TU Dortmund

Introductory Case Studies

Project II: Comparison of Multiple Distributions

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

Author: Kunal Kochar

Group number: 22

Group members: Kunal Kochar, Natasha Sahare, Harshini Eggoni, Supritha Palguna

December 10, 2021

Contents

1 Introduction

Renting or buying a new house is always a daunting process that often seems risky. Homeowners, real estate brokers and investors all need accurate rent prices. The value of house is a major component of the household's aggregate expenditure. The share of the value of a house varies by countries and cities. There are several factors which affects the value of a rental property in general. Nowadays, when people are more keen in buying properties, it is of great interest to understand and study how and which factors affect the price of the property. The aim of this project is to compare the rental price per square meter of properties located in four large cities of the Ruhrgebeit (Ruhr area).

To achieve the objectives, first we perform the descriptive analysis to understand the data set. This part is followed by testing the equality of rental price associated with each city. Later, pairwise comparison are conducted to narrow down the pairs where there is a significant difference between the rental prices per square meter. Lastly, the pairwise comparison are adjusted by Bonferroni method in order to control the Type-I error.

In section 2, an overview of the given data in addition to the data quality is provided. Section 3 deals with the explanation of several statistical terms used for analyzing the data. In Section 4, the presented statistical methods are applied to the given data set and the results are interpreted. The concluding Section 5 contains the summary which deals with all the interpretations and findings and possible further investigations.

2 Problem statement

2.1 Description of data set and quality

The data set is provided by the lecturers of the course Introductory Case Studies at TU Dortmund University in the winter semester 2021. The data set is an extraction of the available rental offers data from a website (Kaggle.com) which comprises data of one of the biggest real estate web-portals in Germany: (Immobilienscout24, 2020). It contains 200 observations also known as records and contains three features or variables. Each record of the data represents the following information, *ID* (Integer-type numerical) which corresponds the record number, *sqmPrice* (Decimal-type numerical)

indicates the rental price of the property which is measured in square meter and finally *regio2* (Nominal-type categorical) which indicates to which city the property belongs to. Additionally, the data set contains no missing values or NA values and the quality of data is decent.

2.2 Project objectives

The project focuses on the following objectives, firstly performing descriptive statistics of the data set, following by defining hypothesis for the tests and verifying the assumptions and lastly conducting appropriate tests and interpreting the results. Initially, a test is conducted to check if the average rental price among four cities are equal or not. For this case, one-way ANOVA is used. However, we also need to verify few assumptions before conducting tests such as normality assumption using QQ-Plot, homogeneity of variances assumption and lastly the assumption of independence. Later, pair-wise comparisons among the cities are performed to check whether there is any significant difference. Lastly, we apply Bonferroni adjustment to control the effect of type-I error in pairwise comparisons and the results are compared and interpreted.

3 Statistical methods

A set of inferential and descriptive statistics are presented in this section which will be used for analyzing the data set. The descriptive method helps in describing the characteristics of a given sample, inferential method helps to draw conclusions from a sample and generalize them to the population. For all data processing and visualizations, software R (R Developement Core Team, 2021) of version 4.0.5 is used. Below are the statistical methods which are used to analyze and evaluate the given data set for the current report.

3.1 Statistical Test

Statistical test is a mechanism for making quantitative decisions whether there is enough evidence to 'support' or 'reject' a hypothesis about the unknown parameter to occur in the distribution of a random variable. (Rasch et al., 2020, p. 39)

3.1.1 Hypothesis Testing and Statement

Inferential statistics is all about making inferences or predictions about the value of a particular observation. The major inferential statistics used for making such decisions is to perform a hypothesis test. A hypothesis test involves two hypothesis: the null hypothesis and the alternative hypothesis. The null hypothesis H_0 is a statement to be tested. The alternate hypothesis H_A is a statement that is considered to be the alternative to the null hypothesis. The hypothesis H_0 is right, if H_A is wrong and vice versa. (Rasch et al., 2020, p. 39)

3.1.2 P-value and Significance Level

P-value is used in hypothesis testing which helps to decide whether to support or reject the null hypothesis. It ranges from 0 to 1. The p-value is always being compared with the level of significance in order to decide whether the null hypothesis should be rejected or not. If the p-value is less then the level of significance, the null hypothesis is rejected and we can conclude that there is a significant difference between the certain characteristic of a population and if not, then we fail to reject the null hypothesis. So, the smaller the p-value, stronger the evidence to reject the null hypothesis. The p-value is used in many hypothesis tests such as ANOVA test, T-test which will be discussed later in this section.

