TU Dortmund

Introductory Case Studies

Project 2 – Comparison of multiple distributions

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Philipp Adämmer

Dr. Andrea Bommert

M. Sc. Hendrik Dohme

Author: Jorge Eduardo Durán Vásquez

Group 11

Group Members: Ahmed Amr Mohamed Sami Souidan, Akshatha Krishnananda Shanbhag, Felix Fikowski, Ibtisam Ahmed Queshi

June 11, 2021

Contents

1. Introduction

The analysis of the relation of the height of the players between the different German national teams gives useful information for many fields like the performance advantages of players, the production of clothes, the design of sports equipment, as well as for the research of medical, physiological, or nutritional treatment for groups of players of different sports. In the clothes industry, the analysis of relation of the height between the players gives insights for creating items that can be used for many players in different sports, which through standardization of production can lead to benefits for clothes producers. Furthermore, height differences between the players in different sport disciplines leads to advantages and disadvantages for the players depending on how the sports are designed. This analysis can help the coach of the teams to select the best teams that fulfill specific characteristics. Some advantages of having a height below the average are for example, more agility in certain movements, more balance because the lower center of gravity, faster rotational capability, and among others lower probability to break bones, because in many cases they are denser when the people is smaller (Samaras T., 2007). On the other hand, some advantages for players and sports that have heights above the average are for example that the players have greater reach that allows them to have larger capacity of catching (very important in sports like Basketball and Volleyball), also players with greater height have more visibility, people with larger height also can have greater weight and more momentum that help in some sports like American football or Soccer (Samaras T., 2007).

The purpose of this report is to find differences between population parameters of sports groups based on samples with a small size. To find and evaluate relations between *the height* means of the players, hypothesis testing techniques are used. Hypotheses regarding the relation of the median value of the height between six sports are proposed. Then, specific tests of significance such as the analysis of variance and the two-sample pooled t-test are performed for those hypotheses. These tests are conducted after checking the respective assumptions and conditions within the sample groups that are involved in this analysis. In addition, some techniques used to analyze small sample sizes and improve the significance level of the results are used.

This report is structured as follows, in section 2, information of the data set, as well as the process of data collection and data quality is described. Section 3 explains the software tools that are used and the statistical methods, which include hypothesis testing, the F-test, the Pooled T-test and some additional statistical tools like Bonferroni method and Normal probability

plots. In section 4, the results of the analysis are presented. Finally, section 5 concludes the report, discusses the current constraints of this analysis, and provides recommendations for future research.

2. Problem statement

The data set used in this report was provided by the instructors of the subject Introductory Case Studies of the Master of Science in Data Science at TU Dortmund University. It contains the information of the heights of the players of six German men's national teams. The *names* and the *height* of each player in the data set were collected by hand directly asking the coach of each national team or with information found on the internet.

The considered sample size amounts to 112 observations. Each observation includes the *height* given in centimeters and the *name* of the players of six German men's national teams: Basketball, Handball, Ice Hockey, Soccer, Volleyball, and Water Polo. The *name* and *sport* of each player are string parameters, and the *height* is an integer variable given in a ratio scale. The number of players and information about the distribution of the *height* of the players of each national team are presented in Table 1 in section 4.1.

It is important to remark that the data set contains missing values in the *names* of the 16 players of Water Polo. This is not a problem for the analysis of this report, though, because for this analysis mainly the values of *height* and the classification into different sports are used. To understand the distribution of the *height* for each team descriptive information as provided in Table 1 and the box plot in Figure 1 are used. In addition, to investigate if the *height* of the players differs between the six sports, the hypothesis of the equality of the means of all teams is assessed using the F-test. Finally, to evaluate the hypothesis of pairwise equality of *heights* between the players in each national team a pairwise Pooled T-test is applied.

