TU Dortmund

Introductory Case Studies

Project 1: Comparison of multiple distributions

Lecturers:
Prof. Dr. Crystal Wiedner
Prof. Dr. Rouven Michel
Prof. Dr. Marlies Hafer

Author: Siddhartha Karki (Matr. No : 238092)

Group number: 4

Group members:
Sagar Basnet

November 24, 2024

Contents

1 Introduction **1**

2 Problem Statement **2**

2.1 Data Set and Data Quality 2

2.2 Project Objectives 2

3 Statistical Methods **3**

3.1 Hypothesis Testing 3

3.1.1 Formulating Hypotheses 3

3.1.2 Test Statistic 3

3.1.3 Significance Level (α) 4

3.1.4 P-Value 4

3.1.5 Decision 4

3.2 Shapiro-Wilk test 5

3.3 Leven's Test 5

3.4 One-Way ANOVA (Analysis of Variance) Test 6

3.5 Post-hoc testing 7

3.5.1 Pairwise t-test 7

3.5.2 Bonferroni Correction 8

3.5.3 Tukey's HSD (Honestly Significant Difference) Test 9

3.6 Graphical Methods 9

3.6.1 Q-Q Plot (Quantile-Quantile Plot) 9

3.7 Tools 10

4 Statistical Analysis **10**

4.1 Frequency Distribution of variables 10

4.2 Global Test 11

4.2.1 Assumption 1: Normality 12

4.2.2 Assumption 2: Homogeneity of Varaince 12

4.3 Pairwise Differences 12

4.3.1 Pairwise t-test 12

4.3.2 Bonferroni correction 13

4.3.3 Tukey's Honest Significant Difference (HSD) test 13

4.4 Comparison of Statistical Significance: Non-Adjusted Test vs. Bonferroni and Tukey HSD Corrections 14

5 Summary **15**

Bibliography **16**

Appendix **17**

1 Introduction

Analyzing performance trends in long-distance running offers valuable insights into factors influencing physical endurance. In marathon running, age plays a critical role in determining performance, as physiological capabilities evolve over time. This project focuses on data from the Berlin Marathon, a prestigious annual event, to investigate how finish times vary among women runners across different age categories.

The dataset used in this analysis includes runner finish times (in seconds) and age groups, categorized into six intervals: 30, 35, 40, 45, 50, and 55 years. By restricting the dataset to women, the project aims to reduce variability associated with physiological differences between sexes, enabling a more focused examination of age-related trends. The primary goal of this project is to compare the distributions of finish times across these age categories and determine whether statistically significant differences exist.

Using descriptive statistics, we first summarize the data to identify patterns and central tendencies. Subsequently, statistical tests are applied to determine whether significant differences exist globally and between individual age categories.To ensure the reliability of the results, both Bonferroni and Tukey's HSD corrections are applied to pairwise comparisons. These methods help mitigate the risk of false positives when testing multiple hypotheses. Additionally, comparisons with unadjusted results provide a broader perspective on the impact of these corrections.

The report is structured as follows: Section 2 provides a detailed overview of the dataset. Section 3 explains the statistical tests and correction techniques used. Section 4 presents the results and discusses their interpretation. Section 5 concludes with a summary of the findings, their implications for marathon performance analysis, and potential directions for future research.

2 Problem Statement

2.1 Data Set and Data Quality

The dataset contains information on participants in the Berlin Marathon, focusing on their race finish times and age categories. The data is sourced from Kaggle and is filtered to include only female participants. This is an observational dataset, likely collected through race records, and includes variables relevant to marathon running performance. The dataset represents a stratified sample of marathon runners based on gender (restricted to women) and age groups. It consists of 2,829 observations of two key variables:
time: The total time (in seconds) taken by a runner to complete the marathon.
age_group: A categorical variable representing the age category of the runner, grouped into six intervals: 30, 35, 40, 45, 50, and 55.

No missing values are identified in the dataset. The variable **time** is numerical and measured continuously, while the variable **age_group** is nominal and grouped into fixed categories, which avoids ambiguity. The data is assumed to be accurate due to the structured recording of marathon finish times.

2.2 Project Objectives

The objective of this project is to analyze the finish times of female marathon runners and explore how these times vary across six age groups: 30, 35, 40, 45, 50, and 55. The analysis starts with descriptive statistics, calculating the mean, median, interquartile range, and standard deviation, offering a solid understanding of the data. Visualizations, such as histograms and boxplots, further illustrate trends in finish times across different age groups.

