AdTrans 2023-1-PL01-KA220-HED-000158917

BASICS OF STATISTICS

Overview of basic concepts

Lucie Vavrova

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

Descriptive statistics

Statistics = provides the ability to interpret what is happening around us

Probability vs. statistics

Probability theory studies what will happen in the future with what probability. Statistics, on the other hand, tracks and studies what has happened.

Random variable

A random variable is a quantity that can take on different values with some probability when measured repeatedly (experimented).

Mean and average

The mean is a probability quantity. It is denoted by the Greek letter μ (μí). It shows the middle (average) value of a random variable.

For example, the mean value of a die roll is 3.5. This means that if you were to roll the die repeatedly, the average of your rolls would approach 3.5 as the number of attempts increased.

The mean is a statistical quantity. It shows the representative value of a group of numbers. It is denoted by a letter with a bar above it (x̄). The mean is an estimate of the middle value. It is calculated as the sum of the given numbers divided by the number of these numbers.

E.g.: $\bar{x}$ {3, 4, 9, 20, 22}= $\frac{3+4+9+20+22}{5}$= $\frac{58}{5}$= 11.6

In general: $\bar{x}$ {x_1, x_2, … x_n}=$\frac{\sum_{i=1}^{n} xi}{n}$

Median

The median is a statistical quantity. It is the middle value of a number series, which is created by arranging all the numbers from smallest to largest.

E.g.: median {3, 4, 9, 20, 22} = 9

Sometimes it is appropriate to use the median instead of the average. Why? Because the median value can better characterize a statistical sample. For example, imagine a country in which three people earn 20,000 CZK per month and one person earns 100,000 CZK per month. The average salary is then 40,000 CZK per month. Based on this value, someone might think that the majority of the population of this country is very rich. However, the lost information would be that three quarters of the people of this country earn much less (20,000 CZK). The median salary in this country is 20,000 CZK/month. This value better describes the salary situation.

Mode

The mode is a statistical quantity. It is the value that occurs most frequently in a given group of numbers.

E.g.: mode {1, 2, 8, 29, 29} = 29

Variance and sample variance

Variance is a probability quantity. It is a characteristic of the variability of a random variable. The higher the variance, the further apart the values of the random variable will be. Variance is denoted by σ^2 (sigma squared).

The sample variance is a statistical quantity. It is an estimate of the variance based on measured values. The sample variance is denoted by ^{2} and is calculated as follows:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Standard deviation and sample standard deviation

The standard deviation is a probability quantity. It is the square root of the variance, so it is also a characteristic of the variability of a random variable. The higher the standard deviation, the further apart the values of the random variable will be. The standard deviation is denoted by σ.

The sample standard deviation is a statistical quantity. It is the square root of the sample variance. It is an estimate of the standard deviation based on measured values. The sample standard deviation is denoted by and is calculated as follows:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

For example: $\bar{x}$ {4, 5, 6} = 5, with {4, 5, 6} = 0.82, and on the other hand $\bar{x}$ {1, 5, 9} = 5, with {1, 5, 9} = 3.27. We see that both sets of numbers have the same mean of 5. However, the second set is more scattered (deviates more from the mean). This fact is seen in its higher standard deviation.

Random variables

Random variable

A random variable is a quantity that can take on different values with some probability when measured repeatedly (experimented).

Discrete random variable

A discrete random variable is a random variable that can only take on a finite number of values.

For example: the value that can be rolled on a dice is a discrete random variable. There are only 6 possible outcomes, a finite number.

Continuous random variable

A continuous random variable is a random variable that can take on an infinite number of values.

Probability distribution

A probability distribution is a scheme that assigns to each possible event or interval of events the probability with which that event will occur. A probability distribution is related to some random variable – either discrete or continuous.

For example: if we roll a dice, we have a discrete random variable Dice (K). This random variable has the following distribution (the letter P denotes probability):

P (K=1) = 1/6 or 16.67%

P (K=2) = 1/6 or 16.67%

P (K=3) = 1/6 or 16.67%

P (K=4) = 1/6 or 16.67%

P (K=5) = 1/6 or 16.67%

P (K=6) = 1/6 or 16.67%

Probability density of a continuous random variable

To express the probability distribution of a continuous random variable, it is appropriate to use the so-called probability density. The probability density of a continuous random variable is defined by the function (f). For this function, the probability that a given random variable will take on a value in a certain interval (a, b) is equal to the integrals (i.e. the area under the curve) of the function f on this interval.