Significance level or also known as level of significance denoted by α is the probability to make an error of the first type, which means rejecting the null hypothesis while it holds true. Besides an error of the first kind, an error of the second kind may occur if we fail to reject the null hypothesis although it is wrong; the probability that this occurs is the second kind risk. The level of significance ranges from 0 to 1. We have to decide the level of significance before starting the experiment and must not be changed afterwards in order to obtain the desired hypothesis. (Rasch et al., 2020, p. 39)

3.1.3 One-way ANOVA Test

One way analysis of variance or One way ANOVA is a statistical technique that is used to check all the sample means at one time. The null hypothesis H_0 is valid when all the sample means are equal and the alternate hypothesis H_A is valid when at least one of

the sample mean is different from the rest of the sample means and we can't tell which one specifically.

In order to check whether the means of the samples are equal, the following hypothesis are formulated:

$H_o : \bar{x_1} = \bar{x_2} = ... \ \bar{x_k}$

H_A : The means are not all equal.

(Christopher Hay-Jahans, 2018, p. 291)

There are several assumptions which need to be taken into consideration before the hypothesis are tested. Firstly, the one-way ANOVA test is applicable only if population variances are equal. Secondly, it is assumed that the population and the observed response are independent and lastly the underlying random variables are normally distributed. All these assumptions will be verified in the next section. ANOVA uses the F-test to determine whether the group means are equal.

In one-way ANOVA test, the F-statistics is the calculated as below:

SSC (Sum of square) = $\sum_{j=1}^{c} n_j(\bar{x_j} - \bar{x})^2$
SSE (Sum of square of errors) = $\sum_{i=1}^{n_c} \sum_{j=1}^{c} (x_{ij} - \bar{x_j})^2$
MSC (Mean square of columns or groups) = $\frac{SSC}{df_c}$
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
c= number of groups or samples
N = total number of observations
df_c = c-1 and df_e = N-c
n_j = number of observations in a group or sample
$\bar{x}$ = overall mean
x_j= a particular group mean
x_{ij}= individual value

If the F-value computed by ANOVA is greater than the critical value obtained from F-distribution table, we reject the null hypothesis otherwise we fail to reject it. (Ken Black, 2010, p. 409)

3.1.4 Two sample T-Test

The two sample T-test is used to compare the means of two different samples. A set of simultaneous pairwise comparison is normally conducted through tests of hypothesis having the following form,

$H_o : \mu_j = \mu_k$

$H_A : \mu_j \neq \mu_k$

where μ_i and μ_j are the population mean of any two groups j,k.

There is an assumption underlying this technique which needs to be taken into consideration before the hypothesis are tested. The assumption is that the comparison group being studied is normally distributed for the population and the variances are equal. The assumption will be verified in the next section.

The formula to calculate the t-statistics for a two sample t-test is mentioned below:

$$t = \frac{\bar{x_j} - \bar{x_k}}{\sqrt{\frac{{s_j}^2(n_j-1)+{s_k}^2(n_k-1)}{n_j+n_k-2}}\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}}$$

where
n_j, n_k are number of observations
$\bar{x_j}$, $\bar{x_k}$ are estimates of μ_j, μ_k
${s_j}^2, {s_j}^2$ are sample variances of the two independent samples.
The test values are t-distributed with $n_j + n_k - 2$ degrees of freedom which indicates that we can accept or reject H_o by using $(1-\alpha)$-quantile of t-distribution. The null hypothesis will be rejected if:
(a) $t < t(n_j + n_k - 2, \alpha)$
(b) $t > t(n_j + n_k - 2, 1 - \alpha)$
(c) $|t| < t(n_j + n_k - 2, 1 - \frac{\alpha}{2})$
If H_o is rejected it can be concluded that the two means are different.(Rasch et al., 2020, p. 63)

