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: Alp Yalçın

Group number: 5

Group members: Ezgi Osmanoğlu, Elif Aykut, Yazan Alhalabi, Marges Pinderi

June 11, 2021

Contents

1 Introduction

In Germany, following sports competitions and doing these sports are extremely significant for people and become a culture for the society especially 6 sports which are basketball, handball, ice hockey, soccer, volleyball and water polo. It is obvious that national competitions of these sports and national player's physical properties arouse interest to followers and analysts. Especially the heights of the players are one of the most exciting features of these physical properties.

The main aim of this report is compare the heights of the players between these six sports. The report is focused on to understand whether the players of different sports has same heights or they differ. Furthermore, the report is aimed to compare pairwise differences between the heights of the players for each six sports. It is focused on to get some connections between heights for each pairs of these sports.

The second part of the report, data set and data quality are clarified. Additionally, objectives of the project are explained. In third part, statistical methods are explained in detail and presented. All of the tests and assumptions which are used in the statistical analysis part are demonstrated. In the fourth section of the report, statistical analysis which are explained in the previous part are performed. Boxplots, QQ Plots, grouped statistical summary are used to understand the dataset, observe the relations of the variables and also for checking whether the dataset is compatible with the assumptions of the tests which are used for making inferences. One-way ANOVA global test is used to understand whether the heights of the players differ between six sports. Additionally, pairwise t-test is used to analyse whether pairwise differences occurs between the heights of the players. Furthermore, Bonferroni method is used to adjust the t-test result to address the multiple testing problem.

2 Problem statement

2.1 Data set and data quality

This report is compiled from the dataset which is provided by TU Dortmund University Introductory to Case Studies course lecturers.

The data set comprises the heights of the players of six German men's national teams which are volleyball, handball, soccer, basketball, ice hockey and water polo. The height which is measured in centimeters for each player of each team is given in the data set.

The data set include 112 observations of 3 variables. It has one character variable which includes the full names of the players and named as "Name". Also it has one integer variable which shows the weights of the players and called as "Height". Last variable of the data set is factor variable with 6 levels which shows the sport categories of the players and named as "Factor". These 6 levels are volleyball (15 observations), handball (21 observations), soccer (23 observations), basketball (12 observations), ice hockey (25 observations) and water polo (16 observations).

Overall, data set is clean and easy to be worked on it. However, name variable observations of the of ice hockey group is totally missing. It does not have any effect on this reports which focus on comparison of the heights of the players because only the names are missing for ice hockey.

2.2 Project objectives

The main objective of the project is comparing of multiple distributions. Therefore, firstly, dataset is separated with 6 groups to sports. Then, grouped statistical summary is conducted. The box plots and q-q plots are used to compare the heights differences between the six sports. ANOVA test is conducted to see the result. Also pairwise differences between the heights of the players are examined. All pairs of sports are considered and two sample t-test is conducted. Additionally, in order to address the multiple testing problem, Bonferroni adjustment method is used. Last but not least, these two results with and without adjusting of multiple testing are compared.

3 Statistical methods

For all analyses and calculation, the software (R Core Team, 2021) with packages **dplyr** (Wickham et al., 2021), **formattable** (Ren and Russell, 2021), **gridExtra** (Auguie, 2017), **ggplot** (Wickham, 2016), **xtable** (Dahl et al., 2019)

3.1 Hypothesis Testing

Hypothesis testing is a way to test the results of the specific hypothesis to see whether there are meaningful results.

Statistical test is a method which enables a decision for decision for accepting or rejecting a hypothesis about the not known parameter to happen in the distribution of a random variable

(Rasch et al., 2019, p. 39)

It can be supposed that two hypothesis are possible, null hypothesis H_0 and the alternative hypothesis H_A .Null hypothesis is a statement of there is no disparity between a sample mean and a population mean. Alternative hypothesis is an assertion about the population is incompatible to null hypothesis and it is concluded when null hypothesis is rejected. Therefore, if null hypothesis is right, the alternative hypothesis is wrong, and vice versa. Hypothesis may be composite or simple.