3. Statistical methods

3.1 Hypothesis testing

A statistical hypothesis is a possible affirmation about the distribution of one or more random variables (Mood, 1974). This hypothesis can be evaluated using the statistical technique of hypothesis testing. In this technique two hypotheses are specified. First, the null hypothesis (H_0) which is a possible affirmation regarding a population model parameter (Sharpe et al, 2012). On the other hand, the alternative hypothesis (H_A) contains the values of the parameter

that are considered plausible if the null hypothesis is rejected (Sharpe et al, 2012). The hypothesis test follows a strict path that can be divided in hypotheses, model, mechanics, and conclusion (Sharpe et al, 2012).

First, for the hypotheses step the null and the alternative hypothesis are stated (Sharpe et al, 2012). Second, for the model section the sampling distribution of the statistic that will be used to test the null hypothesis is specified (Sharpe et al, 2012). In the same step the condition and the assumptions of the selected statistical distribution are evaluated (Sharpe et al, 2012). Third, "the mechanics" is the step in which the test statistic and the p-value are calculated. Finally, the conclusion is a rejection or non-rejection of the null hypothesis based on the results of the hypothesis test and the context of the analysis (Sharpe et al, 2012).

Important terms that are used in the hypothesis testing process are the degrees of freedom, the p-value, confidence level, and the critical value. First, "degrees of freedom" (df) is a parameter that defines which distribution of a family of distributions must be used (Sharpe et al, 2012). The shape of each statistic is affected by the degrees of freedom and each of them has its own way to calculate the df and in this report, the df will be specified when the statistic is defined. Second, the p-value is the probability that the observed statistic value could occur if the null hypothesis was correct (Sharpe et al, 2012). After performing each test, the p-value is calculated and displayed in the summary results of the respective test.

Finally, the critical value is the value in the distribution that is calculated with the corresponding significance level (Sharpe et al, 2012). If the p-value obtained from the test is lower than the significance level, the null hypothesis is rejected. But if the p-value is greater or equal to the significance level, the null hypothesis is not rejected (Sharpe et al, 2012). In this report the confidence level used is 95% which corresponds to a significance level of 0.05.

3.2 F-test

The F-test is used when metrics of three or more groups are analyzed (Sharpe et al, 2012). To test whether all *"k"* means of *"k"* groups are equal, the hypothesis becomes:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k \quad H_A: at\ least\ one\ mean\ is\ different$$

The F-statistics compares two quantities that measure variation, called mean squares, namely the variance of the means and the expected value of that variance (Sharpe et al, 2012). In the numerator of the next formula is the Mean Square due to Treatment (*MST*), which measures

variation between the groups. In the denominator is the Mean Square due to Error (*MSE*), which measures the variation within the groups (Sharpe et al, 2012). The F-statistic is given by:

$$F_{k-1,N-k} = \frac{MST}{MSE}$$

where the subindices in the F represent df. $k-1$ corresponds to the df of the *MST* with the k being the number of groups in the experiment. Additionally, $N-k$ is the df of *MSE*, with the N in this case being the total number of observations (Sharpe et al, 2012). For the F-test the p-value of the statistics is found in the corresponding distribution that has both df. The null hypothesis is rejected when the value is smaller than the significance level, this analysis is called an Analysis of Variance (ANOVA) (Sharpe et al, 2012). The *MST* is obtained as follows:

$$MST = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2}{k-1}$$

where $\bar{y}_i$ is the mean of the group *i*, n_i is the number of observations in group *i* and $\bar{\bar{y}}$ is the overall mean of all observations (Sharpe et al, 2012). Furthermore, MSE is calculated with:

$$MSE = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N-k}$$

where s_1 is the sample variance of group *i*. The two vales, *MSE* and *MST*, estimate the same variance if the null hypothesis is true. If it is false, the *MST* has a higher value (Sharpe et al, 2012). For small values of the F-statistic the probability of rejecting the null hypothesis is also small.

