TU Dortmund

Introductory Case Studies

Project 2: Comparison of multiple distributions

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Bhavesh Jain

Group number: 14

Group members: Akanksha Tanwar, Bhavesh Jain, Sohith Dhavaleswarapu, Elif Ilgin Yildiz, Sükrü Bakan

December 10, 2021

Contents

1 Introduction

One of the major expenditure when living in any city is the rental expenditure. One way the rent is calculated is on the basis of the area of the property. This helps in deciding how much a person could afford for the rent on the basis of the area of the property. In recent years, the rental expenses played a influenced the decision for acquiring a property on rent in a particular city. Generally, properties with more area have many advantages but comes with a higher rental expense, but sometimes apartments with less area would still have a higher rent.

The goal of the project includes to determine are there any significant differences in the mean of the rent per square meter value of the four largest cities of the Ruhrgebiet(Ruhr area). For the first task, we perform an analysis of variance test and based on its output, we concluded that the overall rent per square meter means of the cities are not similar and we reject the null hypothesis. For the second task we perform multiple pairwise t-tests for each 4 cities by obtaining 6 different combinations, which provided significant evidence to not reject the null hypothesis for few pairs. Then, the test results are updated using the Bonferroni correction method. Further, using the updated critical value we can find the existence of any significant differences or variations between the group.

In this report, the section 2 consists the details of the data set provided and the objectives of the project. In section 3, the inferential statistical methods such as analysis of variance (ANOVA), multiple pairwise t- testing and the Bonferroni correction method, which are used to accomplish the tasks, are stated and explained in detail. This section also includes details about how the test works and what are the outputs generated after performing tests, while the assumptions before performing the tests are validated. The section 4 consists of the inference determined about the population. Methods from the section 3 are applied and the test results helps to determine insights about the population. In this section, various tests are performed and they provide insights about the larger population. Finally, possible conclusions and views about further investigations are stated in the summary.

2 Problem statement

2.1 Data set and data quality

The official data set is provided by the instructors of the course Introductory Case Studies at TU Dortmund University in the winter semester 2021/2022. The data set contains 3 variables with 200 observations. The first variable is *ID*, which is numeric and is a unique representation of each record in the data set. The second variable is *sqmPrice* which refers to the price per square meter. The third variable is *regio2*, which has 4 different categories, it refers to the city for which the record belongs to. There are no missing values in the data set provided and *sqmPrice* and *regio2* are the crucial variables considered for the analysis of the project.

2.2 Project objectives

The primary objective of this project is to compare means of multiple random samples drawn from a larger population. For the first task, to determine if the mean *sqmPrice* of the groups are similar on the whole set or not, we do this by performing the analysis of variance test (ANOVA). Before proceeding with the test, the assumptions such as normality, homogeneity of variance and independence in the value of the data set are validated in order to perform the ANOVA test. Then, the p-value obtained after the test provides evidence to either reject the null hypothesis or failing to reject it. After the first task, the second objective is to perform a multiple pair wise t-test in order to check and control the probability of the occurrence of type I error. Here, p-values are determines for t-tests conducted on each pair, and this p-value helps us to determine the rejection decision of the null hypothesis. After performing multiple tests, the inflation of type I error is controlled by the Bonferroni correction method. The significance level is predetermined and is set to 5% while performing analysis in this project.

3 Statistical methods

In this section, several methods from inferential statistics are described. These methods are useful for providing information about the population, while using just a sample

data from the population given. These methods are used in the further analysis of the provided data set.

3.1 Statistical hypothesis testing

For testing the occurrence of an unknown parameter θ in the distribution of the random variables, we make use of statistical hypothesis testing. The distribution is derived from a larger population. Hypothesis testing checks the validity of the hypothesis which means having an assumption for the parameter of the population and then check for its validity. In statistical hypothesis testing, two types of hypotheses are considered, the first one is the null hypothesis, and the second one is the alternative hypothesis These two hypotheses complement each other. For statistical testing we assume random sampling of the data set from the population. If the value of the parameter θ is uniquely prescribed i.e., $H_0 : \theta = \theta_0$, then that hypothesis is known as simple hypothesis and in composite hypothesis, it can have several values for θ. The null hypothesis is the main hypothesis and is denoted by H_0 and the other hypothesis is the alternative hypothesis, which is denoted by H_A. For any statistical test, if H_0 is right, H_A is wrong and if H_A is right, H_0 is wrong.