(Rasch et al., 2019, p. 39)

Hypothesis testing is approached based on random samples. Test statistic, which is a specific random variable, is derived from random sample. Null hypothesis is rejected if the eventuation of the test statistic has some connection to a real number, let assume it exceeds some quantile. This quantile is selected so that the probability which the random test statistic surpasses it if the null hypothesis is correct is equal to value $0 < \alpha < 1$; this is fixed beforehand and it is called significance level. Significance level equals the probability to make an error of the first kind, for example not to accept the null hypothesis if it is correct. Generally, significance level is set to 0.05 or 0.01 and denoted by α.

(Rasch et al., 2019, p. 39)

At the end of the classical statistical test we determine one of two possibilities, which are accepting the null hypothesis or rejecting the null hypothesis. Tests for a given significance level α are called α-tests; generally, there are many α-tests for a given pair of hypothesis.

(Rasch et al., 2019, p. 39)

When wrong assumptions made about the hypothesis, error types arise. Type 1 error is rejecting the null hypothesis where it is true, and Type 2 error is accepting the null hypothesis where it has to be rejected. (Rasch et al., 2019, p. 39)

3.2 QQ-Plots

Normal probability QQ-plots procure a means for comparing the distribution of a sample against the standard normal distribution, the related function is

qqnorm(y, main = "Normal Q-Q Plot", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",...)

Where y typifies the sample, which should be passed into function call. These other arguments have default values which may be switched.

(Hay-Jahans, 2019, p. 146)

Alternatively, a sample's distributional specifications can be compared with any other theoretical distribution with

qqplot(x, y, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)) ...)

Where y typifies the sample quantiles and x represents the quantiles which is used for analogy. These x and y are essential, other arguments are optional.

(Hay-Jahans, 2019, p. 146)

Normal Probability QQ-Plots

The following part addresses the continuum of the involved in the structure of a normal probability QQ-plot.

Sample of interest is denoted by $(y_1, y_2, y_3, ..y_n)$, and start by sorting this sample in ascending order. Sorted data is denoted by $(y_{(1)}, y_{(2)}, y_{(3)}, ..y_{(n)})$ and referred as the *sample* or *observed quantiles*. Then, calculate what are referred to as *probability points*, p_i.

(Hay-Jahans, 2019, p. 147)

The probability points, p_i, are used to calculate theoretical quantiles, x_i, corresponding to each concatenated sample quantile, $y_{(i)}$. For normal probability QQ-plots this involves finding x_i for $i = 1, 2, ..., n$ such that $\mathrm{P}(\mathrm{X} \leq x_i) = p_i$, where $\mathrm{X} \sim \mathrm{N}(0,1)$. The x_i are referred to as the theoretical, or expected quantiles.

(Hay-Jahans, 2019, p. 147)

3.3 The Two-Sample t-Test

Two sample t-Test analyses the given null hypothesis for the case that the variance σ_1^2 and σ_2^2, in turn, of the relevant variables in the two populations which are unknown; this is the ordinary case. However, it supposes that these variances are equal in both populations.

$$t = \frac{\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2}{\sqrt{\frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1+n_2}}$$

(Rasch et al., 2019, p. 63)

The test statistic is t-distributed with $n_1 + n_2 - 2$ degrees of freedom. It can be determined the null hypothesis by using the $(1 - \alpha)$ - quantile of the (central) t-distribution t($n_1 + n_2 - 2$, $1 - \alpha$) and t ($n_1 + n_2 - 2$, $1 - \alpha/2$).

H_0: $\mu_1 = \mu_2 = \mu$; will be rejected if:

(a) $t<t(n_1 +n_2 -2,\alpha)$

(b) $t>t(n_1 +n_2 -2,1\text{-}\alpha)$

(c)$|\ t\ |>t\ (\ n_1+n_2-2,1-\alpha/2)$

If the null hypothesis is not accepted, the two expectations differ considerably.

(Rasch et al., 2019, p. 63)

3.4 ANOVA

ANOVA, abridgment of Analysis of Variance, is a statistical test which used to analyse the difference between the means of multiple groups. There are two types of ANOVA, one-way ANOVA which uses one explanatory variable and two-way ANOVA which uses two explanatory variables.

It is assumed that the error terms, ϵ_{ij} , are independently and identically distributed with $\epsilon_{ij} \sim$ N (0, σ) where σ^2 symbolizes the (unknown and common) population variance.

(Hay-Jahans, 2019, p. 271)

The null hypothesis of ANOVA is that there is no difference among group means. The alternative hypothesis is that at least one group differs considerably from the total mean of the response variable.

