TU Dortmund

Introductory Case Studies

Project II: Comparison of multiple distributions

Author: Swapnil Srivastava

June 11, 2021

1 Introduction

While height doesn't necessarily reflect the ability of any person, because hard work and skill are the most important factors, height can influence what sports he can choose and where he can channel his efforts to get better. It therefore makes sense to study this difference in heights of players of various sport teams.

In this project, 6 German men's national team is analyzed with respect to the heights of each player. For this purpose, first we perform the descriptive analysis to understand the data set. We then run ANOVA test to understand whether or not there exists a significant difference among the means in the heights of the players of various sports. If yes, we then try to narrow down the pairs where there is a significant difference using pairwise testing. Furthermore, we apply Bonferroni corrections to our results in order to control the Type I error which may have appeared.

Apart from the Introductory section, the project report deals with additional 4 sections. The Section 2 deals with the description of the data set used in terms of the definitions of the variables and the quality of the data provided. The section 3 deals with the explanation of various statistical terms used for analyzing the data like QQ Plots, various tests used for hypothesis testing etc. Section 4 then deals with the interpretation of the analysis by using the statistical methods (Section 3) on the given data set. The last section i.e. Section 5 contains the summary which deals with all the interpretations and results, also providing an insight for future analysis.

2 Problem statement

2.1 Description of data set and quality

The data set was compiled by the instructors of the course *Introductory Case Studies* at TU Dortmund University (Summer Semester 21). It was collected by various sources like web, their personal contacts etc. The Data set contains the names and the heights (in *cm*) of the players of six German men's national teams. We see that the data type of columns *Name* and *Sport* is **character**, while for *Height* is **integer**. The Data set contains overall 112 players information, segregated as:

Table 1: Overview of data set

Sport	Number of observations
basketball	12
handball	21
ice hockey	25
soccer	23
volleyball	15
water polo	16
Total	**112**

While the height is available for every player, the names are *missing* for all the players of water polo team. Since, the names do *not* hold a significant value in our analysis, we will ignore these missing values for this project.

2.2 Project objectives

The main objective of the project is to perform hypothesis testing and interpret the results accordingly. For this, we are provided with the data set (as described above) dealing with heights of players of six German men's national teams. Our first objective is to find if there is at all any difference (in terms of mean heights) present among the various sports or not. If the difference exist, we need to then narrow down and find the pair-wise differences among the sport groups. Since we are performing multiple analysis on same variable i.e. Height, it will increase our chances of committing *Type I* error. Hence, to control this type of error, we further apply Bonferroni adjustment. We then again analyze the differences between heights of the groups using the new p-values which are obtained after the Bonferroni adjustments and there we get our final conclusions.

3 Statistical methods

The below mentioned statistical methods are used to evaluate the given data set, for the current project.

3.1 Graphs used in the Descriptive Statistics

QQ Plot: QQ plot or **Quantile-Quantile** plot is a graphical way based on analysis of the estimates of quantiles. The Normal probability QQ plot provides means for comparing the distribution of a sample against the standard normal distribution.

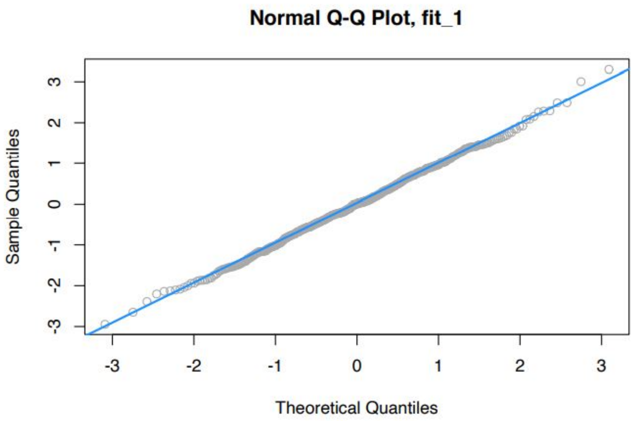

Figure 1: Normal Probability QQ Plot.(Alex Stepanov, David Unger, James Balamuta, 2021)

