by α_{FW}

To reduce Type I error, we might test each hypothesis at a different threshold of significance. That is α/k, where the number of hypotheses is denoted by k. To reduce the false positives, the α levels is adjusted by Bonferroni correction. The adjusted α level formula is given by,

$$\alpha_{adjusted} = \frac{\alpha}{k}$$

Each p-value is multiplied by k before being compared to the significance level. If the multiplied value is greater than 1, then we use an adjusted p-value of 1. (Hay Jahans, 2019, pg.274)

3.3.2 Holm-Bonferroni correction

Holm's procedure is a sequential approach whose goal is to increase the power of the statistical tests while keeping FWER Type I error under control. The first step in Holm's procedure is to perform the tests to obtain their p-values, then we order the tests with smallest p-value to the one with largest p-value. The test with the smallest probability will be tested with a Bonferroni for the FWER tests. If the test is not significant, then the procedure stops. If the first test is significant, then the test with the second smallest p-value is corrected with a Bonferroni for a (FWER) - 1 tests. The procedure stops when the first non-significant test is obtained or when all the tests have been performed. Formally, we assume that the tests are ordered according to their p-values.

When using the Bonferroni correction with Holm's approach, the corrected p-value for the i^{th}-test, denoted $p_{Bonferroni}, i|C$ is computed as:

$$p_{Bonferroni}, i|C = (C - i + 1) * p$$

Just like the standard Bonferroni procedure, corrected p-values larger than 1 are set equal to 1 (Abdi, 2010)

3.4 Quantile–Quantile Plots (Q-Q Plots)

The Q-Q plots are a graphical tool for determining if a sample's distributional characteristics are likely to have originated from a hypothetical distribution like the Normal distribution. Take into account these sample data: $y_1, y_2, ..., y_n$ and sort them in ascending order. After sorting, it is denoted by $y_{(1)}, y_{(2)}, ..., y_{(n)}$ and is referred to as the sample or observed quantiles. The probability points p_i are determined using the following formula for $i = 1, 2, .., n$.

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

The theoretical quantiles, x_i, and their corresponding sorted sample quantiles, y_i, are calculated using probability points p_i. The theoretical quantile for $i = 1, 2, ..., n$ should be defined as x_i such that $P(X \leq x_i) = p_i$, where $X \sim N(0, 1)$. The graph displays the ordered pairs $(x_i, y_{(i)})$, and it also allows for the overlay of a reference line. In a normal probability Q-Q plot, the sample mean is the y-intercept, and the standard deviation is the slope of the reference line. For non-standardized data, the reference line has the formula $y = \mu + \sigma z$, where z is the transformation calculated using $(y - \bar{y})/s$. The mean and standard deviation of the proposed theoretical normal distribution are denoted by μ and σ, respectively. If the two distributions under comparison are comparable, then the points in the Q-Q plots will almost certainly lie on the line $y = x$.

Once the plot is finished and the reference line is superimposed on it, it is simple to visually check whether the data points match the theoretical distribution. If the data points follow the reference line, they are normally distributed. Any deviation from the reference line deviates from normality and causes a concave trend to form. Data

skewness is indicated by the concave trend in the Q-Q plots. The extreme points of the distribution are known as outliers. (Hay Jahans, 2019, pg.147-152)

4 Statistical analysis

Statistical hypothesis testing on a given data-set will be carried out in this section using the statistical techniques covered in Section 3 and their interpretation. Before we conduct the hypothesis tests we have removed the duplicate entries for groups which had same swimmers names so there that there is no issue while assuming the assumptions.

4.1 Descriptive Analysis

Before performing hypothesis testing, we perform descriptive analysis in this subsection to acquire a general understanding of the data. Table 1 presents Frequency , Mean, Standard Deviation, and the Interquartile Range (IQR) of the women swimmers swimming style namely, Backstroke, Breaststroke, Butterfly, Freestyle, and Medley. The frequency column shows that there are different number of women participants in each category, with minimum of 14 samples in categories Backstroke, Breaststroke and Butterfly to the maximum of 16 samples in category Medley. In total including all swimming categories we have 73 samples to analyse.

