TU Dortmund

Introductory Case Studies

Project 2: Comparison of rental prices in the Ruhr area

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

Author: Tadeo Hepperle

Group number: 28

Group members: Dhanya Zacharias, Imene Kolli, Vanlal Peka, Sara Zhara, Tadeo Hepperle

December 8, 2021

Contents

1 Introduction

Everyone needs to have a place to live. That is why rent and housing prices affect almost everyone and play a huge role in ones life. Data suggests, that the average German spends XXXX percent of their income on rent and mortgage (source: XXXXX). This makes local rental prices an important factor in deciding where to live, especially in times where real estate prices kept rising since 20XXX (source: XXXXX). In this project we will take a look at rental prices in the 4 largest cities of the ruhr area (Dortmund, Duisburg, Bochum and Essen) and analyze whether or not rental prices differ between those regions. For this, data from the web platform Immobilienscout24 was taken. Immobilienscout24 brings landlords and real estate firms together with potential tenants and provides transparency in terms of prices. For each of the 4 cities, 50 properties were sampled. An Analysis of Variance (ANOVA) was run on this data, to see if any of the cities differ in the mean reantal price. Result: XXXXXX. After that, we conducted 6 bonferroni-corrected two group ANOVAs for all possible pairs of cities, to find out if there are significant differences between the rental price of two designated places. This yielded the result, that XXXXXXXXXXXXXXXXXXXXXXXXX. First, in Section 3 an overview is given on how the data was collected and its quality is asessed. Also we present the goals of the project. In Section 3 the statistical and computational methods will be discussed, in particular the concept of hypothesis testing and how an Analysis of Variance works. Section 4 displays the data analysis results, where we find out which differences in rental prices between cities turned out to be significant. Finally Section 5 gives a brief summary and highlights potential implications of the findings for people that might want to rent a property in the Ruhr area.

2 Problem statement

We analyzed a dataset containing the rental prices of 200 properties in the Ruhr area to find out if there are differences in the mean rental prices between Dortmund, Duisburg, Bochum and Essen.

2.1 Description of the dataset

The data used in our analysis is a subset of a dataset from kaggle (source: XXXXXXX). It was originally scraped from www.Immobilienscout24.de, an online real estate marketplace on February 20, 2020. Immobilienscout24 earns money by ... It is one of the largest real estate marketplaces in germany according to XXXXXXXXXX with a revenue of XXXXXXX in XXXXXXyear. The dataset consists of rental prices for 200 properties in the Ruhr area. The data was randomly sampled in such a way that we have exactly 50 properties for each of the large cities Dortmund, Bochum, Duisburg and Essen. The respective city is the indepentent variable in the context of this study. The rental price in euros per square meter (shortcode "price") resembles a metric depentant variable. The data was gathered in a purely observational way, which introduces a couple of biases. First of all we do not know if prices from Immobilienscout24 are representative for the entire market. For example it could be, that there is a portion of rental contracts that have been running for years or decades and therefore may have lower rent then newly issued contracts. Increasing Rent is a common occurance when tentants change, according to XXXXXXXXXX. Also we do not know anything about the type of properties and the position with respect to the city center. We can just assume the random sampling kept those factors somewhat constant for all 4 cities, even though we do not even know if the original dataset our sample comes from may have been biased in the first place. In Addition to that it is unknown to us how balconies, roof slopes and gardens have been factored into the calculation of the rental price in Euros per square meter. The rental price relates to the net rent: service charges come on top of that. Unfortunately we cannot even make the assumption that net rent and service charges are independent from each other, as it is common practice to advertise a property by stating a lower net rent while inscresing certain service charges. There is no missing data, but having just 50 objects per group is not a lot. The randomness of the sampling alone can be responsible for some variation between the groups. Therefore the quality of the dataset is quite weak and the results should not be overinterpreted. Despite that, the averages over the 50 properties for each city should make a rough estimation quite possible.

2.2 Objectives of the report

The main objective of the report is, to find out if the mean rental price differes between the 4 cities Dortmund, Bochum, Duisburg and Essen. For this we conduct an ANOVA

which yields if there are any significant differences at all, but does not tell, between which cities they are. The Anova is basically asking if a substantial amount of variance in the data can be linked to differences in the means of the 4 groups. After this we want to figure out, if any pair of cities has a significant difference in their mean rental prices. For those pairwise comparisons we also run an ANOVA with just the two data vectors of the resepctive two cities involved. The alpha level for those two group anovas is bonferroni corrected to not accumulate an alpha error. With this report we hope to find out if the differences in the data just come from random variation or can be actually linked to the city as a factor. This could in turn help people make decisions about moving and provide a better understanding of the real estate market in the Ruhr area.

3 Statistical Methods

First a brief overview of hypothesis testing as a method for finding significant effects is presented. After that the used testing-methods and how to check their requirements is explained.