In the above figure 1, we can observe that since the plots of the plot closely follow the straight line, it would suggest that the data comes from a normal distribution. The calculations required to create the plot vary depending on the implementation, but essentially the y-axis is the sorted data (observed, or sample quantiles) and the x-axis is the values we would expect if the data did come from a normal distribution (theoretical quantiles). (Alex Stepanov, David Unger, James Balamuta, 2021)

3.2 Inferential statistics

Inferential Statistics can be defined as the part which makes out inferences or predictions from the data set using the descriptive statistics. The major inferential statistics used in this project can be summarized as below:

Hypothesis can be defined as *tentative explanations of a principle operating in nature.*(Charles Merriam, George Merriam, 1993)

3.2.1 Statistical hypothesis analysis

Null Hypothesis: The null hypothesis can be defined as the default status i.e. it is a type of hypothesis that signifies that there is *no difference* between a certain characteristics of a population. It is represented as : H_o

Alternative Hypothesis: As opposed to null hypothesis, alternative hypothesis signifies that *there is a significant difference* between the characteristics. If we get any significant evidence to *reject* the null hypothesis, alternative hypothesis comes into picture in such cases. It is represented as: H_1 or H_a

Rejection and Non Rejection Regions: When there is a study to be conducted, we divide the outcomes of that study as following:

- Those that cause the rejection of the null hypothesis. (Rejection Region)
- Those that do not cause the rejection of the null hypothesis. (Non - Rejection Region)

Type I and Type II Errors: A Type I error is committed by rejecting a true null hypothesis while Type II is committed when we fail to reject a false null hypothesis.

Level of significance: The level of significance (denoted by α) can be defined as the probability of committing a Type I Error. For instance, if $\alpha = 0.05$ it means that there is a risk of 5% of making the following inference: There exists a difference when there is actually *no* difference.

Degree of Freedom: Degree of freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample. It can be calculated as: $D_f = N - 1$ where N is the number of sample.

3.2.2 Statistical hypothesis testing methods

Using t-statistic to test hypothesis: An assumption underlying this technique is that the measurement or characteristic being studied is normally distributed for the populations and the variances are equal (${\sigma_1}^2 = {\sigma_2}^2$). The pairwise t-statistic can be calculated by using the following formula :

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{x_1{}^2(n_1-1)+x_2{}^2(n_2-1)+..+x_k{}^2(n_k-1)}{n_1+n_2+..+n_k-k}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+..+\frac{1}{n_k}}}$$

where n_k represents the number of observations in sample k, $\bar{x_1}$ and $\bar{x_2}$ are the means of first and second samples respectively. If this t-statistic value lies within the *non-rejection* region, we fail to reject the null hypothesis or else we reject the null hypothesis.(Black, 2010)

Using the p-value to test hypothesis: Another approach to test the hypothesis can be to calculate the p-value.It represents the probability of occurrence of the given event which is being taken into account. If the p-value is less than α, we get a strong evidence to reject the *null* hypothesis and if not, then we fail to reject the null hypothesis. We can calculate the p-value by using the t-distribution table. In this we need to find the value corresponding to our t-statistic value (as calculated above) and degree of freedom value. The analysis can be well understood as below:

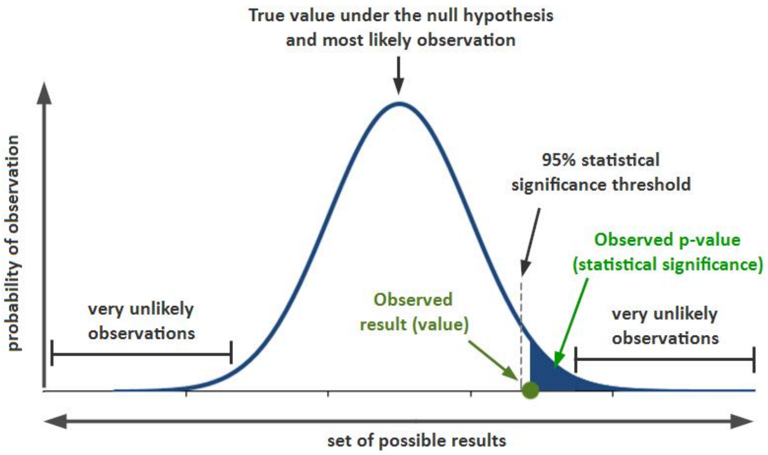

Figure 2: Demonstrating p-value analysis with $\alpha = 0.05$.(Web)