Table 1: Table showing Frequency, Mean, Standard deviation and Interquartile range for swimming style categories

Category	Frequency	Mean	Standard Deviation	Inter Quartile Range
Backstroke	14.0	131.38	1.85	1.52
Breaststroke	14.0	146.31	1.51	2.00
Butterfly	14.0	131.66	2.61	4.42
Freestyle	15.0	119.36	1.56	2.23
Medley	16.0	134.04	1.59	2.47

The mean value varies from 131.38 to 146.31 and the standard deviation varies from 1.51 to 2.61. We can observe the standard deviation for Backstroke, Breaststroke, Freestyle, and Medley to be almost similar but the category Butterfly shows a value of 2.61 which is a little bit higher. For the interquartile range Butterfly category shows value of 4.42 which is almost double when compared to remaining 4 categories.

4.2 Global test

To determine whether we can utilize a ANOVA test, we shall first examine it's underlying assumptions. We can infer that the variance is not uniformly distributed among the three groups by examining the box plot's in figure 1 of the appendix. As seen in figures 2, 3, and 4 in the appendix, we generated QQ-plots to see if the samples for swimming styles are normally distributed. The sample quantile of swimming time in seconds is plotted on the y-axis, and the theoretical quantile of a normal distribution is plotted on the x-axis. As discussed in section 3.4 we can clearly see that the points follow the reference line and if the data points follow the reference line, they are normally distributed. The individual observations are different within and between the group as the duplicate observations between the group are eliminated. Hence, all the 3 assumptions of ANOVA are satisfied.

The null hypothesis H_0 defined for this test is that there is no difference in the mean value of observations for different swimming categories. The alternate hypothesis H_1 is that atleast one mean value of observations is different for the swimming categories.

Table 2 contains the results of the ANOVA test. The degree of freedom for variability between group is calculated by subtracting the number of groups which are the 5 swimming categories by 1, so $5 - 1 = 4$. The degree of freedom for the within-group (Residual) variability is the difference between the total number of observations and the number of groups, so 73 - 5 = 68. The F-value is high and takes a value of 385.92 because between-group variability is greater than within-group variability. The mean squared value for within-group variability is 3.45, while the mean squared value for between-group variability is 1331.81.

Table 2: Table showing ANOVA test results

	Sum of Squares	Mean Squared	Degree of freedom	F-value	p-value
Category	5327.24	1331.81	4.0	385.92	0.0
Residual	234.67	3.45	68.0		

Given that the F-value is large, there is probably a variation in the means of the swimming times between the 5 swimming categories. The p-value, which is less than the chosen significance value α of 0.05, is considerably very low so its rounded to 0.0. The p-value is lower than significance value α, indicating that there are significant differences between the timings in various swimming categories, hence the null hypothesis can be rejected.

4.3 Multiple two sample test

In subsection 4.2 we conducted global test ANOVA and we found that the null hypothesis is been rejected and there is difference between the means values of time for the swimming categories. We will examine the pairwise variations in swimming timings across all swimming categories provided in the dataset. On each paired combination of swimming categories, we will do a two sample t-test with significance level α of 0.05 along with test results adjustments with Bonferroni adjustments and Holm-Bonferroni adjustment methods. Here, we suppose the following null and alternate hypothesis,

Null Hypothesis H_0: There is no difference in the distribution of the swimming time between two swimming categories.

Alternate Hypothesis H_1: Difference exists in the distribution of the swimming time between two swimming categories.

Table 3 provides the results of the paired t-test with the Bonferroni correction. The pairs of swimming categories that are tested are listed in the table's first two columns. The p-values are displayed in the third column. The adjusted p-values using Bonferroni corrections are displayed in the fourth column. We are able to see that the pair Backstroke and Butterfly has the p-value 0.7492 and the adjusted p-value 1.0 which is greater than the significance value α 0.05.

Table 3: Table showing results of Multiple t-test with Bonferroni Corrections

Group 1	Group 2	p-value	Adjusted p-value	Reject
Backstroke	Breaststroke	0.0	0.0	True
Backstroke	Butterfly	0.7492	1.0	False
Backstroke	Freestyle	0.0	0.0	True
Backstroke	Medley	0.0002	0.0022	True
Breaststroke	Butterfly	0.0	0.0	True
Breaststroke	Freestyle	0.0	0.0	True
Breaststroke	Medley	0.0	0.0	True
Butterfly	Freestyle	0.0	0.0	True
Butterfly	Medley	0.0048	0.0479	True
Butterfly	Medley	0.0	0.0	True

Here, we fail to reject the null hypothesis for categories Backstroke and Butterfly. But for the remaining pairs of categories we successfully reject the null hypothesis as the p-

values and the adjusted p-values are smaller than significance value α 0.05. We are able to observe that there is no considerable change in the adjusted p-values so the results remain the same after the error adjustments.