The assumptions and conditions for ANOVA are almost the same as in the two sample T-test that is described in the next section. The first one is the independence assumption. This assumption is verified when it is possible to explain that the groups and the observations are independent from each other (Sharpe et al, 2012). The probability that the groups are independent from each other increases if the observations in each group are randomly selected (Sharpe et al, 2012). The second assumption is the condition of equal variance. ANOVA assumes that the true variances of the groups are equal (Sharpe et al, 2012). Finally, the third assumption is the normal population assumption, which means that it is reasonable to think that the population distributions of each group are normally distributed (Sharpe et al, 2012).

3.3 Two sample Pooled T-test

The two sample Pooled T-test is used when the samples have different but small sizes and are completely independent of each other. The objective of this model is to compare two means of two different samples that must be independent from each other. The Pooled T-test compares

the ratio of the difference in the means from the samples to its standard error with the critical value from a student's t-model (Sharpe et al, 2012). The hypotheses for this test could be, for example:

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

where μ_1 and μ_2 represent the mean of the population of group one and two respectively (Sharpe et al, 2012). The T-statistic uses the observed means that are represented by $\bar{y}_1$ and $\bar{y}_2$, and is given by:

$$t_{df} = \frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y}_1 - \bar{y}_2)}$$

where SE is the standard error. For this case of Pooled T-test there exists the condition that the variances of the groups must be equal, for that reason the SE is as follows (Sharpe et al, 2012):

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{pooled}}{n_1} + \frac{s^2_{pooled}}{n_2}}$$

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

The values of $s_1$and s_2 represent the sample estimation of the standard deviation. They are calculated using the same formula of the standard deviation (Sharpe et al, 2012). s_{pooled} represents the sample standard deviation, taking both groups as one bigger group. The values of n_1 and n_2 represent the sample size of each group (Sharpe et al, 2012). Finally, the df for this statistic is calculated with $df = (n_1 - 1) + (n_2 - 1)$.

Before performing the two-sample T-test, the assumptions and conditions must be assessed. The first assumption is the independence assumption, it states that the data in each group must be drawn independently and obtained randomly from each population (Sharpe et al, 2012). The second condition is that the groups must follow a nearly normal distribution (Sharpe et al, 2012). The third condition is that the members of each group must be independent, and one member cannot affect other members of the other group (Sharpe et al, 2012). Finally, the fourth assumption is the equal variance between the groups that are analyzed (Sharpe et al, 2012).

3.4 Bonferroni method

As more tests are made to look for the difference between groups, the probability to make a Type I error grows, that means that the probability to reject the null hypothesis by mistake is bigger (Sharpe et all, 2012). To avoid this problem, it is possible to use the so-called multiple comparison methods. These methods require that, first, the null hypothesis of the ANOVA test is rejected, and then, the pairwise comparison of groups is made (Sharpe et al, 2012). One of

the multiple comparison methods is called Bonferroni method. This method adjusts the p-value of the test and its confidence interval to have a wider margin of error (Sharpe et all, 2012).

The Bonferroni method divides the error rate among the "r" confidence intervals (number of tests that are made), finding each interval at confidence level of $(1-\frac{\alpha}{r})$ instead of the original $(1-\alpha)$ (Sharpe et al, 2012). This means that for each of the *"r"* simultaneous tests that are performed, the significance level to compare the p-value changes to α/r (Hay-Jahans, 2019). Alternatively, each adjusted p-value for each test is obtained by multiplying the number of tests *"r"* by the p-value of the test and then compare it against the joint significance level α. In the cases where $(r * P-value) > 1$, the method suggests to assign the value of 1 adjusted p-value (Hay-Jahans, 2019).