To determine if the differences in finish times are statistically significant, a one-way ANOVA is performed to assess whether the mean finish times vary between the six age groups. The assumptions of normality and homogeneity of variances are checked using QQ plots and boxplots, ensuring the results are valid.

After performing the ANOVA, pairwise comparisons are conducted using t-tests to identify specific differences between age groups. Multiple testing corrections, including the

Bonferroni correction and Tukey's Honest Significant Difference (HSD), are applied to adjust p-values, controlling for the error rate and providing reliable confidence intervals.

The results from the Bonferroni and Tukey HSD methods are compared with unadjusted p-values to evaluate their impact on the findings. This comparison provides valuable insights into the strengths and limitations of each correction method and its effect on identifying significant differences in finish times across age groups. Through this analysis, the project aims to uncover patterns that highlight the influence of age on marathon performance.

3 Statistical Methods

3.1 Hypothesis Testing

Hypothesis testing is a statistical technique used to evaluate assumptions about a population using sample data (Ning-Zhong Shi, 2008).

3.1.1 Formulating Hypotheses

A hypothesis is a claim about a population that can be tested. The two main types of hypotheses are:

Null Hypothesis (H_0): Suggests no effect or difference exists. For example,

$$H_0 : \mu = a \quad \text{(The population mean equals } a\text{)}.$$

Alternative Hypothesis (H_a): Proposes that there is an effect or difference. It can be:

$$H_a : \mu \neq a$$

3.1.2 Test Statistic

The test statistic compares the observed sample data to the expected data under the null hypothesis. Common test statistics include:

t-score: Applied when the sample size is small or the population standard deviation is unknown. The formula is:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where:

- $\bar{x}$ = sample mean
- μ = population mean (under H_0)
- s = sample standard deviation
- n = sample size

3.1.3 Significance Level (α)

The significance level, typically set at 0.05 or 0.01, defines the threshold for rejecting the null hypothesis. It represents the maximum acceptable probability of committing a Type I error (incorrectly rejecting H_0).

3.1.4 P-Value

The **p-value** represents the probability of obtaining a test statistic as extreme as or more extreme than the one observed, assuming the null hypothesis is true. A **small p-value** (typically $\leq \alpha$, where α is the significance level, usually 0.05) suggests that the observed data is inconsistent with the null hypothesis, leading us to reject it in favor of the alternative hypothesis. Conversely, a **large p-value** indicates insufficient evidence to reject the null hypothesis.

3.1.5 Decision

Based on the test results:

- **Reject H_0:** The evidence supports H_a, suggesting a significant effect or difference.
- **Fail to reject H_0:** There is not enough evidence to support H_a, and the null hypothesis stands.

3.2 Shapiro-Wilk test

The Shapiro-Wilk test assesses whether a data sample follows a normal distribution by comparing the observed data to a normal curve. The test generates a statistic W, which quantifies the alignment between the sample and the normal distribution. The null hypothesis (H_0) assumes that the data is normally distributed, while the alternative hypothesis (H_a) suggests that the data is not normally distributed (SHAPIRO and WILK, 1965).

Test Statistic:
The Shapiro-Wilk statistic is calculated as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where $x_{(i)}$ are the ordered sample values, a_i are constants, and $\bar{x}$ is the sample mean.

Interpretation:
If the p-value > 0.05, fail to reject H_0, indicating normality.
If the p-value ≤ 0.05, reject H_0, indicating non-normality.
A test statistic W close to 1 suggests the data is normally distributed.

3.3 Leven's Test

The test is designed to check whether the variances of the groups are equal, which is an important assumption in many statistical tests, such as ANOVA. The null hypothesis of Levene's test states that the variances across all groups are equal, while the alternative hypothesis suggests that at least one group has a variance different from the others(Brown and Forsythe, 1974).

Test Statistic:
The Levene's test statistic is calculated as:

$$W = \frac{(n-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - z_i)^2}{\sum_{i=1}^{k} n_i(z_i - \bar{z})^2}$$

where, n is the sample size of data, k is the number of groups within the variable, and z_{ij} is the deviation of the j-th observation from the group mean (or median, or trimmed mean). Specifically, $z_{ij} = |y_{ij} - y_i|$, where y_{ij} is the individual data point and y_i is the

mean (or median, or trimmed mean) of the group i. z_i is the mean (or median) of group i, and $\bar{z}$ is the overall mean of the deviations.

