TU Dortmund

Introductory Case Studies

Project 1: Assessing the Reliability of Statistical Tests in Group Comparison Analysis

Lecturers:

Dr. Crystal Wiedner

Dr. Marlies Hafer

Dr. Rouven Michels

Author: Mohaiminul Islam

Group number: 5

Group members: Md Zihad Hossain, , Mosfiqun Nahid Hassan, Ilham Pambudi

November 24, 2024

Contents

1 Introduction

The Berlin Marathon, held annually in Berlin, Germany, since 1974, is one of the most prestigious and well-known long-distance races, attracting thousands of runners of all genders from around the world. This project focuses exclusively on six age groups of female participants to examine the differences in their average finish times. To address this objective, statistical methods such as one-way ANOVA and two-sample t-tests are employed. The underlying assumptions for these methods are evaluated using appropriate statistical tests, including the Shapiro-Wilk test and Levene's test, as well as visualization technique like Q-Q plot. Following the initial analysis, post-hoc methods, namely Tukey HSD and Bonferroni correction, are applied to further investigate specific group differences with stricter adjustments to control the overall family-wise error rate.

The report is structured as follows: Section 2 provides an overview of the dataset and outlines the project's objectives. Section 3 details the statistical methods and their assumptions. Section 4 applies these techniques to the dataset and interprets the results, while Section 5 summarizes the findings, discusses implications, and offers suggestions for future research.

2 Problem statement

2.1 Data Description and Data Quality

The dataset analyzed in this report was provided by the instructors of the "Case Studies I" course at TU Dortmund University during the winter term 2023/24. It represents a reduced version of the Berlin Marathon data, containing two integer-type variables: `agegroup` and `time`.

`agegroup` is a categorical variable with six unique values, each representing a distinct age group of participants. `time` is a continuous variable indicating the finish times of runners, measured in seconds. The dataset contains one thousand observations with no missing values; however, some extreme values, particularly in the younger age groups, are present and remain unadjusted in this project.

2.2 Project Goals

The primary objective of this report is to investigate differences in average finish times across six distinct age groups of female marathon participants using appropriate statistical methods. The analysis aims to examine variation within each group, explore trends in finish times relative to age, and assess whether significant differences exists between the groups. Additionally, the report evaluates the normality of the data distributions using tests such as the Shapiro-Wilk test. The effectiveness of various statistical methods, including ANOVA, t-tests, and post-hoc tests, is also compared to identify these differences.

3 Statistical Methods

3.1 Hypothesis Testing

Hypothesis testing is a statistical method used to evaluate claims about a population based on sample data. The process begins with two contradictory hypotheses: the null hypothesis (H_0), which assumes no effect or no difference, and the alternative hypothesis (H_a), which represents a claim contradictory to H_0. Sample data is analyzed using appropriate statistical techniques to determine whether sufficient evidence exists to reject H_0. The outcome of the test leads to one of two decisions: either reject H_0 if the evidence supports H_a, or fail to reject H_0 if the evidence is insufficient. This systematic approach ensures conclusions are data-driven and statistically sound while accounting for potential errors.(Ott, 2016, p. 505-506)

3.1.1 Type I and Type II Errors

In hypothesis testing, two potential errors can occur: Type I error and Type II error. A Type I error happens when the null hypothesis H_0 is rejected, even though it is true, resulting in a false positive. The probability of making a Type I error is denoted by the significance level (α). Alpha (α) is determined before conducting the test (typically set at 0.05) and serves as a threshold for deciding when to reject the null hypothesis. In contrast, a Type II error occurs when the null hypothesis is not rejected, even though it is false, leading to a false negative. The probability of a Type II error is denoted by β, with its complement $(1-\beta)$ representing the power of the test, or the ability to detect a

true effect. Balancing these errors is crucial, as lowering α to reduce the risk of a Type I error can increase the likelihood of a Type II error, and the other way around.(Ott, 2016, p. 508-509)

3.1.2 Test Statistic

The test statistic is a random variable calculated from the sample data, which is used to evaluate the null hypothesis (H_0) within the framework of the chosen statistical test. The type of test statistic depends on the hypothesis being tested and the nature of the data. Each test statistic is generally associated with a specific probability distribution. For example, when we have a sample with a large number of observations and a known population standard deviation (σ), the z-statistic is used to test whether the sample mean significantly differs from the population mean, i.e.,