A probability density graph characterizes a given random variable in such a way that we are able to determine the probability with which the random variable falls into a certain interval. On an interval where the value of the density function is high, the corresponding random variable has a higher probability of falling into it.

Normal distribution and central limit theorem

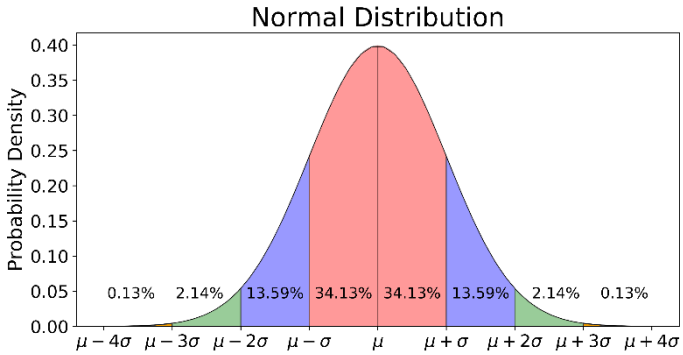
Normal distribution

The so-called normal distribution is one that has the following density function:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ is the mean (average) value and σ is the standard deviation. These two values can be chosen. Thus, there are infinitely many different normal distributions. The graph of this function (the density of this distribution) is shown in Figure 1.

Figure 1.



This shape is also known as a Gaussian curve. Why is the normal distribution important? One reason is that many things in the real world have a normal distribution.

For example: A random variable that describes the probability of an adult's height has a normal distribution. The probability that a person will have a height close to the mean value is higher than the probability that they will be very above average or very below average. The average height of women in the Czech Republic is 168 cm. The random variable "Woman's height" therefore has a normal distribution with a mean value μ=168.

Central limit theorem

Central Limit Theorem - If we want to express its content as simply as possible, we can say that the sum of many random variables converges to a normal distribution. There are several necessary assumptions, but for our purposes it is enough to know that these assumptions are mostly met. What does convergence to a normal distribution mean? If our random variable is the sum of many random variables, then the probability density (graph) of this sum becomes more and more similar to the density (graph) of the normal distribution as the number of random variables added increases.

The central limit theorem states that as we increase the number of random variables, their sum behaves like a normally distributed random variable. The properties and applications of the normal distribution can therefore be applied to sums of different random variables. We will see later that

this is a very important conclusion, because the normal distribution (or an approximate normal distribution) is an important assumption in many statistical tests and models.

Correlation (Pearson's correlation coefficient)

The correlation coefficient is a measure of the linear relationship between two variables. It is usually denoted by ρ (rho). The correlation coefficient between two variables X and Y is calculated as follows:

$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$,

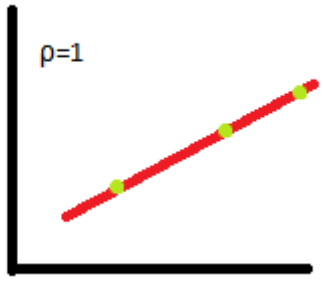
cov(x,y) =

$$\frac{1}{N-1}\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})$$

where cov(x,y) is the covariance of x and y, σ_y is the standard deviation of x and ρ_y is the standard deviation of y.

If there is an exact positive linear relationship between two variables (i.e. if they lie on the same straight line - see Figure 1), then the correlation is 1. We say that there is a perfect positive correlation between the variables. The green points in the figure show all observations with values of the first characteristic (x-axis) and values of the second characteristic (y-axis). The red line is their connection.

Figure 1.

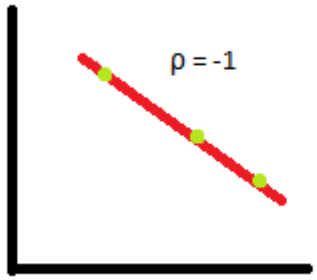


If there is a not-quite-exact, but still positive linear relationship between the variables (see Figure 2), the correlation coefficient is in the interval (0,1). We say that there is a positive correlation between the variables (the variables correlate with each other).