Interpretation:
If the p-value > 0.05, fail to reject H_0, indicating that the variances are equal across groups. If the p-value ≤ 0.05, reject H_0, indicating that at least one group has a variance different from the others. A test statistic W significantly larger than the critical value suggests that the variances are unequal across groups.

3.4 One-Way ANOVA (Analysis of Variance) Test

One-way ANOVA is a statistical test used to compare the means of three or more groups to determine if there is a significant difference between them(Illowsky and Dean, 2017). We chose the one-way ANOVA, which tests the null hypothesis that the mean finish times are equal across all age groups. It is appropriate here as we have one independent variable (age group) and one dependent variable (finish time), making it ideal for assessing the impact of age on performance.

The null hypothesis (H_0) assumes that the means of all groups are equal, represented as...

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

where k is the number of groups, while the alternative hypothesis (H_a) suggests that at least one group mean is different from the others.

Test Statistic:

Between-Group Sum of Squares (SSB):

$$SSB = \sum n_j(\bar{y}_j - \bar{y})^2$$

where n_j is the number of observations in group j, and $\bar{y}_j$ is the mean of group j.

Within-Group Sum of Squares (SSW):

$$SSW = \sum \sum (y_{ij} - \bar{y}_j)^2$$

where y_{ij} is an individual observation in group j.

F-Statistic:

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

where $MSB = \frac{SSB}{k-1}$ and $MSW = \frac{SSW}{n-k}$.

Find the Critical Value or P-Value Compare the F-statistic to the critical value from the F-distribution table or use statistical software to compute the p-value.

Make a Decision

- If $p < \alpha$ (e.g., 0.05), reject H_0.
- A significant result indicates that at least one group mean is different from the others.

3.5 Post-hoc testing

Post-hoc testing refers to statistical tests that are conducted after an initial hypothesis test, like ANOVA, to further investigate where the differences lie between specific groups. In our project comparing finish times between age groups of women marathon runners, a one-way ANOVA show a significant difference in finish times across different age groups. Post-hoc tests will allow us to identify which exact age groups have different mean times, providing insights into which age categories differ in performance.

3.5.1 Pairwise t-test

A pairwise t-test is used to compare the means of two groups at a time to check if there is a statistically significant difference between them. The steps for conducting a pairwise t-test involve stating the hypotheses (Ross and Willson, 2017).

The null hypothesis (H_0) assumes that there is no difference between the means of the two groups, represented as:

$$H_0 : \mu_1 = \mu_2$$

The alternative hypothesis (H_1) suggests that there is a significant difference between the means of the two groups, represented as:

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

Test Statistic:
The formula for the t-statistic in a pairwise t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups.
- s_1^2 and s_2^2 are the sample variances of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

Calculate the Degrees of Freedom (df):
The degrees of freedom for the pairwise t-test are calculated using:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

This formula adjusts for the sample sizes and variances in both groups.

Determine the p-value:
After calculating the t-statistic and degrees of freedom, use a t-distribution table or statistical software to find the **p-value**. The p-value tells you the probability of observing the data assuming the null hypothesis is true.

Make a Decision:

- If the **p-value** is less than the significance level α, **reject the null hypothesis** and conclude that there is a significant difference between the two group means.
- If the **p-value** is greater than α, **fail to reject the null hypothesis**, meaning there is no significant difference between the two group means.

3.5.2 Bonferroni Correction

The **Bonferroni correction** is a statistical method used to address the problem of multiple comparisons. When performing multiple hypothesis tests, the chance of committing a Type I error (false positive) increases. The Bonferroni correction adjusts the

significance level (α) by dividing it by the number of tests being performed. This ensures that the overall Type I error rate remains controlled (Abdi, 2007).

The adjusted significance level is:

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

where:

- α is the original significance level (e.g., 0.05),
- m is the number of comparisons being made.

3.5.3 Tukey's HSD (Honestly Significant Difference) Test

Tukey's HSD test is a post-hoc analysis used after an ANOVA to find out which specific group means are different. Unlike the Bonferroni correction, which adjusts the significance level, Tukey's test computes the HSD statistic, which helps to identify which pairs of group means differ significantly while controlling the overall Type I error rate (Kramer, 1956).