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0$$

$$\text{Test Statistic } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

(Ott, 2016, p. 511)

3.1.3 P-Value

The P-value represents the probability of obtaining test statistics as extreme, or more extreme, than the one observed in the sample data, assuming that the null hypothesis is true. After computing the test statistic, the p-value is obtained by determining the area under the associated distribution curve corresponding to the extreme values of the test statistic.For example, the z-statistic follows the standard normal distribution $N \sim \mathcal{N}(0, 1)$. The p-value for 1 can be calculated as:

$$p - \text{value} = 2 \times P(z \geq |z_{\text{obs}}|)$$

(Ott, 2016, p. 512-513)

3.1.4 Decision and Conclusion

The decision to reject or fail to reject the null hypothesis (H_0) is based on comparing the p-value with the significance level (α). If the p-value is less than or equal to α, H_0 is rejected, indicating sufficient evidence to support the alternative hypothesis (H_a). Conversely, if the p-value is greater than α, H_0 is not rejected, indicating insufficient evidence to support the alternative hypothesis.(Ott, 2016, p. 515)

3.2 Q-Q plot

A Normal Q-Q plot assesses whether a dataset follows a normal distribution by comparing its sample quantiles with the theoretical quantiles of a normal distribution. The data points are first sorted, and theoretical quantiles are computed based on the normal cumulative distribution function. These quantiles are then paired with the corresponding sorted data values and plotted. If the data is normally distributed, the points align closely with a straight reference line. Deviations from this line indicate departures from normality.(Pandit and Infield, 2018) For example, a Q-Q plot generated for a sample drawn from a normal distribution is shown below:

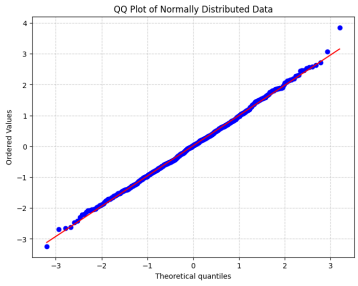

Figure 1: Normal Q-Q Plot for a Sample Dataset

3.3 Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical method designed to test the composite null hypothesis that a sample of n real-valued observations, $X_1, X_2, \ldots, X_n$, is drawn independently and identically from a normal distribution $N(\mu, \sigma^2)$. The test statistic W is used in the Shapiro-Wilk test to measure the degree of linearity between the ordered sample values and their expected counterparts under a normal distribution. The formula for W is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

where the weights a_i are determined based on the expected values of the order statistics of a standard normal distribution $N(0, 1)$. The exact distribution of W is non-standard. The p-value can be computed through numerical approximation, and by comparing it with the significance level (α), the decision on whether to reject the null hypothesis is made.(González-Estrada et al., 2022)

3.4 Levene's Test

Levene's test is used to assess the equality of variances across multiple groups. The null hypothesis (H_0) states that the variances are equal across all groups, while the alternative hypothesis (H_a) posits that at least one group has a different variance. The test begins by calculating the absolute deviation from the median (or mean) for each observation within a group. For each observation x_{ij} in group j (where $j = 1, 2, \ldots, k$, with k being the number of groups), the absolute deviation is calculated as:

$$y_{ij} = |x_{ij} - \text{median}(x)|$$

where y_{ij} is the absolute deviation of the i-th observation in group j, and median(x) is the median (or mean) of the group. These absolute deviations y_{ij} are then used to calculate the dispersion within each group.

The test statistic W is calculated as the ratio of the between-group variance in deviations to the within-group variance in deviations:

$$W = \frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}$$

where $\bar{y}_j$ is the mean of the absolute deviations for group j, $\bar{y}$ is the overall mean of the absolute deviations across all groups, n_j is the number of observations in group j, k is the number of groups, and N is the total number of observations.