One way ANOVA testing: One way **analysis of variance** or ANOVA is a technique that analyzes all the sample means at one time. This means, if k samples are being

analyzed, the following hypothesis are being tested:
$H_o = \bar{x_1} = \bar{x_2} = ...\ \bar{x_k}$ where $\bar{x_k}$ represents the mean of k^{th} sample.
$H_a =$ Atleast one of the sample's mean is different from the rest.

We compute the following :
SSC (Sum of square) $= \sum_{j=1}^{k} n_j(\bar{x_j} - \bar{x})^2$
SSE (Sum of square of errors) $= \sum_{i=1}^{n_k} \sum_{j=1}^{C} (x_{ij} - \bar{x_j})^2$
MSC (Mean square of columns or groups) $= \frac{SSC}{df_k}$
MSE (Mean square of errors) $= \frac{SSE}{df_e}$
F - value $= \frac{MSC}{MSE}$

where i = particular member of the group or sample
j= group or sample level
k= number of groups or samples
n_j = number of observations in a group or sample
$\bar{x}$ = overall mean
x_j= a particular group mean
x_{ij}= individual value

As mentioned for t-value, similarly, we read the critical **F-value** from the F-distribution table and compare the F-value computed from ANOVA testing to it. This helps us to determine whether there is a significant difference between the groups or not. If the F-value computed by ANOVA is greater than the critical value obtained from F-distribution table, we reject the null hypothesis or else we fail to reject it. (Black, 2010)

Bonferroni Adjustments: When we conduct multiple analysis, it increases the chance of committing *Type I* error, in turn increasing the likelihood of getting the false significant result by chance. In order to control the possibilities of Type I error, we adjust our results using Bonferroni adjustment. It increases the p-value, hence making it less likely to commit such error. Even though, we are able to control Type I error using this, it creates more vulnerability to Type II errors. Bonferroni adjustment computes the new p-value as:

$$p_{adjusted} = p * n$$

where p is the raw p value and $p_{adjusted}$ is the p-value after Bonferroni correction and n is the number of tests performed. (Black, 2010)

4 Statistical analysis

All the following measures and plots were created using the software R in version 4.1.0. (R Development Core Team, 2021)

Description of data set: The Statistical methods described in Section 3 are implemented on the whole data set to get the better understanding of data. It can be summarized as below:

Table 2: The table describes the data set.

Sport	Min	Q1	Median	Mean	Q3	Max
volleyball	190.00	192.50	199.00	200.20	207.50	211.00
handball	178.00	190.00	194.00	193.81	198.00	207.00
soccer	175.00	180.00	183.00	185.00	190.00	195.00
basketball	188.00	191.75	203.50	200.67	207.00	211.00
ice hockey	170.00	180.00	183.00	183.28	187.00	194.00
water polo	180.00	188.00	192.00	192.81	199.25	203.00

In order to maintain homogeneity throughout, we would be considering all the data values up to 2 decimal places

Assumptions: In order to perform the tests on the data set provided to us, we would be checking for the underlying assumptions which are made by us. We would be performing *ANOVA* test and *Pairwise t test*, both of which are based on our following assumptions:

- **Assumption of independence** : Our tests require that all the observations are taken randomly and that such samples extracted from the population are independent of each other. Since the data set was collected by our professors of this course, we are assuming that this condition *holds* true.
- **Assumption of homogeneity of variance** : Our tests require that the variance of distributions in the *population* are equal. Since, the null hypothesis states that all observations come from same underlying group with same degree of variability, hence if we do not assume that all the variances of distributions are equal, our null hypothesis will be rejected before hand only. Hence for any population, the variances ${\sigma_1}^2$, ${\sigma_2}^2$, ${\sigma_3}^2$, ...,${\sigma_k}^2$ are assumed equal. For our current project, we need to check this variability for the data grouped by sport. A quick way to understand the variability is by using the box plot :