3.1.5 Bonferroni Adjustments

The Bonferroni method adjusts probability values because of the increased risk of type-I error when making multiple statistical tests. When we conduct multiple tests the chance of committing type-I error increases and thus increasing the likelihood of generating

the false significant result by chance. It is a conservative test that, although protects from Type-I error, however vulnerable to type-II error (being failed to reject the null hypothesis when it is in fact false for the underlying population) since it keeps a reciprocal relationship with the type-I error.

In order to get the adjusted p-value, we divide the family-wise type-I error by number of simultaneous comparisons (m) in order to provide the adjusted error. As a result, the significance level becomes lower and a reference to which the p-values are multiplied. Bonferroni adjustment computes the new p-value as below:

$\alpha_{adjusted} = \alpha/m$
or
$p_{adjusted} = p.m$

The assumptions for the bonferroni's procedure are the same as for the one-way ANOVA. That is, the underlying random variables are independent and normally distributed with equal variances. (Christopher Hay-Jahans, 2018, p. 293)

3.2 Graphical Tools

This part introduces to the graphical tools which are used to display distributions and assess the assumptions.

3.2.1 QQ-Plot

QQ-Plot or quantile-quantile plot is a graphical tool to help us assess if a set of data came from some theoretical distribution such as normal or exponential. A QQ-Plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a straight line. For example, if we run a statistical analysis that assumes that the dependent variable is normally distributed, we can use a normal QQ-Plot to check that assumption and if we see that the data points follow straight line, we could say that data is normally distributed and the assumption holds true. We would be using the QQ-Plot in the next section to verify our assumptions. (Clay Ford : Understanding the Q-Q Plots)

4 Statistical analysis

In this section, the statistical methods explained above are applied to the given data set and results are interpreted. Moreover, before applying tests on the data set, we will verify whether all the assumptions holds on test data.

4.1 Descriptive Analysis

Initially, a descriptive analysis is being conducted on the data set. The sample size is n = 200 and there are no missing values. The data set contains a categorical variable *regio2* , which contains four different factor levels such as Dortmund, Bochum, Duisburg and Essen. The number of observations for each factor/level is 50 which can also be seen from Table 1. Furthermore, the data set contains one numerical variable *sqmPrice* for which the effects of the levels are investigated.

Table 1: Number of Observations for each region

Region	Number of observations
Dortmund	50
Bochum	50
Duisburg	50
Essen	50
Total	**200**

Descriptive analysis of the variable *sqmPrice* is represented in Table 2. According to the table, Dortmund and Essen shows a slightly high rental price with mean values of 9.526 and 9.299 respectively whereas the mean rental price per square meter for Duisburg and Bochum are 8.621 and 9.15. Also, we can see that the region Essen shows a relatively larger variability than the others.

Table 2: Descriptive Analysis of Rental Price for Different Regions

Region	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Dortmund	6.662	8.540	9.552	9.526	10.457	13.629
Bochum	5.842	8.357	9.180	9.150	9.738	12.714
Duisburg	6.667	7.753	8.663	8.621	9.450	11.103
Essen	6.250	8.440	9.278	9.299	10.094	12.282

4.2 Assumptions Verification

As mentioned in earlier section, assumptions need to be verified before conducting the tests.

Normality The validity of the normality assumption can be assessed graphically by means of a normal probability QQ-plot. If the data is normally distributed, the points in the QQ-plot lie on the reference (black) line. Figure 1 represents the normality of each group individually. According to Figure 1, we can see that for region Dortmund and Essen, almost all the points lie on the reference line. For the region Bochum and Duisburg we can see a slight deviation from the reference line at the tails but this does not suggest a violation of normality assumption because the data set is small and it is of a sample and not of the population, hence an acceptable level of deviations are also expected in sample from normally distributed population. Hence, the normality assumption holds true.