The Tukey HSD statistic is calculated as:

$$\text{HSD} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MSE}{n}}}$$

Where:

- $\bar{X}_i$ and $\bar{X}_j$ are the sample means for groups i and j,
- MSE is the mean square error (from the ANOVA),
- n is the sample size for each group (assumed to be equal).

3.6 Graphical Methods

3.6.1 Q-Q Plot (Quantile-Quantile Plot)

A Q-Q plot, or quantile-quantile plot, is a graphical method used to compare the quantiles of a dataset to the quantiles of a theoretical distribution or another dataset. It helps

assess whether the dataset follows a specific distribution, such as a normal distribution, or whether two datasets come from the same population. In this project, Q-Q plots are used to check if the finish times across different age groups in the Berlin Marathon dataset are normally distributed (Efron and Gong, 1983).

3.7 Tools

The analysis was conducted using Python (version 3.10.4) and a range of essential libraries, including Pandas (version 1.3.4), NumPy (version 1.20.3), and Seaborn (version 0.13.0) for data manipulation and visualization. Statistical analysis was performed using SciPy (for tests like t-tests and Levene's test), statsmodels (for performing Tukey's HSD and multiple testing corrections), and Matplotlib (for creating plots). These tools were instrumental in implementing the analysis and generating insightful visualizations.

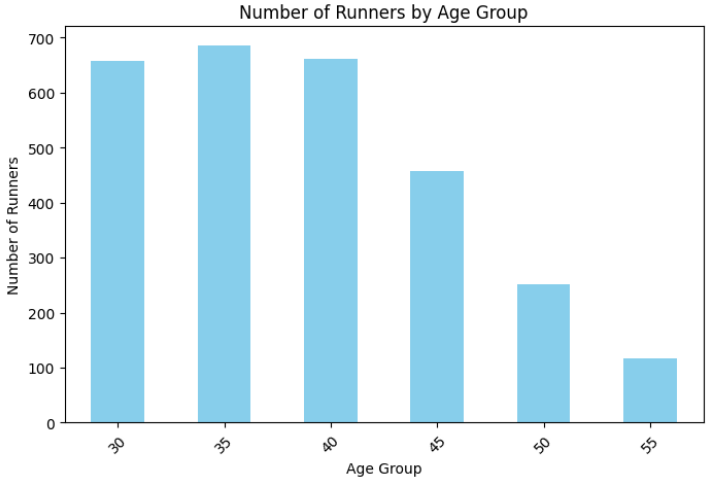
4 Statistical Analysis

4.1 Frequency Distribution of variables

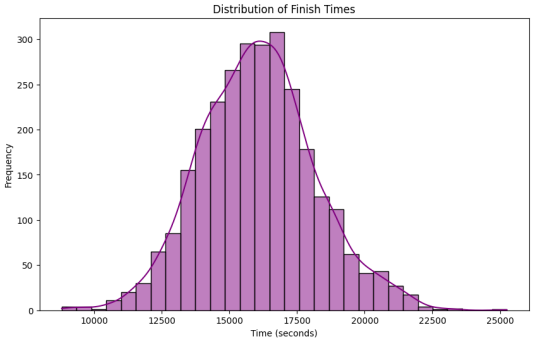
The bar chart (Figure 1a) illustrates the count of participants across different age groups, providing insights into the distribution of runner demographics. Age group 30 has the highest number of participants (657), followed by groups 35 and 40. In contrast, age group 55 has the lowest count, with only 116 participants. This pattern indicates that younger age groups are more prominently represented in the marathon.

The histogram (Figure 1b) visualizes the overall distribution of marathon finish times, showing approximately normal with slightly right-skewed distribution. Most finish times are concentrated between 14,000 and 17,500 seconds, with an average (mean) time of approximately 16,055 seconds. The spread is considerable, ranging from a minimum of 8,802 seconds to a maximum of 25,254 seconds, highlighting the performance variability among runners. The distribution suggests that the central tendency (mean and median) lies around 16,000 seconds, with a moderate tail of slower finishers.