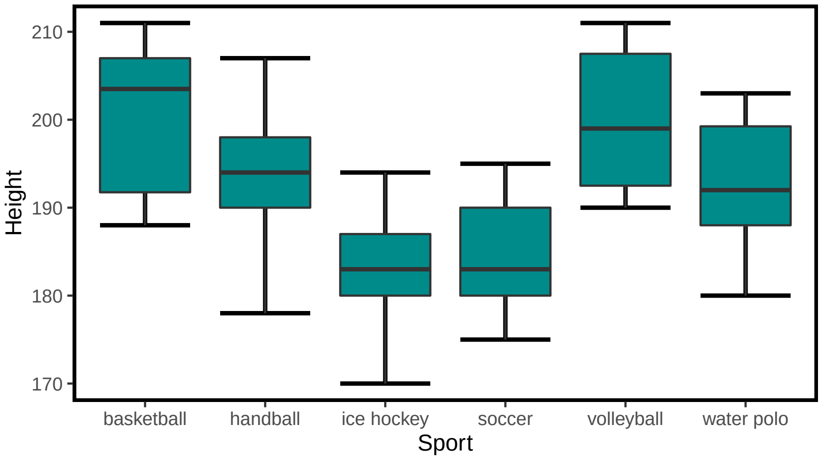

Figure 3: Understanding the variability of the given data set using box plot.

In the box plot, the variability within each group is represented by the vertical size of each box, i.e., the inter-quartile range. The variances of each group can be numerically compiled as :

Table 3: The variances of each group.

Sport	**Variance**
volleyball	66.89
handball	39.96
soccer	39.27
basketball	68.42
ice hockey	28.63
water polo	53.50

As from the above Table 3 and Figure 3, we can observe that the variances are not exactly equally within the group. But since the data which is provided to us is of a sample and not of whole population, we assume that this assumption holds true for the population (unlike what we can see for the sample data provided).

- **Assumption of normality** : Our tests require that the observations are drawn from normally distributed populations. A Quantile-Quantile (or QQ) plot can be used to study the distribution. For our project, the same can be visualized as :

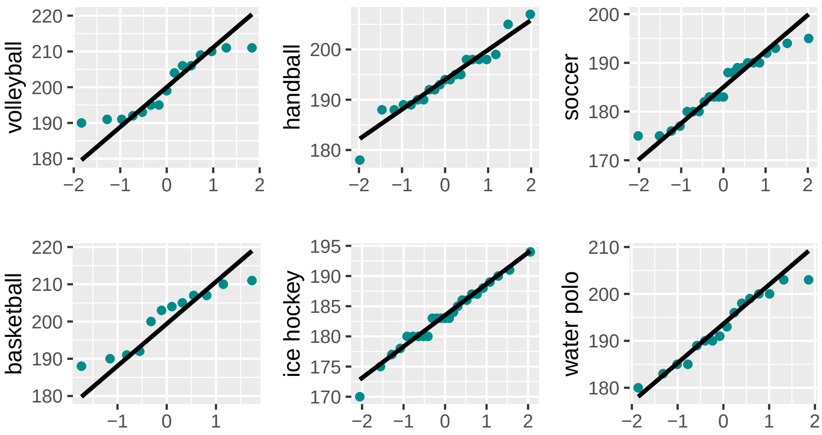

Figure 4: The study of underlying distribution using QQ plots.

In the above Figure 4, we analyze the distribution in each group against the normal distribution. Where the *green* dots represent the dependent variable in our data set, the black line represents the *normal* distribution. Except basketball and volleyball, we can see that for all the other groups, height data, almost *overlaps* the Normal distribution line. Hence, it is safe to conclude for such cases that the height is normally distributed. As stated above, the data is of a sample and not for a population. Hence, we will assume for the basketball and volleyball group that the data is normally distributed for them as well.