Table 4 shows the paired t-test results with Holm-Bonferroni correction method. We can see that the adjusted p-value for the category pair Backstroke and Butterfly has changed from 1.0 to 0.7492. We have our significance value α as 0.05 and the changed p-value is still greater than the significance value hence, we still fail to reject the null hypothesis. We can observe that the number of pairs that reject the null hypothesis are the same after Holm-Bonferroni correction procedure.

Table 4: Table showing results of Multiple t-test with Holm-Bonferroni Corrections

Group 1	Group 2	p-value	Adjusted p-value	Reject
Backstroke	Breaststroke	0.0	0.0	True
Backstroke	Butterfly	0.7492	0.7492	False
Backstroke	Freestyle	0.0	0.0	True
Backstroke	Medley	0.0002	0.0007	True
Breaststroke	Butterfly	0.0	0.0	True
Breaststroke	Freestyle	0.0	0.0	True
Breaststroke	Medley	0.0	0.0	True
Butterfly	Freestyle	0.0	0.0	True
Butterfly	Medley	0.0048	0.0096	True
Butterfly	Medley	0.0	0.0	True

Even after making the correction in p-values, we may still draw the conclusion that there are statistically significant pairwise differences between the swimming times for various swimming categories at a significance level of 0.05.

5 Summary

The given data-set has information about 80 swimmers timings for 5 different swimming categories with 3 variables Name, Category, and Time. There were no missing value in the data-set, so we can say that the overall quality of the data-set was good. The goal of the analysis was to run a global test to determine whether there were any variations in times between the five swimming categories and then use two-sample tests to further

examine pairwise differences. Finally, a correction method was used so that we could account for various testing problems which occurs while repeatedly multiple testing.

We then performed the descriptive analysis on the given data set and checked the variances, means and standard deviations. We found out that the means were not deviating considerably. Swimming category Butterfly showed a slight deviation having standard deviation of 2.61. There were total of 73 observations after we removed the duplicated observations which were repeating in two groups while testing. We removed these entries to prevent the assumptions of ANOVA getting violated. We then checked the normality of the data by plotting the QQ plots and Box plot for every swimming category and found out that there the data is distributed normally. Then, we performed the ANOVA test and checked the null and alternate hypothesis. We supposed the null hypothesis as means of swimming times across all the swimming categories are same. And we considered our alternate hypothesis as the compliment of it, that is atleast one of the swimming category has different mean.

The performed tests gave a result that there is statistically significant pairwise differences between the times in all swimming categories for all comparison. Conducting these multiple tests produced some errors, so to cope with these errors we adjusted the p-value using the Bonferroni correction for the multiple testing problem. After adjusting the p-values we observed that the p-value had increased for the comparison group Backstroke and Butterfly but it was still greater than the significance value of 0.05. The remaining group comparison p-values did not changed considerably and were less than the significance value of 0.05. The same multiple tests has been performed with the Holm-Bonferroni Correction procedure but we get the same results and we were able to reject the null hypothesis expect for the pair Backstroke and Butterfly. Therefore, we can say that there are significant pairwise differences between the times in different swimming categories. We can further use other adjustment methods for multiple testing and analyze our results.

Bibliography

Herv´e Abdi. Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8, 2010.

Ken Black. *Business Statistics: For Contemporary Decision Making, Sixth Edition*. John Wiley & Sons, Inc., 2009. ISBN 9780470409015.

Census. Sustainable development goals and the 2020 round of censuses. https://www.census.gov/content/dam/Census/library/working-papers/2018/demo/sdg-2020 2020. Accessed: 2022-12-06.

European Aquatics. European aquatics championships roma 2022. https://www.roma2022.eu/en/, 2022. Accessed: 2022-12-06.

Federazione Italiana Nuoto. Federazione italiana nuoto. https://roma2022.microplustimingservices.com/indexRoma2022$_w$*eb.php*, 2022. *Accessed* : $2022 - 12 - 06$.