Error types

In making statistical decisions, there are chances of occurrences of 2 types of errors. Type I error - This error occurs when we reject null hypothesis H_0, when it's not meant to be rejected. Type II error – This error occurs when we fail to reject alternative hypothesis H_A, while H_A is meant to be rejected.

	Retain Null Hypothesis(H_0)	Reject Null Hypothesis(H_A)
While H_0 is True	no error	type I error
While H_A is True	type II	no error

Rejection region and critical value

After performing a hypothesis test on a sample distribution of random variables, we obtain an interval R in that sample space Θ of range X where the possible outcomes R

($R \subset X$), are the ones in which we reject the null hypothesis. This region is known as rejection region of X. If $X \notin R$, then we fail to reject null hypothesis.

$$X \in R = \text{rejection of } H_0$$
$$X \notin R = \text{fail to reject } H_0$$

The critical value defines a boundary that differentiates the rejection region and the region where we fail to reject null hypothesis.

Test statistic and significance level

The test statistic is the output obtained after performing a hypothesis test. It is a special random variable which is derived from the random sample on which the test is performed. It is used to decide whether the null hypothesis should be rejected or not Rasch et al. (2019).
The significance level is the fixed value of the probability where we can reject the null hypothesis H_0, given the condition that its true. It is denoted by α and its value ranges from 0 to 1 i.e., $(0 < \alpha < 1)$. It is also known as the upper bound of type 1 error and is given by $\alpha = P\{rejectH_0|H_0true\}$. Typically the value of α is set to 0.05 or 0.01 Rasch et al. (2019).

P value and Confidence interval

The P value is a probability value, where the value measures the existence of the data by a random chance. The P value can take any value from 0 to 1. Smaller P value provides strong evidence to reject the null hypothesis and if the P value is higher it provides significant evidence to retain the null hypothesis and reject the alternative hypothesis. McLeod (2019)
The Confidence level γ is the maximum probability where we do not reject the null hypothesis assuming that the null hypothesis is true. It is given by $\gamma = (1 - \alpha)$. The confidence interval is pre-determined before the test Gillard (2020).

3.2 One-way ANOVA test

The abbreviation ANOVA refers to analysis of variance. This parametric test is used to check if there exists any significant difference in the means of two or more independent groups of same population Hay-Jahans (2019).

Assumptions

There exists 3 assumptions that the sample data must satisfy, in order to perform an analysis of variance (test). The first one is the assumption of normality which states that the population from which the random sample data is drawn must be normally distributed. This assumption can be proven with the help of a QQ plot. The second one is the existence of homogeneity of variance. It is assumed that the standard deviation of the groups is similar and the third one is the independence in the data Hay-Jahans (2019).

Hypothesis

The null hypothesis is H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_m$, where μ is the mean of each individual group and m refers to total number of groups used for comparison. The alternative hypothesis H_A: at least means of two groups are dissimilar Hay-Jahans (2019).

Normal QQ-Plot

A QQ-plot is used to determine how well the distribution of the data given is in the form of standard normal distribution. The figure 1 shows a QQ-plot. We plot the values in our data set against the normal distribution, i.e., keeping theoretical quantiles on the $x - axis$ and the sample quantiles on the $y - axis$. When a reference line is imposed on the plot and if the plotted points lie close to the line, then the data is said to be normally distributed Hay-Jahans (2019).

Working of ANOVA test

An ANOVA performs an F-test, which allows the comparison of means among multiple groups at once. If any of the group means is significantly different from the overall mean,

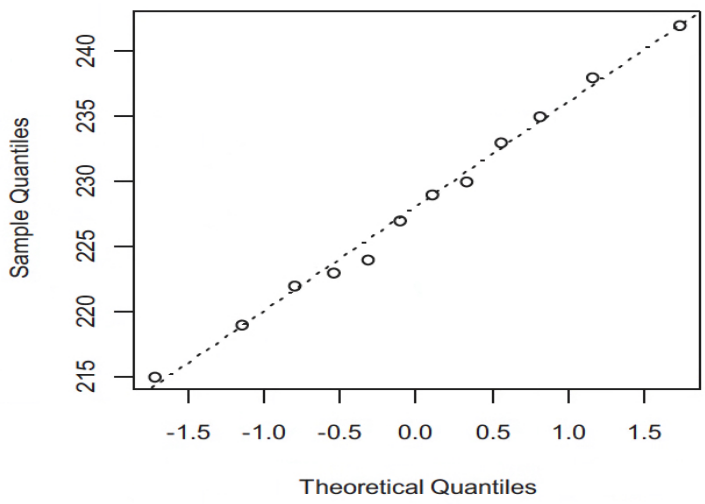

Figure 1: Example of a Normal QQ-Plot
(Hay-Jahans, 2019)