3.5 QQ-Normal plots

QQ-Normal plots are also known as normal probability plots. They are used to test the normality assumption of one set of data. The process involved in the construction of this graph is the following. First, the set of data is sorted from the smallest to the largest value $\{X_{(1)}, X_{(2)}, \dots, X_{(3)}\}$. Each value $X_{(i)}$is called "the order statistics" of the sample (Fox, 2016) that corresponds to the sorted *ith* observation in the data set. This set is also called "sample quantiles" or "observed quantiles" (Hay-Jahans, 2019). Next, the plotting (or probability) points are calculated, using one of the next formulas according to the condition:

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

where n is the sample size (Hay-Jahans, 2019). After obtaining the probability points for each observation, the theorical quantiles (z_i) are calculated using the inverse of the Cumulative Distribution Function $P(x)$, that in this case is the normal distribution function (Fox, 2016).

$$z_i = P^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$$

Now, the ordered pairs $(X_{(i)}, z_i)$ are plotted and can be seen as points in the QQ-Normal plots (Fox, 2016). Finally, for comparison the reference line $X_{(i)} = \bar{y} + sz_i$ is plotted, where $\bar{y}$ is the sample median value and s is the sample standard deviation (Fox, 2016).

If the distribution of the data is normally distributed, the plot shows that the data points fit around the comparison line that are used as a reference (Sharpe et al, 2012). If there are strong deviations from the straight line, it indicates that the data is not normally distributed.

3.6 Software tools

The Software used for this report is R in the version R 4.0.5 GUI 1.74 for MAC iOS.

4. Statistical analysis

The statistical methods explained above are used to evaluate the similarities in the *height* of the players between six German national teams that are Basketball, Handball, Ice Hockey, Soccer, Volleyball, and Water Polo. First, to have a better understanding of the variables, a summary table and the boxplot of the distribution of each sport is presented. Then, the assumptions and conditions of the F-test and Pooled T-test are evaluated. After confirming that these assumptions and conditions are fulfilled, the hypothesis that the mean *height* of the players of the six national teams are equal is evaluated using the ANOVA test. Finally, to have a specific result of the pairwise relations between the *heights* of the six sports the Pooled T-test is applied.

4.1 Summary of the data set

The data set that is analyzed contains the information of the *heights* of the players of six German national teams. Table 1 displays the main information of the six sports presented in the data set which are Basketball, Handball, Ice Hockey, Soccer, Volleyball and Water Polo. Measures of dispersion like standard deviation and the minimum (min) and maximum (max) value of each distribution is presented. Also, some measures of central tendency like the median and the mean are given in the table. Furthermore, the multi-boxplot depicted in Figure 1 allows to compare the behavior of the *height* of the players within national teams and between them in a graphical way. Each box plot represents the distribution of the *heights* of the players in a specific sport.

Table 1. Summary information of the distribution by sport.

	Basketball	**Handball**	**Ice Hockey**	**Soccer**	**Volleyball**	**Water Polo**
Observations	12	21	25	23	15	16
Min [cm]	188.00	178.00	170.00	175.00	190.00	180.00
Median [cm]	203.50	194.00	183.00	183.00	199.00	192.00
Mean [cm]	200.67	193.81	183.28	185.00	200.20	192.81
Max [cm]	211.00	207.00	194.00	195.00	211.00	203.00
Standard deviation [cm]	8.27	6.00	5.35	6.27	8.19	7.31
Total Range [cm]	23.00	29.00	24.00	20.00	21.00	23.00