The Wstatistic follows F-distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = N - k$.The decision to reject or fail to reject the null hypothesis is based on the p-value derived from the F-distribution. If $p < \alpha$ (e.g., 0.05), the null hypothesis of equal

variances is rejected, indicating significant differences in group variances. Levene's test is robust to non-normal data, especially when using the median, making it an effective tool for variance analysis. (Hosken et al., 2018)

3.5 One-Way ANOVA

One-way ANOVA is used to determine if there are statistically significant differences among group means by analyzing variances. It assumes normality (populations are normally distributed), homogeneity of variances (equal variances across groups), and random sampling with independence. These assumptions are crucial for the validity of the test results. For k groups, the null hypothesis (H_0) states that all group means are equal ($\mu_1 = \mu_2 = \cdots = \mu_k$), while the alternative hypothesis (H_a) posits that at least one pair of means differs ($\mu_i \neq \mu_j$ for some $i \neq j$).

The F-statistic for one-way ANOVA is calculated as the ratio of between-group variance to within-group variance:

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{k-1} \div \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{N-k} = \frac{MS_B}{MS_W}$$

where $\bar{x}_i$ is the mean of the i-th group, $\bar{x}$ is the overall mean of all observations, and x_{ij} is the j-th observation in the i-th group. The F-statistic follows an F-distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = N - k$. Based on the p-value derived from this distribution, the null hypothesis is rejected if $p < \alpha$, indicating that at least one of the group means differs.(Ott, 2016, p.744-751)

3.6 Two-Sample t-Test

The two-sample t-test compares the means of two independent groups, assuming normality, homogeneity of variance and independence of observations. The null hypothesis (H_a) states that the two population means are equal ($\mu_1 = \mu_2$), while the alternative hypothesis (H_a) asserts that they are not equal ($\mu_1 \neq \mu_2$).

The test statistic for the two-sample t-test is the ratio of the difference between the sample means to the standard error of the difference- which is derived from the pooled standard deviation. For two samples x_1 and x_2 with sample sizes n_1 and n_2, the formula for the t-statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where s_p is the pooled standard deviation, computed as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Here, s_1^2 and s_2^2 are the sample variances for groups 1 and 2, respectively.

The degrees of freedom (ν) for the test are calculated as:

$$\nu = n_1 + n_2 - 2$$

The t-statistic follows a Student's t-distribution with ν degrees of freedom. Based on the p-value derived from the distribution, the null hypothesis is rejected if the p-value is less than the significance level (α), indicating a statistically significant difference between the two group means. (Ott, 2016, p. 568-571)

3.7 Family-Wise Error Rate (FWER) and Post-Hoc Methods

When multiple statistical tests are performed simultaneously, the risk of Type I errors (false positives) increases. In standard hypothesis testing, the significance level remains at α, and the probability of avoiding a Type I error is $1 - \alpha$. However, when performing k independent tests, the probability of not making any Type I errors becomes $(1 - \alpha)^k$, and consequently, the probability of making at least one Type I error is:

$$\text{FWER} = 1 - (1 - \alpha)^k$$

As the number of tests k increases, the Family-Wise Error Rate (FWER) also increases. For instance, if $\alpha = 0.05$ and $k = 100$, the FWER becomes approximately 0.994, indicating a certainty of rejecting at least one true null hypothesis, when performing 100 independent tests.

To control the FWER and reduce the risk of Type I errors in multiple comparisons, post-hoc methods such as the Bonferroni correction and Tukey's Honestly Significant Difference (HSD) test are employed. These methods adjust the significance level to

account for the number of tests being performed, thereby controlling the overall error rate and ensuring more reliable conclusions.(Bender and Lange, 2001)

3.7.1 Tukey's HSD

Tukey's Honest Significant Difference (HSD) test provides a pairwise comparison of means. In the context of n groups with k observations each, the null hypothesis (H_0) posits that the means of the two groups are equal ($\mu_i = \mu_j$ for all i, j), while the alternative hypothesis (H_1) suggests that the means differ ($\mu_i \neq \mu_j$ for some $i \neq j$).Tukey's HSD test evaluates the absolute difference between the means of two groups, $|\overline{x}_i - \overline{x}_j|$, against a threshold known as the HSD. This threshold is derived using the Mean Square Error (MSE) from the ANOVA table, which represents the variability within the groups. The formula for the HSD when the sample sizes are equal (k observations per group) is:

$$\text{HSD} = q_{n,\alpha}\sqrt{\frac{MSE}{k}}$$

The critical value, $q_{n,\alpha}$, is based on the studentized range distribution, which considers the number of groups (n) and the significance level (α), effectively controlling the Family-Wise Error Rate (FWER). For unequal sample sizes (k_i for group i and k_j for group j), the HSD formula is adjusted to account for the differences in group sizes:

$$\text{HSD} = q_{n,\alpha}\sqrt{\frac{1}{2} \cdot MSE\left(\frac{1}{k_i} + \frac{1}{k_j}\right)}$$