then the null hypothesis is rejected.
Let us consider an independent, identically and normally distributed sample which contains m groups and total number of n observations. Considering sample data set X which has $x_{11}, \ldots, x_{mn}$ observations and each specific value in the data set is denoted by x_{ij}, where $i = 1 \ldots m$ indicates the specific group to which the observation belongs and $j = 1 \ldots n$ refers to each individual observation in that corresponding group Levin (2011).

The initial calculations are performed which are then used to output the F-ratio. The sum of all individual values in a specific i^{th} group is denoted by T_i and is given by $T^i = \sum_{j=1}^{n_i} x_{ij}$, where n_i refers to total number of observations in the i^{th} group. Then we sum of the values of T_i for all m groups in G, where $G = \sum_{i=1}^{m} T_i$. The mean of individual group is given by $\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$ and $\bar{\bar{x}}_i = \frac{\sum_{i=1}^{m} T_i}{n_i}$. The raw sum of squares (RAW S.S) = $\sum_{i=1}^{m} \sum_{j=1}^{n_i} {x_{ij}}^2$. The correction factor (C.F.)= $\frac{G^2}{n}$, the total sum of squares (T.S.S.) = RAW S.S. subtracted by the correction factor (C.F.) Levin (2011).

Firstly, the degree of freedom (df) which is given by $m - 1$ for the groups for the whole data set, it is given by $n - m$. Secondly, the sum of squares are calculated between the groups and within each group. The sum of squares between groups (B.S.S.) explains the total variation between each individual group's mean and the overall mean of the groups

which is given by B.S.S= $\sum_{i=1}^{m} \frac{T_i^2}{n_i}$- correction factor (C.F.). Similarly sum of squares within the groups are summed up in (W.S.S.), where $W.S.S = T.S.S \breve{} B.S.S$. Then after that, we calculate mean sum of squares (M.S.S.), which is the mean of the sum of the squares. For between the groups, it is calculated by dividing the BSS by $m-1$ i.e., Mean sum of squares between groups $(MSB) = \frac{BSS}{m-1}$ and similarly for mean sum of squares within the groups $(MSW) = \frac{BSS}{n-m}$. Finally, the F ratio is determined by dividing MSB with MSW, i.e., F= $\frac{MSB}{MSW}$ Levin (2011).

Rejection of null hypothesis

After determination of the F ratio, we determine the critical value F_{crit} with the help of RStudio. The next step is to compare the F_{crit} with the F ratio obtained. If F_{crit} value is greater than F ratio, then we fail to reject null hypothesis and if F_{crit} value is less than F ratio, then we reject the null hypothesis. An illustration of the same is provided in the figure2 Levin (2011).

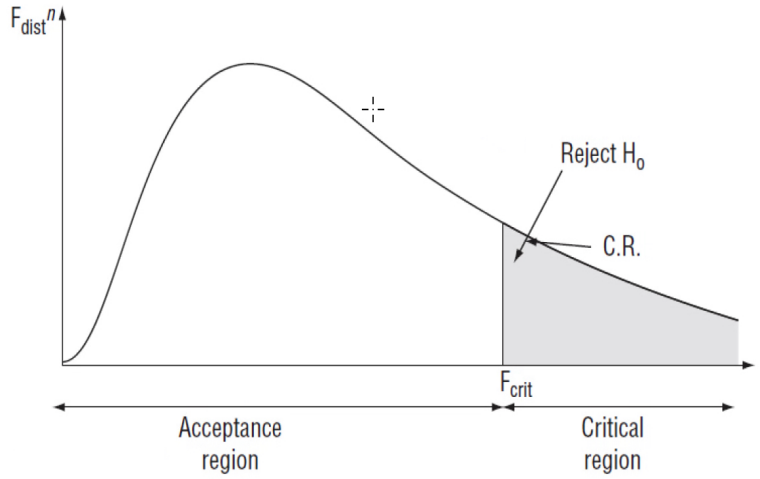

Figure 2: Rejection of null hypothesis after ANOVA test
(Levin, 2011)

3.3 Multiple testing

The scenario where we conduct many individual hypothesis tests is known as multiple testing. There is one problem which arises when conducting a multiple test, at a constant

significance level α for each individual test, there is a high chance of having at least one false rejection. This problem is known as Multiple testing problem Romano et al. (2010).