Figure 2.

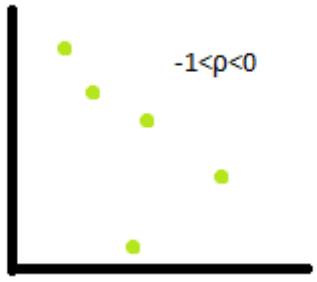
If there is a perfect negative linear relationship between the variables (see Figure 3 – the points lie on a straight line with a negative slope), the correlation coefficient is -1. We say that there is a perfect negative correlation between the variables.

Figure 3.



If there is a not-quite-exact, but still negative linear relationship between the variables (see Figure 4), the correlation coefficient is in the interval (-1,0). We say that there is a negative correlation between the variables (the variables are negatively correlated with each other).

Figure 4.



Correlation is not causation (influence)

One of the most important aspects of the correlation coefficient is that the size of this number indicates the observed relationship, but it says nothing about causality (the influence).

If there is a correlation between two variables (they correlate either positively or negatively with each other), one of the five situations described below occurs:

1) Variable X (the first variable) has an effect on variable Y (the second variable) – the causal effect of X on Y.
2) Variable Y (the second variable) has an effect on variable X (the first variable) – the causal effect of Y on X.
3) Variable X has an effect on variable Y and at the same time variable Y has an effect on variable Y - causal effect of X on Y and Y on X.
4) Variable X has no effect on variable Y and variable Y has no effect on variable X (no causality) and the correlation is just random.
5) Variable X has no effect on variable Y, variable Y has no effect on variable X, and at the same time there is a third variable Z that has an effect on both X and Y.

Confidence intervals

A confidence interval is an estimate that tells us that a given parameter is likely to lie within a certain numerical range. For example, if we have 100 observations of the height of the population of the Czech Republic and their mean is 177.5, then the 95% confidence interval for their mean height could be <173, 182>. This tells us that the height mean is within a given interval with a ninety-five percent probability. The specific interval needs to be statistically calculated using all observations. In our example, we can also see that the calculated mean is in the middle of the confidence interval for this value (the mean).

It is also possible to create narrower or wider confidence intervals. For example, in our example, the 90% confidence interval could be <175, 180>. On the other hand, the 99% confidence interval could be <171, 184>. Of course, we would like the narrowest interval with the highest probability. However, when creating intervals, we have to pay for the narrowing of the interval with a lower probability and for the higher probability with a wider interval. We need to choose the properties of the interval that suit us best.

Statistical testing

A statistical test decides whether the difference found is a statistical inaccuracy or a structural difference . The numbers 167 and 163 (large difference) would probably indicate a structural difference and the numbers 165.6 and 165.5 (small difference) would probably indicate a statistical inaccuracy. However, this is only our assumption and for an unbiased result it is necessary to perform a correct statistical test.

The statistical testing procedure is described below:

1) Determining the research question.
2) Test selection.

A statistical test appropriate for examining our research question is selected.

3) Determining the null hypothesis and alternative hypothesis. The alternative hypothesis is always the negation of the null hypothesis.

This is the traditional way of specifying a test. The result of the test is either "we reject the null hypothesis (in favor of the alternative hypothesis)" or "we do not reject the null hypothesis."

4) Assessment of test assumptions.

In this step, we need to check whether all the mathematical conditions for using our chosen test are met. Choosing a statistical test whose necessary assumptions are not met would lead to inaccurate results.

5) Determining the significance level α

The significance level is the probability limit below which the null hypothesis is rejected.

6) Performing the test

After determining the five specifications, we can move on to testing. First, we calculate the test statistic T based on the measured data. The test statistic T is calculated differently for each test. The calculation of a specific test statistic is the core of the test.