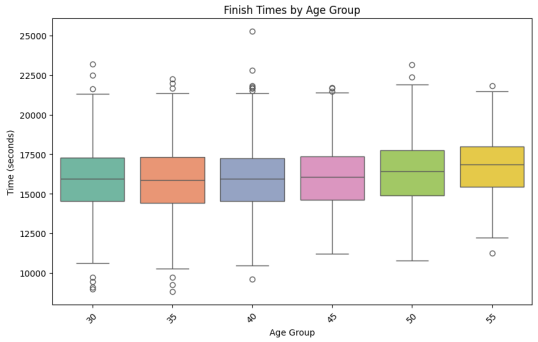
The boxplot (Figure 1c) compares finish times across age groups, highlighting the central tendency and variability within each group. Younger age groups (30 and 35) exhibit lower medians and tighter interquartile ranges, indicating faster and more consistent performances. In contrast, older age groups (50 and 55) have higher medians and greater



(a) Bar Chart of Participants by Age Group Label



(b) Histogram of Finish Times



(c) Boxplot of Finish Times by Age Group

Figure 1: Frequency distribution of variables

variability, reflecting slower and more dispersed finish times. The central tendency within each group confirms that performance generally declines with age, accompanied by increasing inconsistency.

4.2 Global Test

In order to investigate whether the finish times differ across the different age categories, we perform a global test using One-Way ANOVA. Before applying the ANOVA, we first check the assumptions of normality and homogeneity of variances.

4.2.1 Assumption 1: Normality

The normality of the data for each age group was assessed using the Shapiro-Wilk test. The results indicate that the assumption of normality holds for most age groups, as the p-values were greater than the significance level of 0.05, suggesting that the data for those groups follows a normal distribution.A QQ plot (Figure 3) also visually supports this, demonstrating that the data in these groups are normally distributed.

4.2.2 Assumption 2: Homogeneity of Varaince

Levene's test is used to check the assumption of homogeneity of variances across the age groups. The result shows a test statistic of 0.9415 and a p-value of 0.4528, which is greater than 0.05. Therefore, we fail to reject the null hypothesis, indicating that the variances are equal across the groups, satisfying the assumption of homogeneity of variances.

Since, all the assumptions of ANOVA are satisfied, the testing of significance difference can be performed. The results of the one-way ANOVA indicate a significant difference in finish times across age groups. The F-statistic is 5.463, and the p-value is 5.215e-05, which is less than 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference in the finish times between the age groups.

4.3 Pairwise Differences

To investigate pairwise differences in finish times across age groups, we performed two-sample tests for each pair of age categories. To account for multiple comparisons, we applied two methods for adjusting the results: Bonferroni correction and Tukey's Honest Significant Difference (HSD) test.

4.3.1 Pairwise t-test

The analysis examines pairwise differences in finish times across various age groups using two-sample t-tests. The null hypothesis assumes no difference in mean finish times between any two groups, while the alternative hypothesis posits that differences exist. The results from Table 1 in Appendix identifies significant differences in finish times for several pairs of age groups. Specifically, Group 35 significantly differs from

Groups 50 ($p = 0.0012$) and 55 ($p = 0.0001$), and Group 30 differs significantly from Groups 50 ($p = 0.0011$) and 55 ($p = 0.0001$). Additionally, Group 40 differs from Groups 50 ($p = 0.0044$) and 55 ($p = 0.0003$), and Group 45 differs significantly from Group 55 ($p = 0.0015$). These results, highlighted in the table, suggest notable differences in finish times among these specific group comparisons.

4.3.2 Bonferroni correction

The Bonferroni correction identifies significant differences in finish times across specific age group pairs. From Table 2 in Apendix Adjusted results show that Group 35 significantly differs from Groups 50 ($p = 0.0183$) and 55 ($p = 0.0018$), while Group 30 significantly differs from Groups 50 ($p = 0.0169$) and 55 ($p = 0.0013$). Additionally, Group 40 differs significantly from Group 55 ($p = 0.0048$), and Group 45 differs from Group 55 ($p = 0.0228$). These findings, detailed in Table 2, highlight the adjusted significant pairwise differences in finish times.