Family wise error rate

When performing multiple testing, the probability of occurrence of at least one type I error is known as family wise error rate. For n tests and α as significance level for an individual hypothesis test, the family wise error rate can be calculated by Family-wise error rate $= 1 - (1 - \alpha)^2$ Romano et al. (2010).

3.4 Bonferroni correction

Bonferroni correction solves the problem of multiple testing. In multiple testing, where we tend to see an inflation in type I error, with the help of Bonferroni correction, we control the type I error which leads to retaining more null hypotheses. In this method, the P value is corrected by multiplying with the number of tests. For a given set of P values of m number of tests, $P = P_1, P_2, P_3 \dots P_m$, we reject the null hypothesis H_0 if $(P_i.m) < \alpha$, where P_i is P value of an individual test and α is the significance level Wasserman (2010).

3.5 T-test

A t-test is used to determine if there exists any significant difference in the means of a pair of samples provided from the same population. The t-test outputs a t statistic which is basically a ratio of the difference of the means of both the samples considered for the test to the variance that exists within them. The larger the t-value, we can say that the groups are different and the smaller the t value indicates similarity in groups Gillard (2020).

Hypotheses

The null hypothesis H_0 considers that the mean of both the samples are equal and the corresponding alternate hypothesis says that mean of both are samples are dissimilar. For total number of groups m and i and j denotes individual group in the data set, then

H_0 and H_A are as follows Gillard (2020).

$$H_0 : \mu_i = \mu_j; i \neq j; i, j = 1, \ldots m$$
$$H_A : \mu_i \neq \mu_j; i \neq j; i, j = 1, \ldots m$$

Assumptions

To perform a T-test, the following assumptions are considered. The values must be independent, which means that, there shouldn't be any dependency among the values. Data obtained is a random sample from the population. It follows a normal distribution. The values must be continuous and the variances for two independent groups are equal.

Working of t-test

The t score or t- value is the ultimate output of performing the t-test which is calculated by:

$$t = \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{n_1+n_2}{n_1.n_2}}}$$

Consider X and Y as two random samples where X contains $x_1, x_2, x_3 \ldots x_{n1}$ values and Y contains $y_1, y_2, y_3 \ldots y_{n2}$, x_i and y_i denotes an individual value from the sample. The variables n_1 and n_2 are the total number of values in the samples respectively. S_p is the value of the standard pooled error which is given by $S_p = \sqrt{\frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}}$, where s_x and s_y are $\sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2}$ and $\sqrt{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i - \overline{y})^2}$ for X and Y respectively and $\overline{x}$, $\overline{y}$ are the mean for X and Y. The degrees of freedom correspond to the value obtained by subtracting 1 from the total number of observations of both the samples, i.e., $df = n - 1$. The ultimate output of the test i.e., the p value is obtained with the help of RStudio Gillard (2020).

Rejection of null hypothesis

If the P value obtained is greater than the significance level which was predetermined before the start of the test, then we fail to reject the null hypothesis H_0 based on the evidences Gillard (2020).

4 Statistical analysis

In this section, detailed analysis of the data set is provided applying inferential statistical methods which are stated in the section 3. Then the results are interpreted. To perform the analysis, we use R studio software (version 4.1.2) RStudio Team (2021).

4.1 Descriptive analysis

A general descriptive analysis is performed to describe how the data is distributed and various central tendency values are determined. The given sample data has data value of rent per square meter for 200 properties and are equally divided into 4 groups based on the cities they belong. The mean, median and rest other values are summarized in the table1.