A difference between two group means is considered statistically significant if the absolute difference exceeds the calculated HSD:

$$|\overline{x}_i - \overline{x}_j| \geq \text{HSD}$$

(Abdi and Williams, 2010)

3.7.2 Bonferroni Correction

The Bonferroni correction method is used to control the Family-Wise Error Rate (FWER) when conducting multiple statistical tests. To address the issue of increased FWER in the multiple testing, the Bonferroni correction adjusts the significance level for each test

by dividing the desired overall α (typically 0.05) by the number of tests k. The adjusted significance level for each test is:

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

Each p-value is then compared to the adjusted threshold to determine which null hypotheses are statistically significant. This method ensures that the overall FWER remains controlled at the desired level throughout the experiment. However, the Bonferroni correction can be conservative, especially with a large number of comparisons. As k increases, the adjusted significance level becomes smaller, increasing the risk of Type II errors (false negatives).

For confidence intervals, the Bonferroni correction widens each interval to ensure that the overall confidence level remains consistent across all tests, making them more conservative and reducing statistical power. Therefore, this method is particularly suitable when the number of comparisons is small.

(Emerson, 2020)

4 Statistical analysis

The statistical analysis was conducted using Python == 3.12.6(Van Rossum and Foundation, 2023) with VScode == 1.94.2 (Microsoft, 2023) as an IDE. The following Python packages are used: pandas == 2.2.3(McKinney et al., 2023) for data manipulation, matplotlib == 3.9.2(Hunter et al., 2023) and seaborn == 0.13.2(Waskom et al., 2023) for visualization, and statsmodels == 0.14.4(Seabold et al., 2023) and scipy == 1.14.1(Virtanen et al., 2023) for conducting statistical tests. Throughout the report, the significance level (alpha) for statistical tests is set to 0.05.

4.1 Descriptive Analysis

Table 1 illustrates the number of participants across different age groups with the corresponding mean and variance in finish times. The age group 35 has the highest number of participants, with 686, followed closely by age groups 40 and 30, which have 661 and 657 participants, respectively. Participation decreases steadily in older age groups, with the age group 55 having the fewest participants, at 116.

In terms of mean finish times, younger participants generally perform better, as evidenced by the comparatively lower mean finish times recorded for age groups 30, 35, and 40. In contrast, mean finish times increase with age and the highest mean finish time recorded for age group 55 at 16,764.08 seconds. A noticeable increase in average finish time is observed starting from age group 45, suggesting that age-related factors may influence performance.

Age Group	Count	Mean Finish Time (seconds)	Standard Deviation
30	657	15935.20	2091.52
35	686	15928.75	2168.78
40	661	15995.39	2123.20
45	458	16105.97	1972.15
50	251	16448.97	2189.25
55	116	16764.08	2045.04

Table 1: Summary statistics of finish times across age groups

These observations are further supported by the box plot in Figure2, which reinforces the trend of increasing finish times with age, with the highest median observed for age group 55. This trend of increasing median finish time becomes evident from age group 45, aligning with the observed increase in mean finish times from Table 1. The relatively consistent interquartile range (IQR) across all age groups indicates a similar spread of finish times around the median. However, the presence of outliers, particularly in younger age groups, suggests variability in individual performance within those groups.

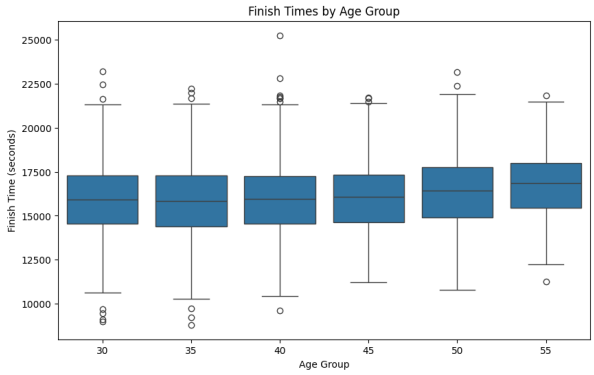

Figure 2: Boxplot of finish times across age groups

4.2 Assumptions for ANOVA and t-test

Statistical tests like ANOVA and t-tests require certain assumptions to be satisfied to ensure the validity of the results. The key assumptions include normality, homogeneity

of variances, and independence of observations. This section examines these assumptions for each of the six groups in the data.