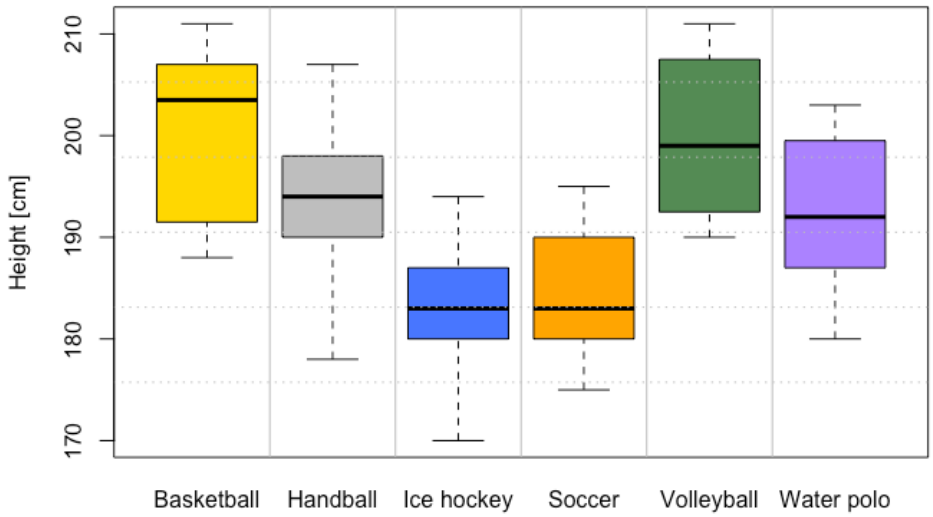

Figure 1. Comparison of the *heights* of the players between the sports.

Figure 1 and Table 1 show that the variability of the *height* of the players, as measured by the total range and the standard deviation, presents a similar behavior. Also, it is possible to see that the difference of the absolute values of the standard deviation are relatively small (maximum 2.92 cm) and the total range of the box plots are similar. Figure 1 also illustrates the differences of the median value of the *heights* of the players between each national team. Some sports present a similar value of the median values, whereas others present a significant difference. For this kind of symmetric distribution, it is possible to say that the behavior of the mean values of the population parameters is comparable to the behavior of the medians. To evaluate the last conclusion in a more formal way, the next sections will formulate the respective hypothesis tests of the comparison of the mean values of the *height* of the players between the sports presented in the data set.

4.2 Comparison of multiple distributions

The analysis in this section is aimed at testing the relation of the *heights* of the players of the six German national teams which is described in the previous section. First, the overall difference between the six teams is tested using the F-test and then, the pairwise differences between the *heights* of the players are evaluated using the two sample Pooled T-test. As is described in the method section both tests have assumptions that must be evaluated and confirmed before performing the tests.

The assumptions and conditions for both tests are, first, the independence assumption. The second condition is to have a nearly normal distribution of the groups. The third condition is

that the members of each group must be independent and, finally, the fourth assumption is the equal variance between the groups. Each assumption is analyzed in the following steps.

Regarding the independence assumption, it can be said that no national team of one sport affects the *heights* of the players of a national team of another sport. For that reason, the independence assumption is confirmed. In addition, regarding the Independent Groups Assumption, the *height* of one player is independent of the *height* of another player in another national team since there are no genetic relations between the players. The process of selection of players in each national team is, for this analysis, considered as a random process in terms of the *height* of the players. The normal distribution of the population of each group is assessed using the QQ-Normal plot described in section 3.5 and the multi-boxplot illustrated in Figure 1.