3.1 The concept of hypothesis testing

A lot of times we want to use gathered data to infer a distribution of the underlying distribution the data was taken from. The problem is, that our sample is just a random fraction of the population and therefore statistics like mean and standard deviation might differ from the true parameters one would be able to observe if the entire population was known. To adress this problem hypothesis testing is used. The basic idea is, that we have a null hypothesis (H_0) and an alternative hypothesis (H_1). They are mutually exclusive (Eid et al., 2017, p. 218). For example a null hypothesis (H_0) could be that the mean of a population is equal to 10. Then the alternative hypothesis (H_1) is that the mean is unequal to zero.

$$H_0 : \mu = 10 \qquad H_1 : \mu \neq 10$$

Now some standardized rules need to be introduced, to decide if a sample is probably from a population where H_0 holds (H_0 gets accepted), or if it is more likely that H_1

reflects the real population (H_0 gets rejected). For this a Test statistic T is calculated from the data. Also we decide a value for α, usually $\alpha = 0.05$ is chosen. Now we say that we reject the null hypothesis, if it would be very unlikely to observe data with statistic T if H_0 was true. In Detail: We reject H_0 if the probability of observing T or an even extreme value is less or equal to α under the assumtion that H_0 is true. To calculate this probability some distribution of T is assumed for the case that H_0 is true, depending on what kind of statistic T is used. So α is the probability of committing a type I error if H_0 is true. That is rejecting the null hypothesis, even though it is correct (Eid et al., 2017, p. 222). Conversely the mistake of not rejecting the null hypothesis, even though it is not correct is called type II error. The probability of observing a value of T or an even more extreme one under the assumtion that H_0 is true is called $p - value$. The range of values for T where H_0 is rejected is the rejection region. A value for the test statistic that lies right on the border of the rejection region is called critical value or T_{krit}. Comparing T_{krit} with the actual T of the data, provides a means of rejecting or accepting H_0: If T is more extreme than T_{krit} the null hypothesis is rejected. The entire process of hypothesis testing can be summarized by 4 steps:

1. Define H_0, H_1 and α.
2. Define how the test statistic T will be calculated, and derive the rejection region and T_{krit} from the assumed distribution of T under the assumption that H_0 holds.
3. Collect data and calculate T from the sample.
4. Reject or Accept H_0 based on T and T_{krit}.

Another way of telling whether or not to reject H_0 is by comparing the $p - value$ to α: if $p \leq \alpha$ H_0 shall be rejected, as in this case T is more extreme than T_{krit}.

3.2 One-way ANOVA

A one-way ANOVA (short for: One-way analysis of variance) is a statistical test that is used to determine if there are differences between the means of k samples. For this the Variance of the samples is decomposed, hence the name (Eid et al., 2017, p. 392). A one-way ANOVA for independent groups requires a metric random variable X and $n = \sum_{j=1}^{k} n_j$ realizations. Those belong to k distinct groups, where n_j denotes the number of datapoints in group j. Now, three different square sums can be defined, QS_{within}, $QS_{between}$ and QS_{total} such that $QS_{within} + QS_{between} = QS_{total}$ (Eid et al., 2017, p. 397). Here x_{mj} is the m-th datapoint of the k-th group, $\overline{x}$ represents the mean

of all datapoints and $\overline{x}_j$ is the mean of all datapoints from group j:

$$QS_{total} = \sum_{j=1}^{k} \sum_{m=1}^{n_j} (x_{mj} - \overline{x})^2$$

$$QS_{within} = \sum_{j=1}^{k} \sum_{m=1}^{n_j} (x_{mj} - \overline{x}_j)^2$$

$$QS_{between} = \sum_{j=1}^{k} \sum_{m=1}^{n_j} (\overline{x}_j - \overline{x})^2$$

The idea is that, in case there are substantial differences between the group means, the $QS_{between}$ will be larger in relation to QS_{within}, than if the group means would be the same and differences stem only from random variation. Therefore the test statistic F has to reflect this relationship. First mean square sums MQS_{within} and $MQS_{between}$ are calculated from QS_{within} and $QS_{between}$ (Eid et al., 2017, p. 397). For the following we assume equal sample sizes:

$$MQS_{within} = \frac{QS_{within}}{n-k}$$

$$MQS_{between} = \frac{QS_{between}}{k-1}$$

Then the resulting test statistic F can be calculated. A larger F signifies that there are differences in the means of the groups.

$$F = \frac{MQS_{between}}{MQS_{within}}$$

This F statistic follows an $F(df_{between}, df_{within})$ distribution with $df_{between} = k - 1$ and $df_{within} = n - k$, in case the null hypothesis is true. Therefore we can calculate the critical value F_{krit} as the $1 - \alpha$-percentile of the $F(df_{between}, df_{within})$ distribution and compare it to our actual F-value received from the sample data.

Requirements:

When an ANOVA is conducted for 2 groups with e

Bibliography

Michael Eid, Mario Gollwitzer, and Manfred Schmitt. *Statistik und Forschungsmethoden*. Beltz, 2017.

Appendix

A Additional figures