	ID	sqmPrice	regio2	region
X	Min. : 169	Min. : 5.843	Length:200	Bochum :50
X.1	1st Qu.: 3213	1st Qu.: 8.286	Class :character	Dortmund:50
X.2	Median : 6101	Median : 9.183	Mode :character	Duisburg:50
X.3	Mean : 6220	Mean : 9.149		Essen :50
X.4	3rd Qu.: 9416	3rd Qu.: 9.896		
X.5	Max. :12067	Max. :13.629		

Table 1: Central tendency values of the data set

4.2 Global test

In this subsection, we conduct a global test using an ANOVA hypothesis testing. To perform this, we first validate the assumptions which were mentioned in the section 3.2. To validate the first assumption that the data is normally distributed, we use QQ plots which were described above in the section 3.2.

From the figure, it can be seen that the plotted lines in the QQ plot for the 4 cities are little scattered for some observations and deviate from the reference line. These deviations might not violate the assumptions as the sample size is smaller than the population. With this we can assume that the population is normally distributed. The second assumption of this test states that the values in the individual group are independently drawn from the population, we validate this assumption as true because the

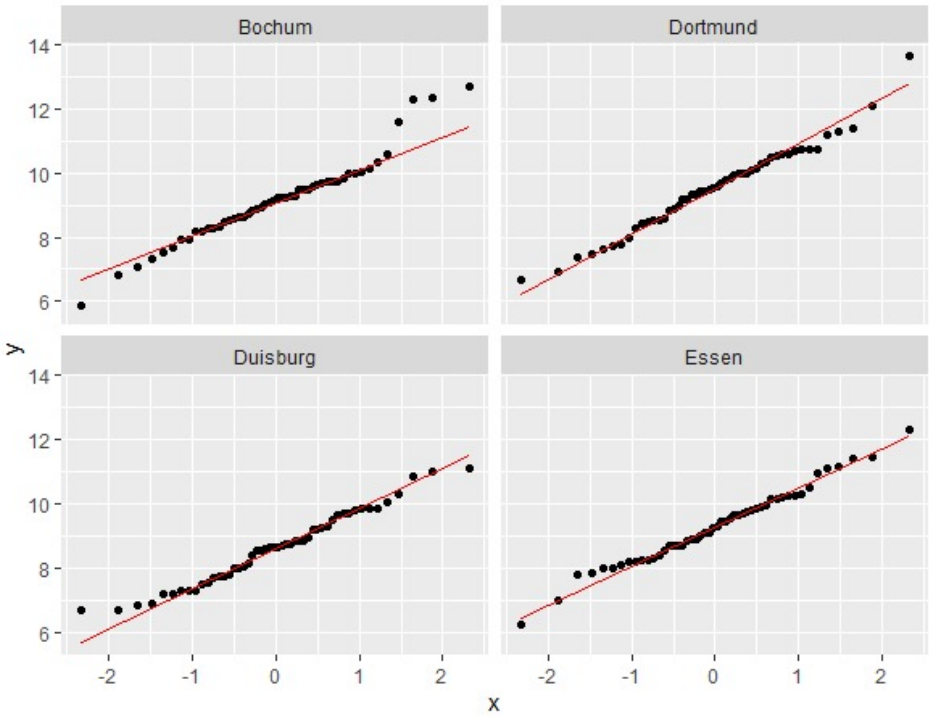

Figure 3: Normal QQ plot showing relation between the sample distribution and Normal distribution

data is collected randomly by the data set providers and the values are independently drawn from the population.

To validate the second assumption which states that the distinct groups must have equal variances, we take the help of a box plot. With the help of box plot in the figure4, it can be seen that there exists homogeneity in the variances between those groups. The median values of the rent per square meter for the four cities lie around 9, this can be visually seen from the box plot.

For the last assumption that the data of one group of the population is independent from other sample. In this case, the rent per square meter of different cities are not comparable or as there is no relation between the cities and hence it is assumed to be independent. The independent data items are not connected with one another in any way, this includes the observations in both between and within groups in the sample. After checking the validity of the assumptions, we proceed on performing the ANOVA test. The null and the alternative hypotheses for this test are $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \& H_A : \mu_i \neq \mu_j$ for all $i \neq j$. Here, μ_i is the mean of the rent per square meter values of the i individual cities of the data set i.e., "Bochum", "Dortmund", "Duisburg", "Essen".

From the table2 The F statistic determined by the test is 4.68 and the p-value is 0.0035.