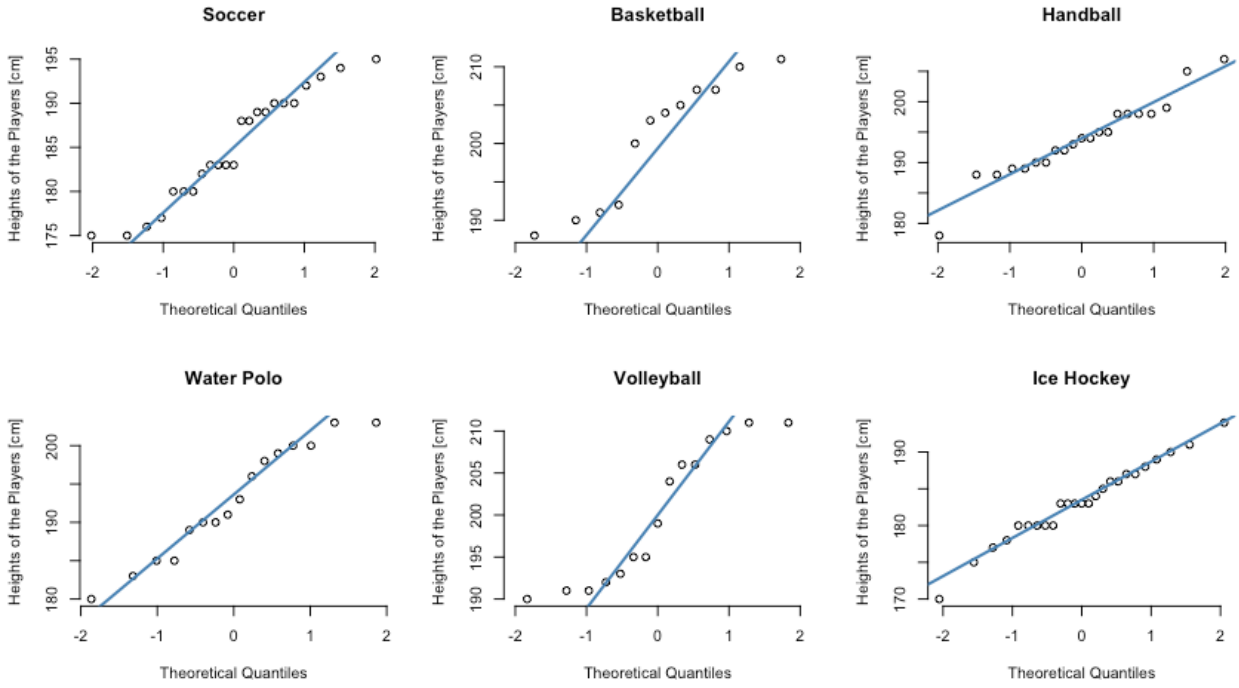

Figure 2. *Normal probability plot of each German national sport team.*

The diagrams in Figure 2 illustrate the Normal probability plot for each of the six German national teams. In all of them, the points fit around the reference line, which hence leads to the conclusion that the population of each sport follows a normal distribution. It is important to point out that for Basketball the values present an unclear behavior around the reference line but, in this case, it could be affected because of the small number of observations presented in the sample group. The conclusion of Normal populations is also confirmed by the boxplots presented in Figure 1. All of them show that the distributions are unimodal, symmetric and without outliers which could affect the assumption of normality.

Finally, the assumption of equal variance is evaluated using the values of standard deviation and total range that are obtained from Table 1. Most of the values of the standard deviation and the total range are close to each other. The values of the standard deviation are between 5.35 cm and 8.27 cm, which is a difference of 2.92 cm. For this case we can say that is a low difference. The same happens with the values of the total range, the largest difference is 9.00 cm. Even though the absolute values for the standard deviation and total range are not the same, the condition of equal variance between the groups is taken as fulfilled because the variation is not high.

4.2.1. Global test

The global test is used to compare one variable, in this case the variable *height* of the players of the six German national teams, between many groups. The test verifies if the mean value of the *height* of the players differs between the six sports. To carry out this evaluation, the ANOVA F-test is used. The following are the null and the alternative hypothesis that will be tested:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \quad H_A: at\ least\ one\ mean\ is\ different$$

where the value of μ_1 represents the mean *height* of the soccer team, μ_2 the mean *height* of the basketball team, μ_3 the mean *height* of the handball team, μ_4 the mean *height* of the water polo team, μ_5 the mean *height* of the volleyball team, and μ_6 the mean *height* of the ice hockey team. Now, after verifying the assumptions and conditions of the F-test, as was showed at the beginning of section 4.2, the F-test is performed giving the following results:

Table 2. *ANOVA Table for the global F-test.*

	Df	**MS**	**F value**	**P-value**
Sport	5	985.20	21.57	7.75 e-15
Error	106	45.70		

In Table 2, "MS" means Mean of Squares. From the p-value of Table 2 (7.75 e-15) it is possible to conclude that the null hypothesis is strongly rejected because the p-value has a small value, smaller than the significance level 0.05. Thus, it is possible to conclude that the mean value of the *height* of the players differs between the six German national teams. This conclusion can also be confirmed for the sample observations in Table 1, where the value of the sample means between the groups differs between the national teams.