H_0 :$\mu_1 = \mu_2 = \cdot\ \cdot\ \cdot = \mu_p$

H_1 : The means are not all equal.

Let p is the population, for $j = 1, 2, ..., p$, sample mean for the j^{th} sample is denoted by y_j, sample size from the j^{th} population is denoted by n_j. Overall mean of the all observed responses are denoted by $\bar{y}$, and the j^{th} sample variance is denoted by ${s^2}_j$. Finally, the test statistic use the test equality of the p means is

$$F^* = \frac{\sum_{j=1}^{p} n_j(\bar{y_j} - \bar{y})^2/(p-1)}{\sum_{j=1}^{p} (n_j - 1)s_j^2/(n-p)}$$

and p-value = P(F $\geq$ F$*$).

(Hay-Jahans, 2019, p. 271)

3.5 Bonferroni's Procedure

ANOVA one-way assumptions are same for the Bonferrroni's procedure. The underlying random variables are independent and normally distributed with equal variances. The formula for the $m = p(p-1)/2$ simultaneous 100(1- α)% Bonferroni confidence intervals for the difference between two treatment is

$$(\bar{y_j} - \bar{y_k}) - t_{\alpha/(2m)} s\sqrt{\frac{1}{n_j} + \frac{1}{n_k}} < \mu_j - \mu_k < (\bar{y_j} - \bar{y_k}) + t_{\alpha/(2m)} s\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}$$

where s is the standard error and $t_{\frac{\alpha}{2m}}$ such that P (t$\geq t_{\frac{\alpha}{2m}}$) $= \alpha/(2m)$ with t $\sim$ t(n$-$p). Two population means, μ_j and μ_k, are evaluated divergent at the individual α/m level of significance if the related confidence interval for $\mu_j - \mu_k$ does not contain zero.

(Hay-Jahans, 2019, p. 274)

The corresponding test statistic is

$$t_{jk}^* = \frac{\bar{y_j} - \bar{y_k}}{s\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}} \sim t(n-p)$$

two sided test, p-value = 2P(t >= $|t_{jk}^*|$) for each comparison. Every m simultaneous tests are used by comparing each p-value against the individual significance level α/m .

(Hay-Jahans, 2019, p. 274)

The pairwise.t.test function use the above described calculations, with the usage definitions with arguments

pairwise.t.test(x, g, p.adjust = "bonferroni")

where x is observed responses, g is allocated the factor including the treatment levels that describe the samples. **p.adjust = "bonferroni"** options are of the form m×p-value .

(Hay-Jahans, 2019, p. 274)

4 Statistical analysis

In this section, the statistical methods explained above are applied to the presented data set and the results are interpreted.

4.1 Comparison of the players heights between the sports

Heights of the players between the sports are compared with box plots, statistical summary table and q-q plots which are used to understand the homogeneity of the variance between groups and whether the data has normally-distributed or not. Our null hypothesis says there is no difference between the heights of the players between the sports.

Box plots which is grouped by the sports is used for the comparison of the player heights between the sports. Volleyball players height is in the range of 190 cm and 206 cm, handball players range is between 178 cm and 207 cm, soccer range is 175 cm and 190 cm, basketball is 188 and 211, ice hockey players are in the range of 175 and 194 and water polo players are in the range of 180 cm and 203 cm. Also, IQR of the dataset for volleyball and basketball groups are wider than the others.

Additionally, grouped statistical summary table is used to compare the standard deviations of heights. Basketball has the highest standard deviations among the sports with 8.27 cm. Basketball also has the highest mean and median of height variable between the sports. On the other hand, ice hockey has the lowest standard deviation with 5.35 cm.

As a consequence, when we look at the box plot tables and grouped statistical summary table, it is obvious to see that all sports have different standard deviations with different ranges. However, it is considered with the equal variances in this report to make test with ANOVA global test.