Here, the p-value is smaller than the significance level of 0.5. With this, we can say that the null hypothesis is rejected which says that we reject the assumption that the mean of the rent per square meter values of the cities are equal.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	22.15	7.38	4.68	0.0035
Residuals	196	309.08	1.58		

Table 2: ANOVA test output

4.3 T-test

In this subsection, we perform t-test and analyse the results. As the test is conducted on a pair, for the given 4 groups there are total of 6 pairs to perform the test. The test is performed using the function *pairwise.t.test* in the RStudio and the results are provided in the table3. For each pair of tests, the null and the alternative hypotheses are $H_0 : \mu_i = \mu_j$ and $H_A : \mu_i \neq \mu_j$, where i and j are the different groups involved in the test. From the table3, it can be determined that there exists 3 pairs where the P-values provides significant evidence to fail rejecting the null hypothesis. The pairs are "Bochum and Dortmund" & "Bochum and Essen" & "Essen and Dortmund", as the p-value is greater than the significance level. We reject the null hypotheses for the pairs "Bochum and duisburg", "Dortmund and Duisburg" and "Duisburg and Essen" as the p values are smaller.

Table 3: Summary of two sample t-test results without adjusting p value

	group1	group2	p-value	result
1	Bochum	Dortmund	0.1367	Do not Reject
2	Bochum	Duisburg	0.0364	Reject
3	Bochum	Essen	0.5558	Do not Reject
4	Dortmund	Duisburg	0.0004	Reject
5	Dortmund	Essen	0.3670	Do not Rejectt
6	Duisburg	Essen	0.0076	Reject

4.4 Modification with Bonferroni correction

After the determination of the pair wise t test results, we correct the p value by multiplying the p value with total number of tests performed. The results of the corrected

p value are summarized in the table4. Where it is seen that, the probability of having equal average rent per square meter for the pair "Bochum and Duisburg" are high and we fail to reject the null hypothesis for that pair. In this way, the type I error is controlled.

Table 4: Summary of two sample t-test results with adjusting p value

	group1	group2	p-value	result
1	Bochum	Dortmund	0.8201	Do not Reject
2	Bochum	Duisburg	0.2182	Do not Reject
3	Bochum	Essen	1.0000	Do not Reject
4	Dortmund	Duisburg	0.0024	Reject
5	Dortmund	Essen	1.0000	Do not Rejectt
6	Duisburg	Essen	0.0456	Reject

5 Summary

As getting to know the average rent per square meter is very important to analyse the price variations among cities. So, in this report, we have b een given a dataset which is an extract from a large dataset which is availble on "www.kaggle.com.". The dataset contains rent per square meter values for 4 different cities of of the Ruhrgebiet (Ruhr area) as of February 20 2020. There were no missing values in the the provided dataset, Our main aim of the project is to perform a comparison of multiple distributions in the data.

Our first task was to verify if the rent prices per square meter differ between the four cities, so we performed a analysis of variance test and it provided significant evidence to reject the null hypothesis, where null hypothesis stated that the average rent per square meter for all four cities are equal. The assumptions regarding the dataset are validated before performing the ANOVA test.

After rejecting the null hypothesis, we performed a t-test on each pair of cities. This resulted to a total of 6 pairs of t-test. For each t-test, the rejection of null hypothesis is decided on comparison with the obtained p value. In this, we see that 3 pairs fail to reject null hypothesis and 3 pairs reject the null hypothesis. Here, the null hypothesis is that the average rent per square meter is equal for both the cities in a pair.

Lastly, to control the type I error, we adjusted the p value with the help of Bonferroni correction method and saw that there was one more pair where we fail to reject the null hypothesis.

For further studies, it would be useful to study what other factors are responsible for the varying rental expenses in the cities and what changes in the factors have caused rent to decline. Finally, we can determine how these changes affect the country's economy in a whole.

Bibliography

J. Gillard. *A First Course in Statistical Inference*. Springer Undergraduate Mathematics Series. Springer International Publishing, 2020. ISBN 9783030395612. URL https://books.google.co.in/books?id=aJneDwAAQBAJ.

C. Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019. ISBN 9780429827273. URL https://books.google.co.in/books?id=8UOCDwAAQBAJ.

R.I. Levin. *Statistics for Management*. Pearson Education, 2011. ISBN 9788177585841. URL https://books.google.co.in/books?id=loa4KcmMmbcC.