Now we can proceed in two ways:

- 1. We determine the critical region (the values of the test statistic at which the null hypothesis is rejected) and see if the value of the test statistic T falls within the critical region. If so, we reject H_0 (and accept H_A). If not, we do not reject H_0.
- 2. Based on the test statistic T, we calculate the so-called **p-value** . The rule here is: if the p-value is lower than or equal to the significance level α, we reject H_0 (and accept H_A). If the p-value is higher than the significance level α, we do not reject H_0. We will use this method for its advantages. The advantage is that we do not have to examine the critical domain. We simply feed the data into the computer and the computer spits out the p-value. This then just needs to be compared with the significance level α. This method is also used by most statistical software.

Non-rejection does not mean confirmation

Note: Non-rejection means nothing more than non-rejection. There is no information about whether the null hypothesis is true or not, and whether the alternative hypothesis is true or not. Non-rejection is therefore a very weak result. In contrast, rejection means a direct rejection of the null hypothesis, followed logically by acceptance of what is left, the alternative hypothesis. Despite this logic, the result “we do not reject H_0” is often taken as an indication that H_0 is likely true.

Random selection

Note: for the application of many tests, it is necessary to use so-called random selection when creating a sample. A random selection from a certain population is a selection in which all units have the same probability of being selected.

P-value

The p-value is defined as the probability that, given the null hypothesis, we will measure data that are more likely to contradict the null hypothesis than the data that we have measured. Equivalently, the p-value can be defined as the smallest level of significance of the test at which we will still reject the null hypothesis on the given data.

Significance level

The symbol α is usually used for the significance level. It is defined as the probability that I reject H_0 if H_0 is true. This would of course be an erroneous result. Rejecting $H_{0 \text{ when } H_0}$ is true is called a type I error or significance level. It is advisable to choose a low significance level. Someone may think to choose it as small as possible (e.g. 0.001 – 0.1%) and their test will then have the lowest probability of a type I error. This is possible – such a test will then be extremely reliable. However, the problem is that its reliability will lie in the fact that the result of almost every test will be the non-rejection of the null hypothesis, i.e. a very cautious procedure. As we already know, non-rejection of the null hypothesis does not mean either acceptance or rejection. The result of such a test will therefore be very weak and meaningless.

The level of the test expresses the probability with which the test is wrong if the null hypothesis is true. Therefore, no test is 100% accurate, which often needs to be taken into account. This corresponds to everyday life: whatever we do for whatever reason based on past observations, we can never be sure that we have interpreted the situation correctly and that we are correctly predicting what will happen.

The significance level α is usually chosen to be 0.05 (5%). This value is a compromise: on the one hand, it is low enough for the test to be reliable, and on the other hand, it is high enough for the test to be able to reject appropriately.

Test strength

The significance level α is defined as the probability of rejecting H_0 if H_0 is true. The power of the test (usually denoted β) is defined as the probability of rejecting H_0 if H_0 is not true ($H_{A \text{ is true}}$). In order to calculate the power of a given test, it is necessary to know the exact value of the relevant quantities. For example, if H_0 is $\mu_1 = \mu_2$, then to calculate the power of the test it is necessary to determine what values of μ_1 and μ_2 we take into account. H_0 does not hold whenever $\mu_1 \neq \mu_2$. This corresponds to an infinite number of possibilities for μ_1 and μ_2. Of course, the further $\mu_{1 \text{ is}}$ from μ_2, the higher the power of the test.

For the values of the relevant quantities under consideration, we naturally want to choose a test with the highest power. So if our situation meets the assumptions of two or more tests, it is appropriate to use the test that has the higher power. Very often, a test that has more assumptions has the higher power. This is one of the reasons why tests with many assumptions are used, even though there are other applicable tests with fewer assumptions. A test with more assumptions is usually better suited to the given situation.

The process of test selection is thus as follows: first, we determine the level of the test. Then, we try to find a test that has the given level, whose assumptions are met, and which has the highest

possible power. Such a test is the most suitable variant. The fact that the specific value of the test power (unlike the test level) depends on the specific values of the relevant variables does not matter. For each test, we can determine the power of the test as a function of the relevant quantities (unknown variables, e.g. in the case of a two-sample t-test μ_1 and μ_2). We then choose a test whose power is as high as possible than the power of other tests for any different values of μ_1 and μ_2 (i.e. for all cases where H_0 does not hold).