As all the three assumptions are met, we will now proceed to perform the tests on our data set.

4.1 Global test

In order to perform a global test to understand whether the mean height of the players differ among the six sports or not, we will first define our Null and Alternative Hypothesis. As stated in Section 3, the Null hypothesis is the default status i.e. it signifies that there is no difference while the Alternative hypothesis states the opposite. We take the value for level of significance (α) as 0.05 and the test which we will be performing is ANOVA Test in our case:

Null Hypothesis : H_o = There lies no significant difference between the mean heights of the players of the six sports.

Alternative Hypothesis : H_1 = There lies atleast one sport whose mean height is different from the rest.

Results : We get the following results upon performing ANOVA test on the given data set:

Table 4: ANOVA Test results

	Degree of Freedom	Sum of Square	Mean Square	F Value	P Value
Sports	5	4926.00	985.20	21.57	< 0.01
Residuals	106	4842.00	45.70		

From the above table we can see that the p-value is *less than* the level of significance i.e. the p-value is statistically significant. This implies a strong evidence *against* the null hypothesis, as there is *less than* 5 % probability that the null hypothesis is correct. Also since the F value ratio is very large (i.e. not close to the F-distribution table value) it also supports the fact that a large variation is present within the groups mean. Hence, after performing ANOVA Test we get enough evidences to *reject* our null hypothesis.

Conclusion : There is atleast one sport whose mean height of the players is different from the remaining sport.

4.2 Pairwise difference

As concluded using the global test, there is a significant difference between the mean heights of the players among the six sports. We still do not know exactly which group (or *groups*) is different from each other. In order to understand this pairwise difference, we will perform pairwise t test. Similar to ANOVA Test, we will first define our null and alternative Hypothesis.

Null hypothesis : H_o = There is no pairwise differences between the mean heights of the players of each sport.

Alternative hypothesis : H_1 = There is lies a significant pairwise differences between the mean heights of the players of each sport.

Results: We get the following results on performing pairwise t test on our data set:

Table 5: P-values using pairwise t test results

	basketball	handball	ice hockey	soccer	volleyball
handball	0.01	-	-	-	-
ice hockey	< 0.01	< 0.01	-	-	-
soccer	< 0.01	< 0.01	**0.38**	-	-
volleyball	**0.85**	0.01	< 0.01	< 0.01	-
water polo	0.01	**0.65**	< 0.01	< 0.01	0.01

Conclusion : As from the above table 5, we see that for the following pairs the p-value (the bold values) comes out to be above than the significant value (0.05):

- basketball - volleyball
- handball - water polo
- ice hockey - soccer

Hence, for the above pairs, we *fail to reject* the null hypothesis. This means that for the above pair of sports there does not exist a significant difference between the mean heights of the players.

On the other hand, for the remaining pairs, we will *reject* the null hypothesis. Stating that except 3 pairs, there exists a significant difference between the mean heights of the players.

4.2.1 Bonferroni adjustments

As stated in section 3, since we are conducting multiple analysis there are more chances of committing *Type I* error. Hence to avoid this as much possible, we apply Bonferroni adjustments on the above pairwise t test analysis. Upon applying this adjustment we can see that the following values for pair wise changed:

Results : We get the following results upon performing pairwise t test on the given data set:

Table 6: P-values after using Bonferroni adjustments to the pairwise t test results.

	basketball	handball	ice hockey	soccer	volleyball
handball	0.01 | **0.09**	-	-	-	-
ice hockey	< 0.01	< 0.01	-	-	-
soccer	< 0.01	< 0.01	0.38 | 1.00	-	-
volleyball	0.85 | 1.00	0.01 | **0.09**	< 0.01	< 0.01	-
water polo	0.01 | 0.04	0.65 | 1.00	< 0.01	< 0.01	0.01 | 0.04

In the above table, the left values in a particular column represents the values before the Bonferroni adjustments, the right value represents after the Bonferroni adjustment.