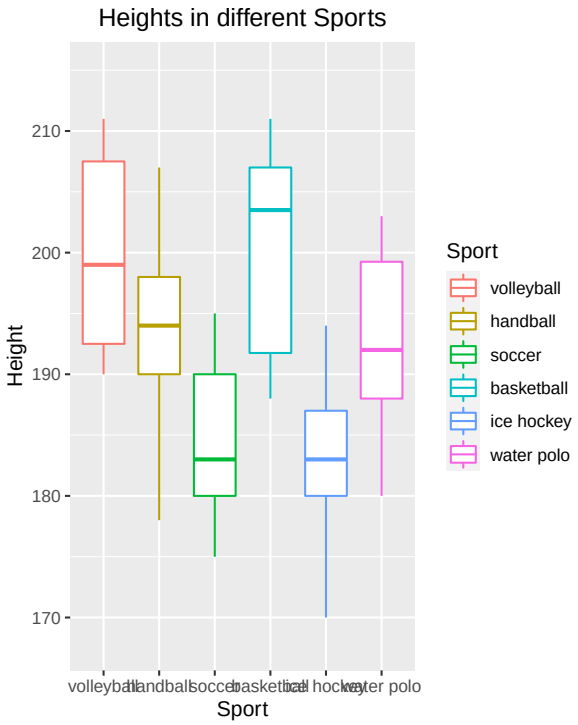

Figure 1: Box Plot

	Sport	n	mean (Height)	Median (Height)	SD (Height)
1	volleyball	15	200.20	199.00	8.18
2	handball	21	193.81	194.00	6.32
3	soccer	23	185.00	183.00	6.27
4	basketball	12	200.67	203.50	8.27
5	ice hockey	25	183.28	183.00	5.35
6	water polo	16	192.81	192.00	7.31

Q-Q plots are used for each sports to observe whether the height variable has normally distributed or not. When we look at each figure, it is crystal clear that height variable is not normally distributed. For volleyball and basketball, data points are digressed from the line with "s shaped" whereas soccer and water polo differentiated from the line with similar shape. Ice hockey is the only sport group which can be assumed normally distributed. To sum up, groups are not normally distributed. However, groups are assumed to be normally distributed to make ANOVA global test.

ANOVA (Global test) The ANOVA output provides an estimate of how much variation in the dependent variable that can be explained by the independent variable. Sum of squares between the group is means and overall mean according to height is 4926.1 cm and mean of the squares is 985.23 cm which is computed with dividing sum of squares by

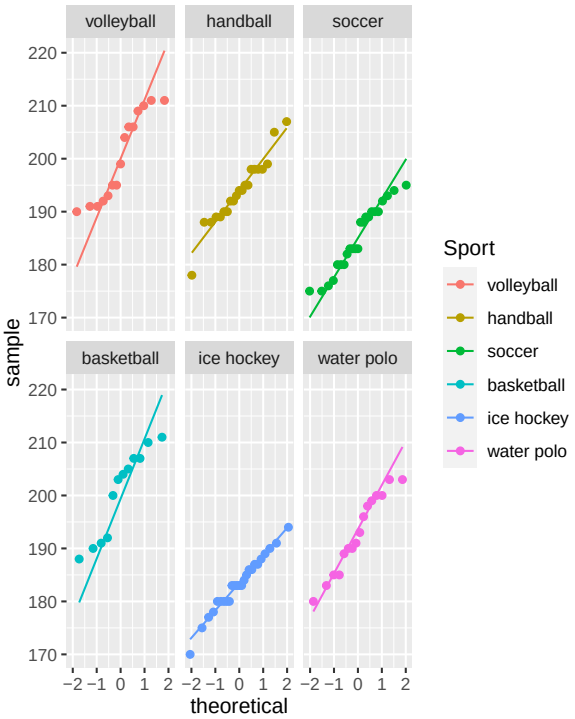

Figure 2: QQ Plot

degrees of freedom. Test statistics from the F-test which is the mean square of variable divided by the mean of squares is 21.57. P-value of the F-statistic from the F test is 7.755e-15 which significance level is high. So we reject our null hypothesis according to ANOVA test. To sum up, our global test indicates that there is a difference between the heights of the players.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sport	5	4926.14	985.23	21.57	0.0000
Residuals	106	4841.78	45.68		

4.2 Pairwise differences heights of the players

This part of the report, pairwise differences heights of the players is examined with two sample t-test and after Bonferroni method is used to make adjustment of the result in order to address the multiple testing problem. All of the 15 pairwise combinations of the sports are tested.