Determining a priori hypotheses

All tests and analyses that we use in research should be planned in advance. If we are investigating a topic for which more than one type of test can be used, it is correct to choose the type of test before the measurement begins. Researchers who do not do this, i.e. those who collect data and then choose a test, are subject to a phenomenon called “confirmation bias”. The problem is that they can look at the results of different tests and choose the type of test that confirms their hypothesis. Choosing a test ex post therefore gives the testers the opportunity to manipulate the results (bias).

Assumptions of statistical tests

If we have data and a research question, we need to choose an appropriate statistical test. Different tests are suitable for different situations. At the same time, each statistical test has specific assumptions. If they are met, the test works well. If they are not, the test does not give us accurate results. The most common assumptions we encounter are:

Normality of observation

Tests that assume normally distributed data are called parametric. Tests that do not make this assumption are called nonparametric. Normality is required in parametric tests because these tests and their applications (calculation of test statistics and p-values) are designed for this type of data. Whether observations are normally distributed can be tested using tests designed for this purpose.

The central limit theorem states that almost all sums of random variables converge to (resemble) a normal distribution as the number of observations increases. Therefore, it is often true that for a given test we do not need to have directly normally distributed observations. It is enough to have a large number of them. Then their sum (which is used in the test) has an approximately normal distribution and the test can be used.

Independence of observation

This assumption describes a situation where the value of each observation is independent of what value we measured in another observation. This assumption is usually not tested because it is clear from the problem statement whether the independence of observations holds or not.

Same variance of two samples

Many tests comparing two samples (e.g. the classic two-sample t-test) require that both measured samples have the same variances. Equality of variances can be tested with a specific

statistical test (e.g. Levene's test). If equality of variances holds, the test can be used. If not, it is necessary to move on to an alternative test that does not require equal variances.

Statistical tests

Test name	**Test type**	**Prerequisites**	**H_o (null hypothesis)**	**H_1 (alternative hypothesis)**	**Typical use**
Shapiro–Wilk	Normality test	None	The data has a normal distribution.	The data is not normally distributed.	Verifying the normality of the distribution
One-sample t-test	Parametric	Normal distribution	$\mu = \mu_o$	$\mu \neq \mu_o$	Comparison of the average with the reference value
Paired t-test	Parametric	Normal distribution of differences	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	Before–after measurements for the same units
Two-sample t-test	Parametric	Normality, equality of variances	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	Comparison of two independent groups
Welch's t-test	Parametric	Normality (inequality of variances)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	Comparing two groups with different variances
ANOVA	Parametric	Normality, equality of variances	$\mu_1 = \mu_2 = ... = \mu_k$	At least one diameter is different	Comparing more than two groups
Mann–Whitney U	Non-parametric	None	The distributions are the same	Distribution varies	Alternative to the two-sample t-test
Wilcoxon Signed-Rank	Non-parametric	None	The distribution of differences is symmetrical	The distribution of differences is not symmetrical	An alternative to the paired t-test
Kruskal–Wallis	Non-parametric	Sameness of variances	All distributions are the same	At least one division differs	Alternative to ANOVA
Levene's test	Test of variances	None	The variances are the same	The variances vary	Testing for homogeneity of variances
Chi-square test	Independence test	Expected frequencies > 5	The variables are independent	Variables are dependent	Dependence between categorical variables
McNemar test	Test for paired binary data	Paired observations	No change	There has been a change.	Before–after analysis for binary characters

Shapiro–Wilk Test

Situation: This is a basic test to verify the normality of the distribution.

Prerequisites: none

H_0: The data has a normal distribution

H_A: The data does not have a normal distribution

Jarque–Bera Test

- It tests the normality of the distribution similarly to the Shapiro-Wilk test and the Kolmogorov–Smirnov Test. The details of this test are the same as the Shapiro-Wilk test.

Kolmogorov–Smirnov Test

- It tests the normality of the distribution similarly to the Shapiro-Wilk Test and Jarque-Bera Test. However, it is necessary to know the mean and variance of the normal distribution, the presence of which we are testing.