Christopher Hay Jahans. *R companion to elementary applied statistics*. Chapman and Hall/CRC, Boca Raton, 2019.

Christian Heumann, Michael Schomaker, and Shalabh Shalabh. *Introduction to Statistics and Data Analysis*. 01 2016. ISBN 978-3-319-46160-1. doi: 10.1007/978-3-319-46162-5.

G. Francis M.H. Herzog and A. Clarke. *Understanding Statistics and Experimental Design: How to Not Lie with Statistics. Learning Materials in Biosciences., Sixth Edition*. Springer International Publishing, 2019. ISBN 9783030034993.

A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduction to the Theory of Statistics*. International Student edition. McGraw-Hill, 1973. ISBN 9780070428645. URL https://books.google.de/books?id=Viu2AAAAIAAJ.

Python Software Foundation. Python, 2020. URL https://www.python.org/.

Taboga, Marco. Test statistic. https://www.statlect.com/glossary/test-statistic., 2021. Accessed: 2022-12-08.

Appendix

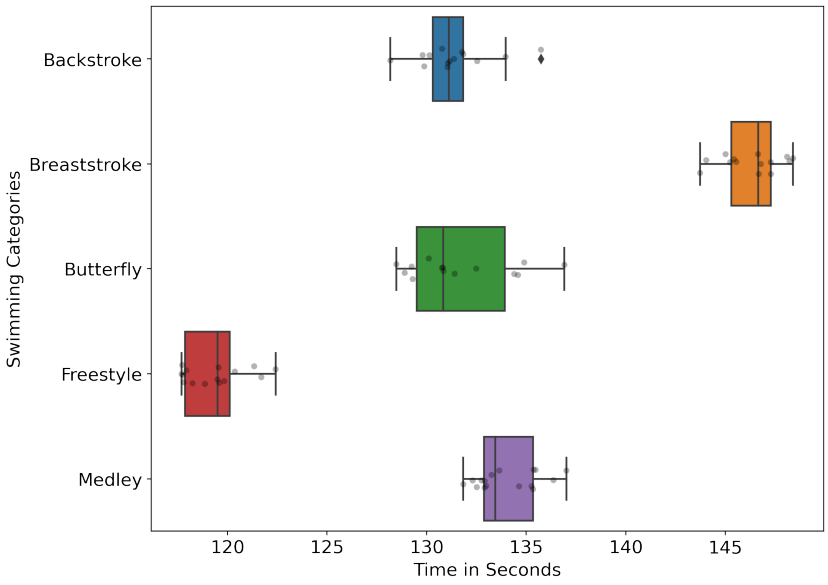
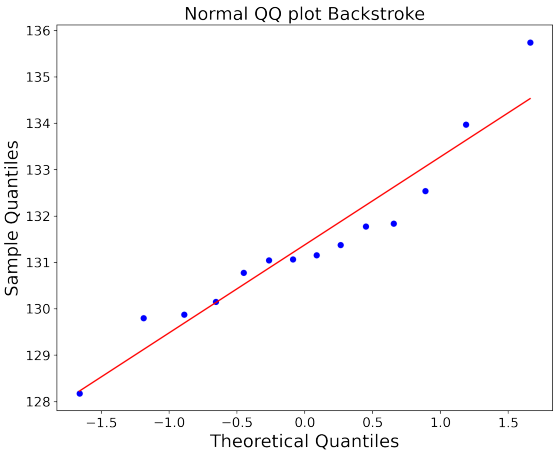
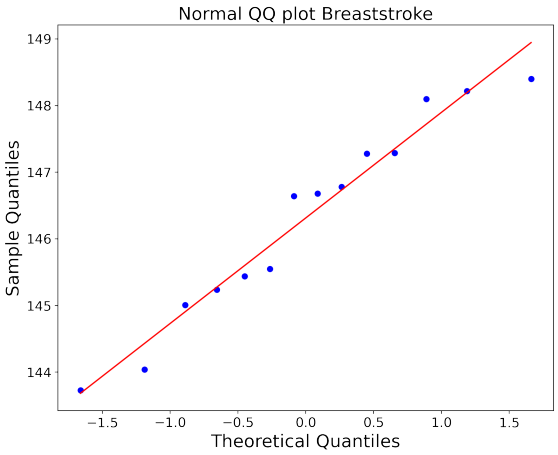