4.2.2. Pairwise differences between the *heights* of the players

Using the two sample Pooled T-test, the difference in the means between pairs of national teams is tested. This test gives information about the sports in which the players have a similar *height*. First, the null and the alternative hypothesis are defined. To facilitate the comprehension of this hypotheses the subindices of the means will change to labels with the first to characters of the *sport name*.

Table 3. *Pairwise hypotheses of the national teams.*

No.	H_0:	H_A:
1	$\mu_{ba} = \mu_{ha}$	$\mu_{ba} \neq \mu_{ha}$
2	$\mu_{ba} = \mu_{ic}$	$\mu_{ba} \neq \mu_{ic}$
3	$\mu_{ha} = \mu_{ic}$	$\mu_{ha} \neq \mu_{ic}$
4	$\mu_{ba} = \mu_{so}$	$\mu_{ba} \neq \mu_{so}$
5	$\mu_{ha} = \mu_{so}$	$\mu_{ha} \neq \mu_{so}$
6	$\mu_{ic} = \mu_{so}$	$\mu_{ic} \neq \mu_{so}$
7	$\mu_{ba} = \mu_{vo}$	$\mu_{ba} \neq \mu_{vo}$
8	$\mu_{ha} = \mu_{vo}$	$\mu_{ha} \neq \mu_{vo}$
9	$\mu_{ic} = \mu_{vo}$	$\mu_{ic} \neq \mu_{vo}$
10	$\mu_{so} = \mu_{vo}$	$\mu_{so} \neq \mu_{vo}$
11	$\mu_{ba} = \mu_{wa}$	$\mu_{ba} \neq \mu_{wa}$
12	$\mu_{ha} = \mu_{wa}$	$\mu_{ha} \neq \mu_{wa}$
13	$\mu_{ic} = \mu_{wa}$	$\mu_{ic} \neq \mu_{wa}$
14	$\mu_{so} = \mu_{wa}$	$\mu_{so} \neq \mu_{wa}$
15	$\mu_{vo} = \mu_{wa}$	$\mu_{vo} \neq \mu_{wa}$

To perform the test, it is important to point out that the main assumptions and conditions for the Pooled T-test are fulfilled. The condition of normality, similar variance, and randomization, as well as the assumptions of independence have been validated at the beginning of the section 4.2. Table 4 shows the results of the Pooled T-test. The first column displays the number of the test according to Table 3, the second and the third column the *names of the sports* that are compared. The fourth, fifth and sixth columns present information of the test, such as the df, the p-value, and the adjusted p-value with the Bonferroni method. Finally, the last column presents the significance of the result, where “Nsig” means not significant, and “Sig” means significant.

Table 4. *Pooled T-test between the six German national teams.*

No.	Group 1	Group 2	df	P-value	Adj. P-value	Signif.
1	Basketball	Handball	31	6.01 e-3	9.01 e-2	Nsig
2	Basketball	Ice Hokey	35	4.89 e-11	7.33 e-10	Sig
3	Handball	Ice Hokey	44	7.44 e-7	1.12 e-5	Sig
4	Basketball	Soccer	33	2.57 e-9	3.86 e-8	Sig
5	Handball	Soccer	42	3.55 e-5	5.32 e-4	Sig
6	Ice Hokey	Soccer	46	3.80 e-1	1.00 e+0	Nsig
7	Basketball	Volleyball	25	8.59 e-1	1.00 e+0	Nsig
8	Handball	Volleyball	34	6.13 e-3	9.19 e-2	Nsig
9	Ice Hokey	Volleyball	38	8.98 e-12	1.35 e-10	Sig
10	Soccer	Volleyball	36	7.15 e-10	1.07 e-8	Sig
11	Basketball	Water Polo	26	2.95 e-3	4.43 e-2	Sig
12	Handball	Water Polo	35	6.58 e-1	1.00 e+0	Nsig
13	Ice Hokey	Water Polo	39	2.53 e-5	3.80 e-4	Sig
14	Soccer	Water Polo	37	5.74 e-4	8.61 e-3	Sig
15	Volleyball	Water Polo	29	2.97 e-3	4.45 e-2	Sig