One-sample z-test (Student's t-test)

Situation: we have one random sample from one distribution, **we know** the variance

Assumptions: the given sample comes from a normal distribution.

$H_0: \mu = \mu_0$ (the mean of the given distribution is equal to μ_0)

$H_A: \mu \neq \mu_0$

One-sample t-test (Student's test)

Situation: we have one random sample from one distribution, **we don't know** the variance

Assumptions: the given sample comes from a normal distribution.

$H_0: \mu = \mu_0$ (the mean of the given distribution is equal to μ_0)

$H_A: \mu \neq \mu_0$

Paired t-test

Situation: we have two equally numerous random samples from two distributions, we do not know the variance, observations from both distributions can be paired (e.g. the same respondent).

Assumptions: both random variables have a normal distribution.

$H_0: \mu_1 - \mu_2 = \mu_0$ (difference of the means of two distributions = μ_0)

$H_A: \mu_1 - \mu_2 \neq \mu_0$

Two-sample t-test

Situation: we have two random samples from two distributions. The numbers of observations from the samples may be different. Observations from both distributions cannot be paired.

Assumptions: both random variables have a normal distribution

$H_0: \mu_1 = \mu_2$ (the mean value from the first sample is the same as the mean value from the second sample)

$H_A: \mu_1 \neq \mu_2$

Example: we are testing the effect of a new drug on people's mood measured on a scale of 0-100, where 0 is the worst

Welch's t-test

- The classic two-sample t-test is used when both distributions have the same variances. If the distributions have different variances, the Welch's t-test should be used. The details of this test are the same as for the classic t-test.

ANOVA (Analysis of Variance)

Situation: we have k (k ≥ 2) random samples from k distributions. We want to find out whether these distributions have the same mean.

Assumptions: all random variables have a normal distribution, all random variables have the same variance.

H_0: $\mu_1 = \ldots = \mu_k$ (all random variables have the same mean value)

H_A: at least one mean value of one random variable differs

F-test (Fisher's test) comparison of variances

Situation: we have random samples from two distributions and we want to find out if these distributions have the same variance

Assumptions: both distributions are normally distributed

H_0: $\sigma_1^2 = \sigma_2^2$ (the variance of the first distribution is equal to the variance of the second distribution – homogeneity of variances - homoskedasticity)

H_A: $\sigma_1^2 \neq \sigma_2^2$ (heterogeneity of variances - heteroskedasticity)

Levene's test

Situation: we have k (k ≥2) random samples from k distributions. We want to find out if these distributions have the same variance.

Prerequisites: none

H_0: $\sigma_1^2 = \ldots = \sigma_k^2$ (variances of all distributions are equal)

H_A: at least one distribution has a different variance

Bartlett's test

- It tests for equality of variances in a similar way to Levene's test. The details of this test are the same as those of Levene's test.

Mann-Whitney U Test (Wilcoxon rank-sum test)

Situation: we have 2 random samples from two distributions. The random variables from which the random samples come do not have a normal distribution. The number of observations from each random variable is low (e.g. less than 25) and their sum does not have a normal distribution. Therefore, we cannot use the t-test. We want to test whether the two distributions are identical.

Prerequisites: none

H_0: random samples come from the same distributions (same distribution function)

H_A: random samples come from different distributions (different distribution functions)

Wilcoxon Signed-Rank Test

Situation: we have 2 equally large random samples from two distributions. Observations from both distributions can be paired (e.g. same respondent). The random variables from which the random samples come do not have a normal distribution. The number of observations from each random variable is low (e.g. less than 25) and their sum does not have a normal distribution. Therefore, we cannot use a paired t-test. We want to test whether the two distributions are identical.

Prerequisites:

H_0: the distributions from which the random samples come are the same

H_A: the distributions from which the random samples come are not the same

Kruskal–Wallis Test (one-way ANOVA on ranks)

Situation: we have k ($k \geq 2$) random samples from k distributions. Each random sample contains a small number of observations. The random samples do not come from a normal distribution. The distributions have approximately the same variance. We want to test whether all distributions, including medians, are the same.