4.2.1 Normality

To detect the normality of each age group, the Shapiro-Wilk test is conducted. The null hypothesis (H_0) states that the data follows a normal distribution, while the alternative hypothesis (H_1) suggests the opposite. From Table 2, it can be observed that for each age group (except Age Group 40), the corresponding p-value is higher than the significance level of 0.05, indicating that the normality assumption for these groups cannot be rejected based on the evidence. However, for Age Group 40, the p-value is 0.0010, which is less than 0.05, providing significant evidence against the null hypothesis. Therefore, the normality assumption for Age Group 40 fails to be accepted based on the data, indicating that this group significantly deviates from normality.

Age Group	**p-value**
Age Group 30	0.6471
Age Group 35	0.0595
Age Group 40	0.0010
Age Group 45	0.0785
Age Group 50	0.3964
Age Group 55	0.9255

Table 2: Shapiro-Wilk Test Results for Each Age Group

To further explore the deviation from normality for Age Group 40, a Q-Q plot is utilized. The Q-Q plot, shown in Figure 3, reveals that the data points align closely with the theoretical quantiles, with only slight deviations at the tails. This suggests that the deviation from normality is minimal and primarily influenced by extreme values.Therefore, the Q-Q plot suggests that the assumption of normality holds well for this group.

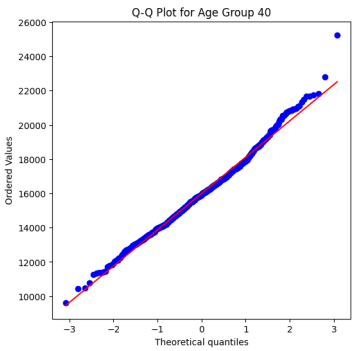

Figure 3: Q-Q Plot for Age Group 40

4.2.2 Homogeneity of Variance

Variance homogeneity across the six age groups is assessed through Levene's test, which tests the null hypothesis (H_0) that the variances are equal across these six different age groups, while the alternative hypothesis (H_1) suggests that the variance of at least one group is different. From Table 3, the p-value of 0.4528 indicates the lack of evidence to reject the homogeneous variance assumption. Therefore, the variance homogeneity assumption holds across the groups.

Number of Age groups	**Test Statistic**	**p-value**
6	0.9415	0.4528

Table 3: Levene's Test for Homogeneity of Variances Across Age Groups

4.2.3 Independence

For the analysis, we assume that the data set has been selected randomly and that each observation is independent.

4.3 One-Way ANOVA

An Analysis of Variance (ANOVA) is conducted to test whether the mean finish times across the six age groups are equal. The null hypothesis (H_0) states that the average finish time is equal across all age groups, while the alternative hypothesis (H_1) suggests that at least one group differs. From Table 4, the p-value of 0.000052 provides strong evidence to reject the null hypothesis (H_0), indicating that at least one group has a significantly different average finish time compared to the others.

Source of Variation	Sum of Squares (sum_sq)	Degrees of Freedom (df)	F-statistic (F)	p-value (PR(>F))
Between Groups	1.211834e+08	5.0	5.463353	0.000052
Within groups	1.252347e+10	2823.0	–	–

Table 4: One-Way ANOVA Results for Finish Times Across Age Groups

4.4 Two-Sample t-Test

A two-sample t-test is conducted to determine which pairs of age groups differ in average finish times. The null hypothesis (H_0: $\mu_i = \mu_j$, $i \neq j$) asserts that the average finish

times between any pairs of the six groups are equal, while the alternative hypothesis (H_1: $\mu_i \neq \mu_j$) asserts that the average finish times are different for the corresponding pair.