Table 4 gives evidence to conclude in which national teams the mean value of the *heights* of the players could be equal or in which they differ. By analyzing the adjusted p-values from Table 4, it can be concluded that the null hypotheses of the pairs of sports Basketball and Handball, Basketball and Volleyball, Handball and Water Polo and Handball and Volleyball cannot be rejected because they are not significant. That means that there is not sufficient statistical evidence to conclude that the population mean of the *height* of the players of these German national teams differ. Furthermore, there is a similar situation for the pair Ice Hockey and Soccer, where the null hypothesis is not rejected and the means of the *heights* for the players of those national could be equal to each other. For the other pairs it is possible to conclude that the means values of the population mean for the *heights* differ between them.

By comparing the adjusted p-value and the p-value, it can be seen that most of the tests that are not significant, are still not significant after the adjustment. In addition, for other relations like the relation of Basketball and Water Polo or Volleyball and Water Polo the adjustment degrade the significance and put them near to the non-rection threshold.

5. Summary

The analysis of the relation of the *height* of the players between the different German national teams gives useful information for many fields like, for example, the performance advantages of players, the production of clothes, the design of sports equipment, as well as for the research

of medical, physiological, or nutritional treatment for groups of players of different teams. The reason is that having significant relations of the *heights* of the players gives insights regarding the possibility of standardization of treatments or of clothes and equipment production processes for different sports disciplines.

The data set used in this report contains the information of the *heights* of the players of six German men's national teams. The sample size amounts to 112 observations that include the *name* and the *heights* in centimeters of the players of six German men's national teams. There are 12 players for Basketball, 21 for Handball, 25 for Ice Hockey, 23 for Soccer, 15 for Volleyball, and 16 for Water Polo.

To understand the distribution of the *height* of the players for the six teams, a boxplot for each national team is created. To know if there is a difference between the mean population *heights* of the players of the six sport disciplines, a F-test also known as analysis of variance (ANOVA) is done. In addition, to find pairwise differences or similarities between the six national teams, the two sample Pooled T-test is performed.

This report shows that there is sufficient statistical evidence to conclude that there is a clear difference between the population *height* mean values of the players in the six national team that have been analyzed. However, analyzing the sports by pairs with a Pooled T-test with the null hypothesis that the population means of pairs constituted by two of the six German national teams are equal. It is possible to see that the null hypothesis between the pairs Basketball and Handball, Basketball and Volleyball, Handball and Water Polo, Handball and Volleyball as well as between Ice Hockey and Soccer is not rejected, that means that is not possible to conclude that the means are different. In addition, the Pooled T-test also gave sufficient statistical evidence to reject the null hypothesis of the rest of pairs in the possible combinations of the six German national teams. Even though there are limitations because of the small sample size, the use of statistical tools such as Pooled standard deviation or Bonferroni method help to find significant relations between the groups.

For further analysis, it is recommended to add more data either from players of previous years in every national team or from players from other countries. Furthermore, finding relations to other variables like the number of goals, points, or scores that each player has made during his career, his age, or the performance in the team could lead to more insights.

Bibliography

Fox, J. (2016). Applied Regression Analysis and Generalized Linear Models. United States of America, SAGE Publications, Inc.

Hay-Jahans, C. (2019). *R Companion to Elementary Applied Statistics*. United States of America, Taylor & Francis Group.

Samaras, T. (2007). *Human body size and the laws of scaling*. New York: Nova Science. pp. 33–61. ISBN 978-1-60021-408-0.

Sharpe, N. R., De Veaux, R. D. & Velleman P. F. (2012). *Business Statistics, 2nd Edition*. United States of America, Pearson Education, inc.

Mood, A. M., Graybill, F. A., Boes, D. C. (1974). *Introduction to the Theory of Statistics*. United States of America, McGraw-Hill.