Conclusion : We can see from above table 6, even though the value of p increased for every pair, for the ones in bold, it was significant change. Meaning, earlier the corresponding p-value was below the level of significance and now it is above it. Hence, for such cases also we would now *fail to reject* the null hypothesis.
Hence, we can summarize that there exists significant difference in the mean heights of the players of the following pairs of sport group: *basketball - ice hockey, basketball - soccer, basketball - water polo, handball - ice hockey, handball - soccer, ice - hockey - volleyball, ice - hockey - water polo, soccer - volleyball, soccer - water polo, volleyball - water polo*

5 Summary

This project report basically deals with hypothesis testing of the data set given. The data set was compiled by the instructors of the course Introductory Case Studies at TU Dortmund University (Summer Semester 21). It was compiled by various sources like web, personal contacts etc. It has overall *112* observations dealing with name and heights of players of 6 German men's national sports team.

The report deals with basic aim to find whether the height of players play an important role in determining which sport he should be playing or not. Since there are many factors which support that a taller person may be beneficial in a particular sport like basketball, it is of great interest to understand whether this difference actually exists or not. Before we begin to conduct our tests, we first check whether all the assumptions which are required are met or not. After analyzing the data set, we see that all the required three assumptions (the samples collected from population are independent of each other;the variance of distributions in the population are equal;the dependent variable is also normally distributed in each group) hold true. We now first run a global test (one way ANOVA) on the data set provided to us. It is here, that we get significant evidences to reject our null hypothesis i.e. we conclude that there exists atleast one particular sport whose mean value of heights is different from the rest. Furthermore, to narrow down the pairs of such significantly different mean heights, we conduct *pairwise t testing* which is nothing but comparing two sports at a time. Since we know that conducting multiple analysis increases the chance of committing *type I* error, we try to control this using the Bonferroni corrections. We then observe that five of the pairs namely *basketball-handball; ice hockey-soccer;basketball-volleyball; handball-volleyball and handball-water polo* have no significant difference among the means of height of the players while rest of the pairs do have significant differences. Although we cannot overlook the fact that it may increase the vulnerability of *type II* errors which are not taken into account here.

Hence to conclude, we can see that while few sports have no significantly different mean heights, most of them do have. This may be due to the biological fact involved in sports like few sports may be easier for a taller person. It may be noted here that since for this project, we only considered the height and not other factors like weight of the players etc, it would be of great interest if we do consider them as well for our future analysis. It would give us more data to support or reject our conclusions.

Bibliography

Alex Stepanov, David Unger, James Balamuta. *Applied Statistics with R.* University of Illinois at Urbana-Champaign, Champaign, IL, United States, 2021.

Ken Black. *Business Statistics For Contemporary Decision Makin.* University of Houston—Clear Lake, United States of America, 2010.

Charles Merriam, George Merriam. *Merriam Webster's Collegiate Dictionary, 10th ed.* Encyclopaedia Britannica, Massachusetts, U.S., 1993.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

Web. p-value. URL `https://www.simplypsychology.org/p-value.png`.

Appendix

A Additional figures

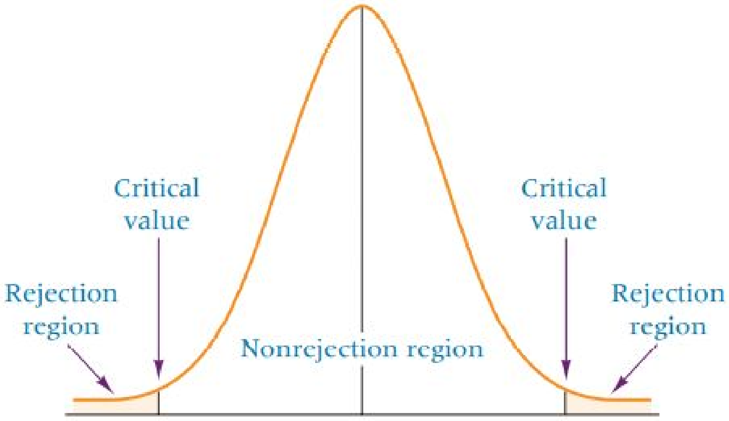

Figure 5: Demonstrating the rejection and non-rejection regions