From Table 5, it is observed that for the following pairs of age groups: 30-50, 30-55, 35-50, 35-55, 40-50, 40-55, 45-50, and 45-55 - the t-statistic yields p-values less than the significance level of 0.05, indicating significant difference in the mean finish times between these pairs. For the other pairs of age groups, there is not enough evidence to reject the null hypothesis, hence no significant difference in their average finish time.

Comparison	**t-statistic**	**p-value**	**Bonferroni-adjusted p-value**
30-35	-0.055453	0.955786	1.000000
30-40	-0.518464	0.604222	1.000000
30-45	-1.372913	0.170056	1.000000
30-50	-3.267573	0.001126	0.016883
30-55	-3.948013	0.000086	0.001290
35-40	-0.569656	0.569006	1.000000
35-45	-1.403680	0.160686	1.000000
35-50	-3.243406	0.001223	0.018344
35-55	-3.867530	0.000119	0.001783
40-45	-0.881718	0.378119	1.000000
40-50	-2.856671	0.004379	0.065679
40-55	-3.615907	0.000319	0.004780
45-50	2.128910	0.033607	0.504101
45-55	-3.186399	0.001519	0.022787
50-55	-1.308564	0.191505	1.000000

Table 5: Two-Sample t-Test Results with Bonferroni Adjustment

4.5 Post-Hoc Analysis

Following the two-sample t-test, post-hoc analysis is conducted using Tukey's HSD test and the Bonferroni correction method to account for multiple comparisons and control the family-wise error rate (FWER). From the Appendix Table A1, the Tukey HSD test identified significant differences in mean finish times for the following seven pairs of age groups: 30-50, 30-55, 35-50, 35-55, 40-50, 40-55, and 45-55. For each of these pairs, the confidence intervals (CIs) do not include zero, confirming the presence of significant differences in the mean finish times between the groups.

In contrast, from Table 5, the Bonferroni correction revealed significant differences for the exact same pairs of age groups, with one exception: the 40-50 group. Unlike the Tukey HSD test, which identified this pair as significant, the Bonferroni correction did not. Overall, the results show that both methods are largely consistent, with only one test result differing between the two.

4.6 Comparison of t-Test with Post-Hoc Analysis

When comparing the results from Table 5 with those from Appendix Table A1, the t-test identifies significant differences in mean finish times for eight pairs of age groups. In contrast, the Tukey HSD method detects one lesser significant pair, and the Bonferroni correction identifies two lesser pairs. Specifically, while the t-test finds the 40-50 and 45-50 pairs to be significant, these pairs are not considered significant under the Bonferroni correction. Moreover, the Tukey HSD test identifies the 40-50 pair as significant, but does not find the 45-50 pair to be significant.

Apart from these differences, the methods are consistent in the specific pairs they classify as significant. These differences arise due to the issue of multiple comparisons, where the risk of Type I errors increases as more comparisons are made. The t-test does not adjust for this, leading to a higher family-wise error rate (FWER), which increases the likelihood of rejecting the null hypothesis when it is actually true. In contrast, both the Tukey HSD test and the Bonferroni correction adjust for multiple comparisons. The Bonferroni correction, however, is more conservative, applying a stricter threshold dividing the significance level(α) by the number of comparisons. As a result, it reduces the overall probability of rejecting the null hypothesis, leading to fewer significant findings. Because of this conservative approach, the Bonferroni correction results in the lowest number of significant pairs, while the t-test, being the least conservative, identifies the highest number of pairs as significant.

5 Summary

This report analyzed the average finish times of six age groups in the Berlin Marathon using both inferential and descriptive statistics. The analysis revealed consistent variability in finish times around the central value across all age groups, with an increasing

trend in average finish times as age increased. ANOVA testing indicated significant differences in mean finish times among the groups.

To identify specific pairs of age groups with significant differences, two-sample t-tests were performed, resulting in eight pairs with statistically significant differences. However, conducting multiple t-tests increased the Family-Wise Error Rate (FWER), leading to a higher risk of false positives. To address this, post-hoc methods, including Tukey's HSD and the Bonferroni correction, were applied. Tukey's method identified one fewer significant pair than the t-tests, while the Bonferroni correction, being more conservative, identified two fewer pairs.Further studies could explore methods like bootstrapping or Bayesian analysis for comparing finish times while controlling for the FWER.