4.3.3 Tukey's Honest Significant Difference (HSD) test

The Tukey's Honest Significant Difference (HSD) test reveals significant pairwise differences in mean finish times across several age groups. From table 3 in Apendix, Group 30 shows significant differences from Groups 50 ($p = 0.0131$) and 55 ($p = 0.0013$), with mean differences of 513.77 and 828.88, respectively. The confidence intervals (CI) for these differences range from 68.08 to 959.45 for Group 30 vs. Group 50, and from 223.98 to 1433.78 for Group 30 vs. Group 55, indicating that the true difference in means lies within these intervals. Group 35 differs significantly from Groups 50 ($p = 0.0107$) and 55 ($p = 0.0011$), with mean differences of 520.22 and 835.33, respectively. The CIs for these comparisons range from 77.14 to 963.29 for Group 35 vs. Group 50, and from 232.35 to 1438.31 for Group 35 vs. Group 55, further supporting the significant differences. Additionally, Group 40 significantly differs from Groups 50 ($p = 0.0430$) and 55 ($p = 0.0040$), with mean differences of 453.57 and 768.68, and CIs of 8.26 to 898.89 for Group 40 vs. Group 50, and 164.06 to 1373.31 for Group 40 vs. Group 55. Lastly, Group 45 significantly differs from Group 55 ($p = 0.0319$), with a mean difference of 658.11 and a CI ranging from 33.80 to 1282.42. The confidence intervals help us understand the precision of the mean differences, with the wider intervals indicating

greater uncertainty, and the narrower intervals reflecting more precise estimates of the true differences in finish times.

4.4 Comparison of Statistical Significance: Non-Adjusted Test vs. Bonferroni and Tukey HSD Corrections

The comparative results of these approaches are summarized in a scatter plot, where p-values for each method are plotted against the group comparisons. In the figure, a black dashed line represents the significance threshold ($\alpha = 0.05$), blue circles indicate p-values from the Paired t-test (Non-Adjusted), red crosses represent p-values from the Bonferroni correction, and green triangles denote p-values from Tukey HSD.

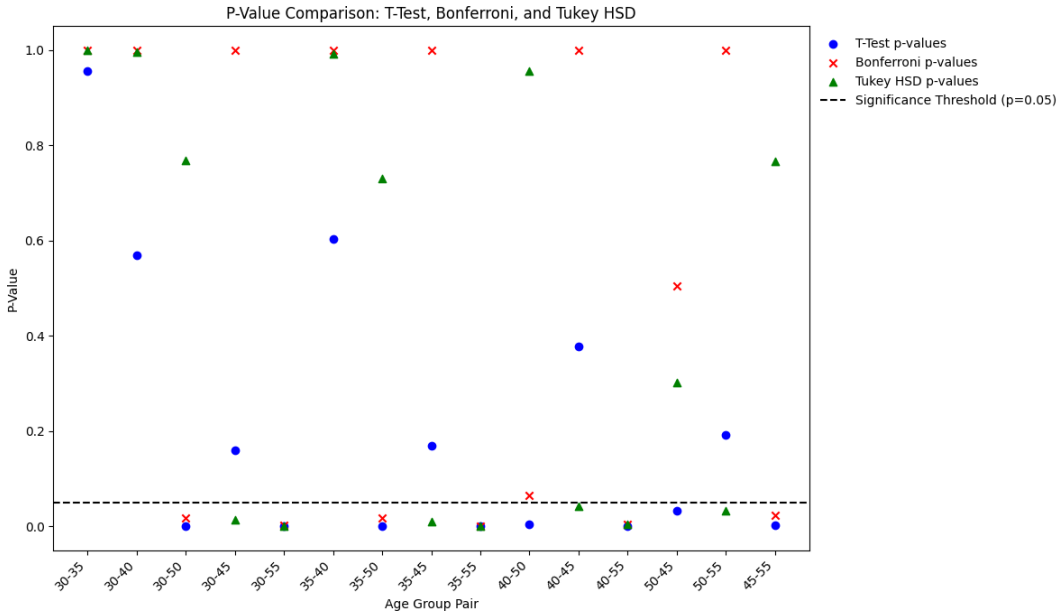


Figure 2: Scatter plot showing comparison of results of three approaches.

The Non-Adjusted test shows the most significant results, as it does not correct for multiple testing, increasing the likelihood of false positives (Type I errors). Bonferroni, being highly conservative, adjusts p-values to control for family-wise error, reducing false positives but increasing false negatives (Type II errors). Tukey HSD, less stringent than Bonferroni, strikes a balance by controlling Type I errors while retaining more power to detect true differences. Overall, the Non-Adjusted test is the most liberal, Bonferroni is the strictest, and Tukey HSD provides a middle ground suitable for grouped comparisons.

5 Summary

This project analyzed the performance of female marathon runners across six age groups (30, 35, 40, 45, 50, and 55) using data from the Berlin Marathon. The primary objective was to investigate whether finish times differed significantly across these groups and compare the impact of various statistical methods. The dataset included 2,829 observations of two variables: finish times (seconds) and age groups. Bar charts revealed that age group 30 had the highest participation, while group 55 had the lowest. Histograms showed a right-skewed distribution of finish times, and boxplots highlighted a decline in performance and increased variability with age.