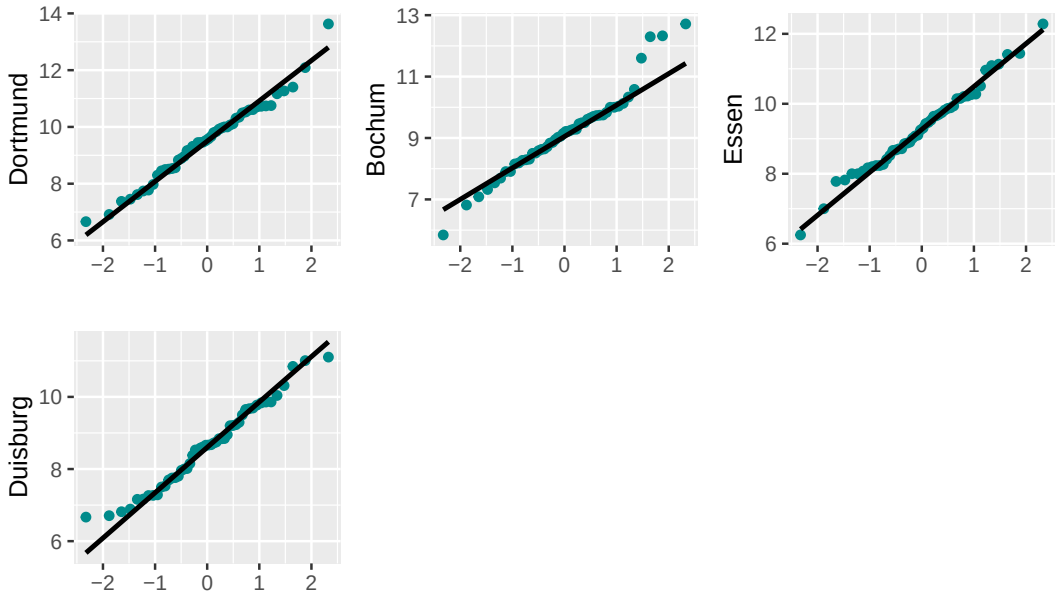

Figure 1: QQ-Plot for Normality Assessment

Data Independence Data Independence is a critical assumption for many tests. Independence means that the sample observations are independent of each other. In this project, it is assumed that a proper randomization is achieved and that all entries are different and independent. Hence, data independence assumption also holds true.

Homogeneity of Variance In order to verify homogeneity of variance assumption we require that the variances of distribution in the population are equal. We can assess this assumption with the help of a box plot which can be referred from the Figure 2.