Figure 1: A Box plot showing the distribution of time in seconds for each swimming category

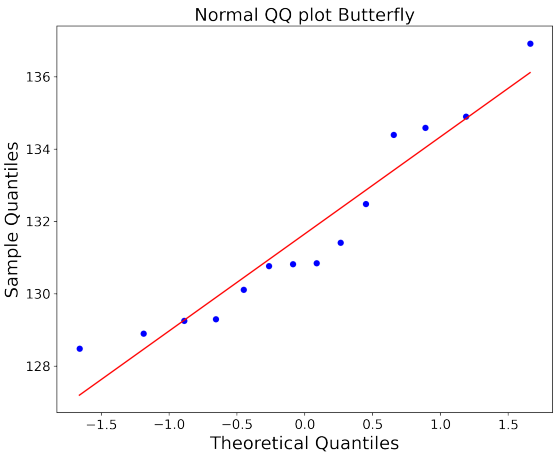


(a) QQ-Plot for swimming style backstroke to check the normal distribution

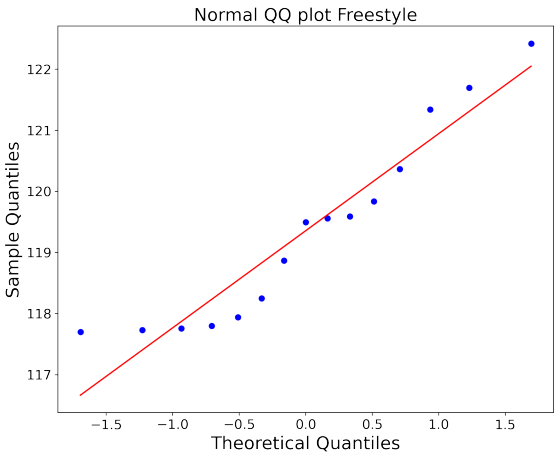


(b) QQ-Plot for swimming style breaststroke to check the normal distribution

Figure 2: QQ-Plot for swimming style backstroke and breaststroke



(a) QQ-Plot for swimming style butterfly to check the normal distribution



(b) QQ-Plot for swimming style freestyle to check the normal distribution

Figure 3: QQ-Plot for swimming style butterfly and freestyle

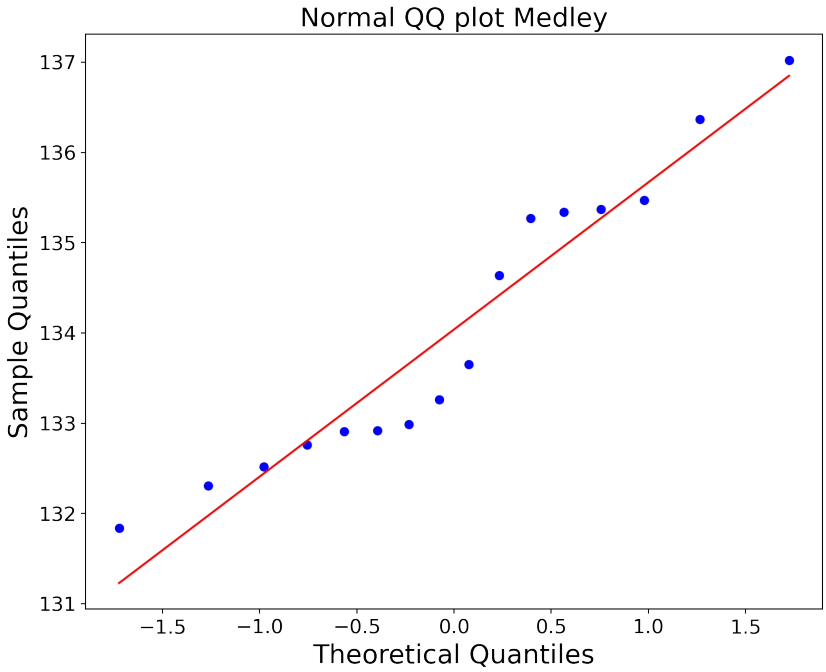


Figure 4: QQ-Plot for swimming style medley to check the normal distribution