Bibliography

Hervé Abdi and Lynne J. Williams. Tukey's honestly significant difference (hsd) test. *Encyclopedia of Research Design*, 3(1):1–5, 2010.

Ralf Bender and Stefan Lange. Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*, 54(4):343–349, 2001. ISSN 0895-4356. doi: https://doi.org/10.1016/S0895-4356(00)00314-0. URL `https://www.sciencedirect.com/science/article/pii/S0895 35600003140`.

Robert Wall Emerson. Bonferroni correction and type i error. *Journal of Visual Impairment & Blindness*, 114(1):77–78, 2020. doi: https://doi.org/10.1177/0145482X20901378. URL `https://doi.org/10.1177/0145482X20901378`.

Elizabeth González-Estrada, José A. Villaseñor, and Rocío Acosta-Pech. Shapiro-wilk test for multivariate skew normality. *Computational Statistics*, 37(4):1985–2001, 2022.

D.J. Hosken, D.L. Buss, and D.J. Hodgson. Beware the f test (or, how to compare variances). *Animal Behaviour*, 136:119–126, 2018. ISSN 0003-3472. doi: https://doi.org/10.1016/j.anbehav.2017.12.014. URL `https://www.sciencedirect.com/science/article/pii/S0003 47217304165`.

John D. Hunter, Michael Droettboom, Thomas A. Caswell, Eric Firing, Benjamin Root, et al. Matplotlib: Visualization with python, 2023. URL `https://matplotlib.org/`. Version 3.7.2.

Wes McKinney, Jeff Reback, Joris Van den Bossche, Tom Augspurger, et al. pandas: Python data analysis library, 2023. URL `https://pandas.pydata.org/`. Version 2.1.3.

Microsoft. Visual studio code, 2023. URL `https://code.visualstudio.com/`. Version 1.83.1.

Lyman Ott. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, Boston, MA, 2016. ISBN 9781305269477.

Ravi Pandit and David Infield. Qq plot for assessment of gaussian process wind turbine power curve error distribution function. In *9th European Workshop on Structural Health Monitoring*, February 2018. URL

http://www.bindt.org/events/ewshm-2018/. 9th European Workshop on Structural Health Monitoring Series (EWSHM) : EWSHM 2018, EWSHM ; Conference date: 10-07-2018 Through 13-07-2018.

Skipper Seabold, Josef Perktold, Nathaniel Smith, Kevin Sheppard, et al. Statsmodels: Statistical modeling and econometrics in python, 2023. URL https://www.statsmodels.org/. Version 0.14.0.

Guido Van Rossum and Python Software Foundation. Python, 2023. URL https://www.python.org/. Version 3.12.6.

Pauli Virtanen, Ralf Gommers, Travis Oliphant, Eric Jones, Charles Harris, Stefan van der Walt, et al. Scipy: Open source scientific tools for python, 2023. URL https://scipy.org/. Version 1.11.3.

Michael Waskom, Martin L. Martin, Anna Lam, Daniel F. Larremore, et al. Seaborn: Statistical data visualization, 2023. URL https://seaborn.pydata.org/. Version 0.12.2.

Appendix

A Additional tables

Table A1: Tukey's HSD Test with Confidence Interval

Comparison	meandiff	p-adj	lower	upper
30-35	-6.4516	1.0000	-334.3197	321.4164
30-40	60.1939	0.9955	-270.6934	391.0812
30-45	170.7685	0.7673	-194.8497	536.3867
30-50	513.7672	0.0131	68.0810	959.4534
30-55	828.8767	0.0013	223.9763	1433.7771
35-40	66.6456	0.9923	-260.7153	394.0065
35-45	177.2202	0.7305	-185.2097	539.6501
35-50	520.2189	0.0107	77.1444	963.2933
35-55	835.3283	0.0011	232.3497	1438.3070
40-45	110.5746	0.9551	-254.5889	475.7381
40-50	453.5733	0.0430	8.2600	898.8865
40-55	768.6827	0.0040	164.0571	1373.3084
45-50	342.9987	0.3014	-128.6942	814.6916
45-55	658.1082	0.0319	33.7986	1282.4178
50-55	315.1095	0.7669	-359.2213	989.4402