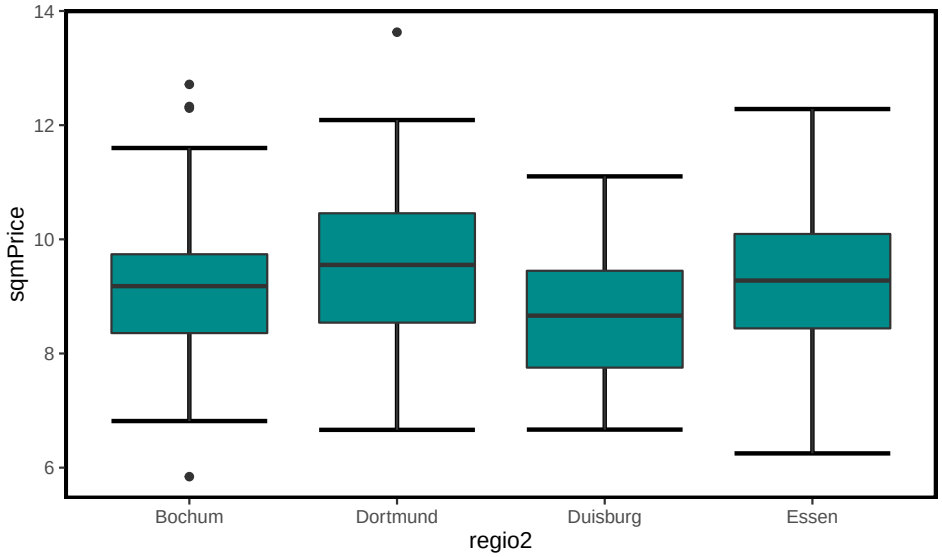

Figure 2: Box-plot for Homogeneity of Variance Assessment

Figure 2 shows the box-plot for the variable 'rental price' in different regions. We can see from the figure that the Inter-quartile range (IQR) within the regions are not equal. Region Dortmund has slightly larger IQR when compared to other regions. While the IQR for the region Duisburg and Essen are almost same. Region Bochum has the smallest IQR comparatively. Overall the IQR range differs among the groups but again the data which is provided is of sample and not of population, we can assume that the assumption holds true and a slight variability is acceptable in the sample data and we cannot claim that the validity of assumption of homogeneity of variance is false.

4.3 Global Test

In this part we address the first objective of the project by applying one-way ANOVA to the data set in order to find if the mean rental prices per square meter differ among the four regions. Regarding the first objective, the hypothesis is formed as stated in section 3 and we already validated that all the assumptions of ANOVA holds.

H_o = There lies no significant difference between the mean rental prices per square meter of the four regions.

H_A = There lies atleast one region whose mean rental price is different from the rest. The R function aov() has been used to perform the test and Table 3 shows the summary of ANOVA test:

Table 3: ANOVA Test Summary

	Degree of Freedom	Sum of Square	Mean Square	F Value	P Value
regio2	3	22.150	7.382	4.681	0.0035
Residuals	196	309.080	1.577		

We can see from the Table 3 that the p-value is less than level of significance 0.05 and hence we have a strong evidence to reject the null hypothesis i.e. mean rental prices differ between the regions which concludes that there is at least one region whose mean rental price is significantly different from the other regions.

When a statistically significant difference between the population means is found, it is of great interest to investigate pairwise differences. Below subsection deals with the pairwise comparisons and their outcomes and interpretations.

4.4 Pairwise Comparison

This part deals with the second objective of the project which requires simultaneous pairwise comparisons of region means. In order to perform the pairwise comparison, we will use pairwise t-test. The hypothesis of t-test is already mentioned in section 3.1.4. Applying the pairwise t-test without adjustment results in the following p-values as shown in Table 4.

Table 4: Non-adjusted p-values for simultaneous pairwise comparisons

	Bochum	Dortmund	Duisburg
Dortmund	**0.137**	-	-
Duisburg	0.036	0.001	-
Essen	**0.555**	**0.367**	0.008

With accordance to the p-values of Table 4 and significance level of 0.05, there are 3 comparisons for which the p-values are above 0.05 or level of significance, which means that there is not enough evidence that the mean rental price of Dortmund- Bochum, Essen- Bochum and Essen - Dortmund are significantly different. The corresponding p-values are 0.1367, 0.5558 and 0.3670 respectively. Hence for the above pairs, we fail

to reject the null hypothesis. On contrary, we reject the null hypothesis for the other pairs which means we have enough evidence that the rental price of Duisburg - Bochum, Duisburg - Dortmund and Essen - Duisburg are significantly different and the p-values are lower than the significance level.

The main the concern of this method is that the type-I error increases as the number of simultaneous pairwise comparisons increases. Hence, to avoid the type-I error we apply Bonferroni adjustments in order to decrease the type-I error.

4.4.1 Bonferroni Adjustments

The Bonferroni's method helps with controlling the type-I error. Hence, we apply this method on the simultaneous pairwise t-test analysis and the summary is presented in Table 5.