Saul McLeod. What a p-value tells you about statistical significance. *Simply Psychology*, 2019. URL www.simplypsychology.org/p-value.html.

D. Rasch, R. Verdooren, and J. Pilz. *Applied Statistics: Theory and Problem Solutions with R*. Wiley, 2019. ISBN 9781119551546. URL https://books.google.co.in/books?id=Ko6pDwAAQBAJ.

Joseph P Romano, Azeem M Shaikh, Michael Wolf, et al. Multiple testing. *The New Palgrave Dictionary of Economics. Forthcoming*, 2010.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2021. URL http://www.rstudio.com/.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer New York, 2010. ISBN 9781441923226. URL https://books.google.co.in/books?id=RMdgcgAACAAJ.

Appendix

A Additional figures

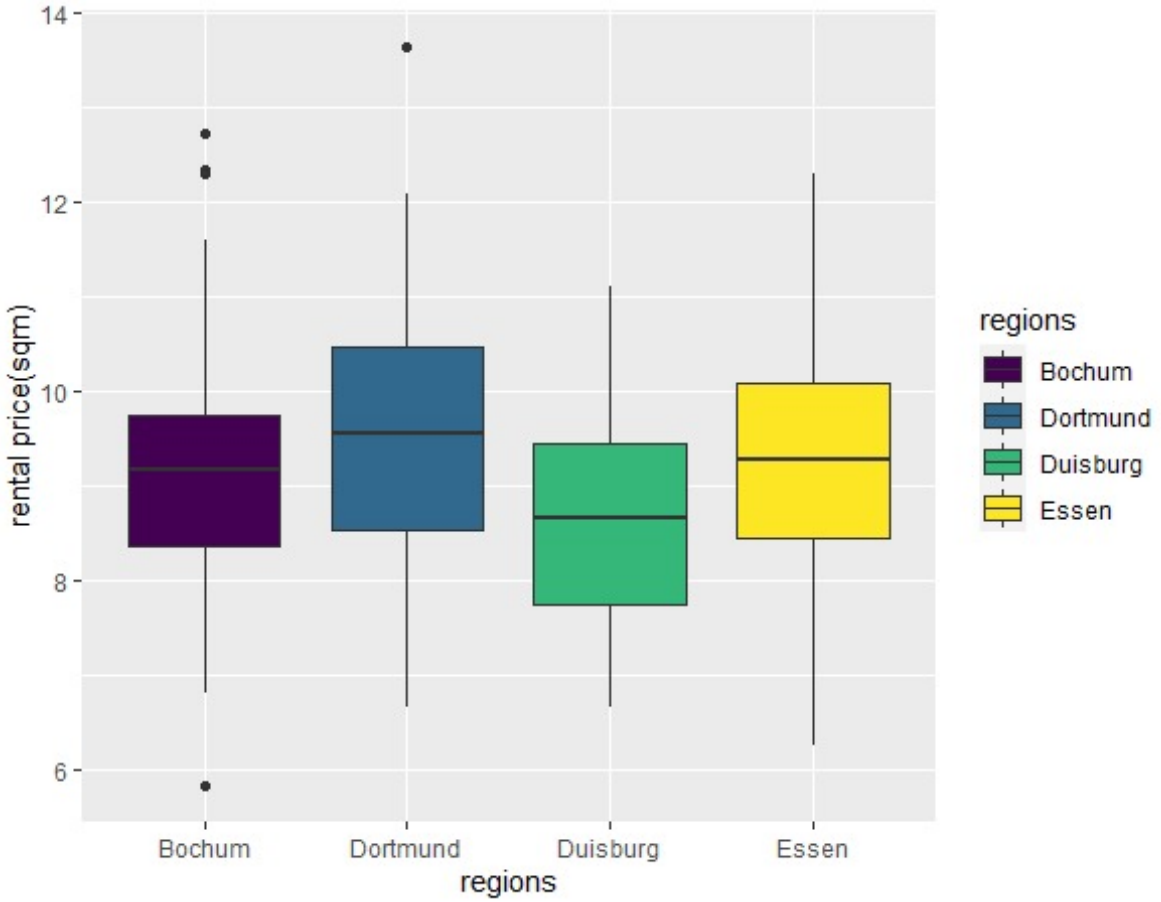

Figure 4: Boxplot of the 4 cities.