Pairwise comparisons using t tests with pooled SD

	volleyball	handball	soccer	basketball	ice hockey
handball	0.00613	-	-	-	-
soccer	7.2e-10	3.5e-05	-	-	-
basketball	0.85884	0.00601	2.6e-09	-	-
ice hockey	9.0e-12	7.4e-07	0.38040	4.9e-11	-
water polo	0.00297	0.65755	0.00057	0.00295	2.5e-05

When the two sample t-test is used, 8 pairs of the sports are highly significant which p-values are smaller than 0.001, 3 of the pairs are significant which p-values are smaller than 0.01 and 5 of the pairs are not significant which p-values are higher than 0.05. These pairs are Basketball-Volleyball, Ice hockey- Soccer and Water Polo-Handball and with the p values of 0.85, 0.38 and 0.65. Therefore, null hypothesis for 12 pairs are rejected and these pairs are not same heights whereas only 3 pairs have the same heights with two sample t-test.

However, there is a crucial fact for this function that we do not compute the standard deviations for each groups individually, we consider as same. Thus, in this method, a single standard deviation is calculated for all groups and used for the t-Test.

Pairwise comparisons using Bonferroni method adjustment

	volleyball	handball	soccer	basketball	ice hockey
handball	0.09191	-	-	-	-
soccer	1.1e-08	0.00053	-	-	-
basketball	1.00000	0.09013	3.9e-08	-	-
ice hockey	1.3e-10	1.1e-05	1.00000	7.3e-10	-
water polo	0.04452	1.00000	0.00861	0.04428	0.00038

After making the Bonferroni adjustment, the new alpha becomes $0.05/15 = 0.0033$. 7 combinations of pairs are significant and less than alpha (0.0033) with Bonferroni adjustment whereas, p values of the other 8 pairs are not significant and hypothesis are accepted for these pairs. Thus, after Bonferroni adjustment, number of rejected hypothesis are decreased. 5 pairs of groups, which are, handball-volleyball, handball-basketball, water polo-basketball, water polo-volleyball, water polo- soccer, becomes not significant after Bonferroni adjustment.

For both tests, there are significant p values which are lower than the alpha and the null hypothesis can be rejected. To sum up, there are pairwise differences between the heights of the players.

5 Summary

The data set analysed in this report was constituted by the instructors of the course Case Studies I at TU Dortmund University in the summer term 2020/21 and it contains heights of the players of six German men's national teams basketball, handball, ice hockey, soccer, volleyball, and water polo. For each player of each team, the height (measured in centimeters) is given.

Firstly, structure of the data set examined and dataset was separated with 6 groups to sports. Then, box plots, qq plots and grouped summary were used to understand the characteristics of the heights between the six sports and also checked the assumptions of ANOVA test and Bonferroni adjustment method. Box plots and grouped summary were used to see the equal variance assumption and qq plots was used to see the normal distribution assumption.

Later, ANOVA test was applied to check whether the heights of the players differ between six sports and p value significance level showed high, and we rejected the null hypothesis. So, the result indicated that heights of the players are different from each other. That was the first task of the report.

In the next part of the report, we checked the pairwise differences between the heights of the players for each groups and we compared 15 pairwise combinations of sports with two sided t test. 8 pairs of the sports emerged significant and 12 pairs of the groups out of these 15 combinations showed that heights were different from each other. Additionally, we made Bonferroni adjustment to address the multiple testing problem and after the adjustment, 8 pairs of combinations were categorized that had not equal heights. For the other 7 pairs, we accepted the null hypothesis. 5 pairs of groups, which are, handball-volleyball, handball-basketball, water polo-basketball, water polo-volleyball, water polo-soccer, had become not significant after Bonferroni adjustment.

To sum up, after the checking both individual and pairwise combinations of the groups, although some pairs of the groups had equal heights, we concluded that sports had different heights from each other.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL `https://CRAN.R-project.org/package=gridExtra`. R package version 2.3.

David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-4.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Dieter Rasch, L.R. Verdooren, and Jürgen Pilz. *Applied Statistics - Theory and Problem Solutions with R*. 07 2019. ISBN 978-1-119-55152-2. doi: 10.1002/9781119551584.

Kun Ren and Kenton Russell. *formattable: Create 'Formattable' Data Structures*, 2021. URL `https://CRAN.R-project.org/package=formattable`. R package version 0.2.1.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.6.

Appendix

A Additional figures

B Additional tables