Table 5: Adjusted p-values by Bonferroni method for simultaneous pairwise comparisons

	Bochum	Dortmund	Duisburg
Dortmund	**0.820**	-	-
Duisburg	**0.218**	0.002	-
Essen	**1.000**	**1.000**	0.046

After applying Bonferroni's adjustments we can see that the p-values increased for all the pairs. According to Table 5, the number of pairs which are significantly different decreases to 2 and as a result of an adjusted type-I error, the mean rental price of Duisburg - Bochum is also not convincing to be different with regards to the p-value of 0.2182. Hence we can summarize that, we reject the null hypothesis for the pairs: Duisburg - Dortmund and Essen - Duisburg which means we have enough evidence that the mean rental price of these pairs are significantly different and the p-value are lower than the significance level, whereas we fail to reject the null hypothesis for the pairs: Dortmund- Bochum, Essen- Bochum, Essen - Dortmund and Duisburg - Bochum which means we do not have evidence that the mean rental price of these pairs are significantly different and the p-values are more than the significance level.

5 Summary

The data set subject to the analysis of this report was compiled by the lecturers of the course Introductory Case Study at TU Dortmund University in winter semester 2021. The data set included three variables, *ID* which corresponds the record number, *sqmPrice* indicates the rental price of the property which is measured in square meter and finally *regio2* which indicates to which city the property belongs to. The data set contains 200 observations and no missing values. The aim of the project was initially to find if the rental prices differ among the four regions, followed by finding pairwise differences between the rental prices and adjusting the test results using Bonferroni method.

In the course of analysis, initially we described all the statistical methods which are required for performing the tests i.e. Hypothesis testing, p-value and level of significance, ANOVA test and pairwise T-test followed by Bonferroni correction. Before we began to conduct our tests we validated the assumptions (the variance of the distributions in the population are equal, the samples collected from population are independent of each other and the dependent variable is normally distributed in each group) which are required for the tests. The global test (one-way ANOVA test) applied in the first part and the test statistics resulted p-value of 0.0035 which is lower than the significance level which revealed that not all the means are equal and there exists atleast one particular region whose mean rental price is different from the others. Simultaneous pairwise t-test was later conducted to find statistically different means. In the first run, for which no type-I error adjustment was fulfilled, the test of three pairs (Dortmund - Bochum, Essen - Bochum and Essen - Dortmund) did not show convincing evidences for different rental price means. The simultaneous pairwise test increases the chance of committing type-I error, therefore the Bonferroni method was applied in the second run. As a result, the adjusted p-value failed to reject the null hypothesis for one additional pair (Duisburg - Bochum). To conclude, the four pairs Dortmund - Bochum, Essen - Bochum, Essen - Dortmund and Duisburg - Bochum have no significant difference among the means of the rental prices whereas the rest pairs (Duisburg - Dortmund and Essen - Duisburg) do have significant differences. Although the bonferroni adjustment has helped in controlling type-I error, it has raised chance of type-II error which are not taken into account in this project.

For further investigations, the data set can be extended to additional samples so that it provides better interpretation. By extending sample size, we can get more precise and accurate inference about the population and gives us more data to support or reject

our conclusions. It may also be noted here that we are considering only one factor i.e. region/city, to understand the rental prices, it would be of great interest to include more factors which may affect the rental price for consistency of the test outcomes.

Bibliography

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Taylor and Francis Group, London, NewYork, 2018.

Clay Ford : Understanding the Q-Q Plots. Understanding q-q plots. URL `https://data.library.virginia.edu/understanding-q-q-plots/`.

Immobilienscout24. *Real estate data base*. Immobilien Scout GmbH, Germany, 2020. URL `https://www.immobilienscout24.de/`.

Kaggle.com. *Apartment rental offers in Germany*. N/A, Germany. URL `https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany`.

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

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

Rasch et al. *Applied statistics : theory and problem solutions with R*. John Wiley and Sons, USA, 2020.

Appendix

A Additional tables

Table 6: Information about different significance level

alpha value	p value	Significance level
0.1	$p<0.1$	statistically low significant
0.05	$p<0.05$	statistically moderate significant
0.01	$p<0.01$	statistically significant
0.001	$p<0.001$	statistically highly significant