A one-way ANOVA was conducted to assess whether the finish times differed significantly across age groups. Before performing the ANOVA, assumptions of normality and homogeneity of variances were checked. The Shapiro-Wilk test and visual inspection via QQ plots confirmed normality for most groups, and Levene's test verified the homogeneity of variances ($p = 0.4528$). The ANOVA results indicated a significant difference in mean finish times between groups ($F = 5.463, p = 5.215 \times 10^{-5}$). Post hoc analysis was carried out to identify specific group differences using pairwise t-tests, Bonferroni correction, and Tukey's HSD test.

The pairwise t-tests revealed significant differences between several age groups, particularly between groups 30, 35, and the older groups 50 and 55. For example, significant differences were observed between groups 30 and 55 ($p = 0.0001$) and groups 35 and 50 ($p = 0.0012$). Bonferroni correction, while more conservative, confirmed significant differences for fewer pairs, such as groups 30 vs. 55 ($p = 0.0013$) and 35 vs. 55 ($p = 0.0018$). Tukey's HSD provided further insights, offering confidence intervals for the mean differences, highlighting consistent significant results across groups such as 30 vs. 55 (CI: 223.98 to 1433.78) and 35 vs. 55 (CI: 232.35 to 1438.31).

These findings underline the decline in performance with age, accompanied by increased variability in finish times. The comparison of statistical methods highlighted the advantages of Tukey's HSD for its balance between Type I error control and statistical power, making it the preferred method in this context. Future research could extend this analysis by considering additional factors like environmental conditions, runner experience, and temporal trends in race performance to provide a more comprehensive understanding of marathon dynamics.

Bibliography

Abdi, H. (2007). The bonferroni and sidak corrections for multiple comparisons. In Salkind, N. J., editor, *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

Efron, B. and Gong, T. (1983). *A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation*. Rage.

Illowsky, B. and Dean, S. (2017). *Introductory Statistics*. OpenStax. Open Access, available at `https://openstax.org/books/introductory-statistics/pages/13-1-one-way-anova`.

Kramer, C. Y. (1956). *Extension of multiple range tests to group means with unequal numbers of replications*, volume 12. AlphaPublication.

Ning-Zhong Shi, J. T. (2008). *Statistical Hypothesis Testing: Theory and Methods*. World Scientific Publishing Co.Pte.Ltd. Accessed: 2024-11-20.

Ross, A. and Willson, V. L. (2017). Paired samples t-test. In *Basic and Advanced Statistical Tests*, pages 123–135. SensePublishers, Rotterdam.

SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Appendix

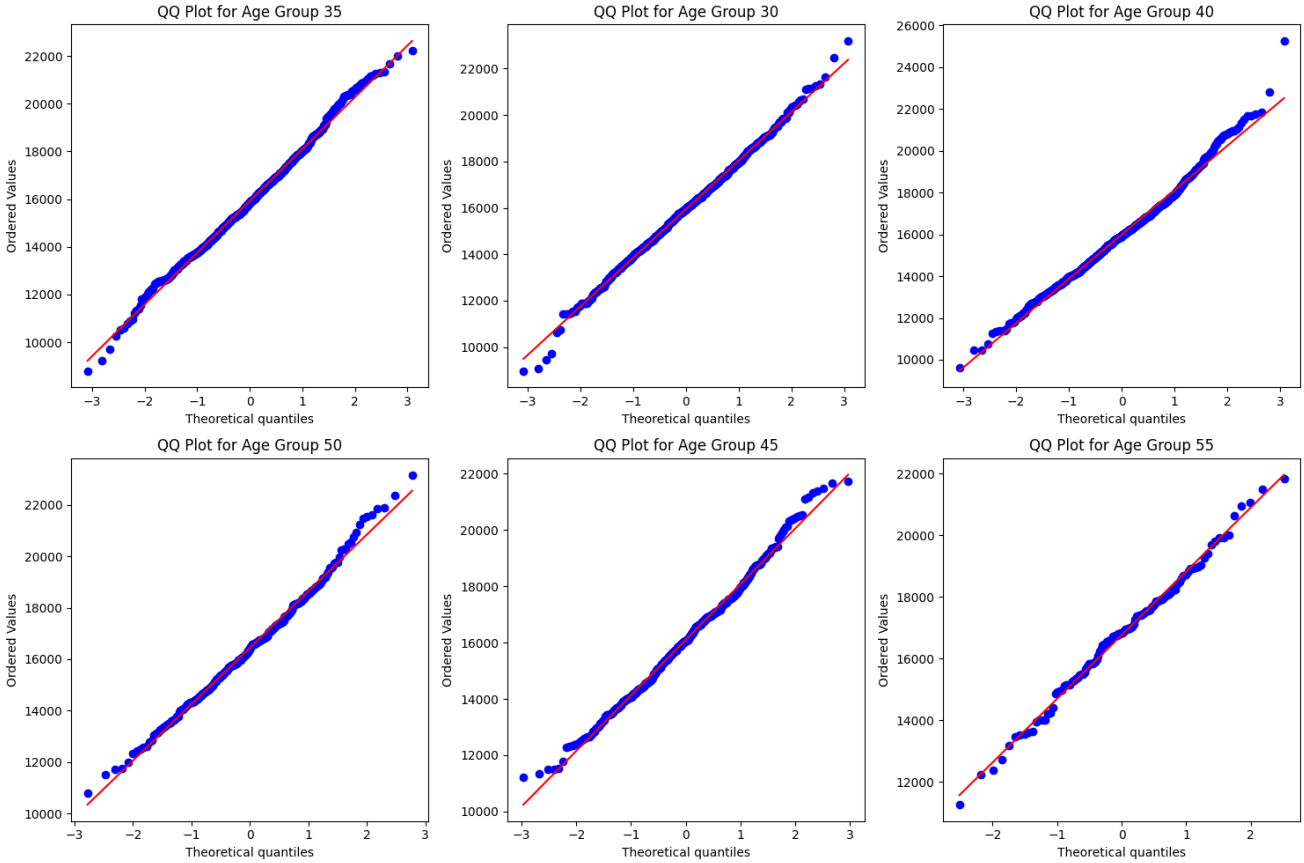


Figure 3: Scatter plot showing comparison of results of three approaches.

Group 1	Group 2	P-Value (Non-Adjusted)	Significant (Unadjusted)
35	40	0.569006	False
35	50	0.001223	True
35	45	0.160686	False
35	55	0.000119	True
30	40	0.604222	False
30	50	0.001126	True
30	45	0.170056	False
30	55	0.000086	True
40	50	0.004379	True
40	45	0.378119	False
40	55	0.000319	True
50	55	0.191505	False
45	55	0.001519	True

Table 1: Non-Adjusted P-Value Results for Finish Times by Age Group

Group 1	Group 2	Adjusted P-Value (Bonferroni)	Significant (Bonferroni)
35	40	1.000000	False
35	50	0.018344	True
35	45	1.000000	False
35	55	0.001783	True
30	40	1.000000	False
30	50	0.016883	True
30	45	1.000000	False
30	55	0.001290	True
40	50	0.065679	False
40	45	1.000000	False
40	55	0.004780	True
50	55	1.000000	False
45	55	0.022787	True

Table 2: Bonferroni Adjusted P-Value Results for Finish Times by Age Group

Group 1	Group 2	Mean Difference	P-Value (Tukey HSD)	Lower CI	Upper CI	Reject
30	35	-6.4516	1.0000	-334.3197	321.4164	False
30	40	60.1939	0.9955	-270.6934	391.0812	False
30	45	170.7685	0.7673	-194.8497	536.3867	False
30	50	513.7672	0.0131	68.0810	959.4534	True
30	55	828.8767	0.0013	223.9763	1433.7771	True
35	40	66.6456	0.9923	-260.7153	394.0065	False
35	45	177.2202	0.7305	-185.2097	539.6501	False
35	50	520.2189	0.0107	77.1444	963.2933	True
35	55	835.3283	0.0011	232.3497	1438.3070	True
40	45	110.5746	0.9551	-254.5889	475.7381	False
40	50	453.5733	0.0430	8.2600	898.8865	True
40	55	768.6827	0.0040	164.0571	1373.3084	True
45	50	342.9987	0.3014	-128.6942	814.6916	False
45	55	658.1082	0.0319	33.7986	1282.4178	True
50	55	315.1095	0.7669	-359.2213	989.4402	False

Table 3: Tukey HSD P-Value Results for Finish Times by Age Group