Assumptions: the variances of the random samples are approximately the same.

H_0: all distributions from which random samples come are the same

H_A: at least one distribution from which one random sample comes is different

Mood's Median Test

Situation: Situation: we have k ($k \geq 2$) random samples from k distributions. Each random sample contains a small number of observations. The random samples do not come from a normal distribution. The distributions do not have the same variance. We want to test whether the medians of all distributions are the same.

Prerequisites:

H_0: the medians of all distributions are the same

HA : there is at least one distribution $_{\text{whose}}$ median differs

One-sample Kolmogorov–Smirnov test (KS Test)

Situation: we have a random sample and we want to test whether it comes from a certain (reference) distribution we specify.

Prerequisites: none

H_0: The random sample comes from a specified reference distribution

H_A: The random sample does not come from the specified reference distribution

Two-Sample Kolmogorov–Smirnov Test (KS Test)

Situation: we have 2 random samples and we want to test whether they come from the same distribution.

Assumptions: the considered joint distribution is continuous

H_0: random samples come from the same distribution

H_A: random samples do not come from the same distribution

Chi-Square Test of Independence

Situation: We have two categorical traits within a population. We want to test whether the occurrence of these traits is independent.

Assumptions: there are specific assumptions for the pivot table used in the test (expected frequency count for each cell in the table)

H_0: two characteristics are independent (no relationship)

H_A: two characteristics are dependent (there is a relationship)

McNemar's test

Situation: we have a population in which we monitor the occurrence of a certain trait. We then monitor the occurrence of the trait even after a change in another trait (the value of another characteristic changes, but the observations before the change and after the change can be paired).

Prerequisites: the values before changing the second character and after changing the second character can be paired.

H_0: the two observed characteristics are independent

H_A: the two observed characters are dependent

Chapter 8 – Regression Analysis

What is regression (regression analysis) for?

The regression analysis method was created to examine the influence of several independent variables (also regressors or explanatory variables) on one dependent (explained) variable.

Two types of regression analysis

1. **linear regression –** The premise of this method is the continuity of the dependent variable, i.e. the fact that it can take on an infinite number of values. The linear regression method is in the vast majority of cases performed using the so-called **OLS approximation** .

2. Nonlinear regression - If the dependent variable takes on only a few discrete values, it is necessary to use **nonlinear regression** .

It is usually done using software (e.g. SPSS)

When interpreting results, we start with the p-value. If it is less than or equal to 0.05, the independent variable is significant, meaning it has an effect on the dependent variable. If the p-value is greater than 0.05, then the variable is insignificant and has no effect on the dependent variable.

Next, we will interpret the Coefficient column. If the variable is insignificant, then it has no effect on the dependent variable and we do not need to examine the coefficient. **For significant independent variables, the size of the coefficient is important. It tells us how many units the dependent variable will increase when the given independent variable increases by one (by one relevant unit) and the other independent variables remain fixed.**

Constant in linear regression

What does the p-value and coefficient of the variable Constant mean? The p-value again shows whether this variable is significant. The coefficient then (in the case of the significance of the constant) shows us the so-called basic value of the dependent variable (intercept). This is the value of the dependent variable in the event that all independent variables (except the constant) are zero.

Dummy variables

A dummy variable is a variable that only takes on the values 0 and 1. Zero indicates a state where the variable is not “active” and one indicates a state where the variable is “active” (here: mice have overpopulated). The interpretation of the coefficient associated with a dummy variable is the same as the interpretation of the coefficients of other variables; the coefficient shows how much the dependent variable changes when the independent (here dummy) variable increases by one. An increase in the variable by one corresponds to a change in state from inactive (not valid) to active (valid).

Regression equation

The general form of the regression equation is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i,\ \ i = 1, \dots, n \quad (1)$$

$$\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$$

y_i is the value of the dependent variable, x_{i1}, …, x_{ip} are p values of the independent variables, ε_i is the measured error (residual), n is the number of observations, β_0 is a constant and β_1, …, β_p are the coefficients belonging to the individual independent variables. The method we use is called linear regression precisely because all the coefficients β_1, …, β_p are in linear form (the first power). The values of y_i and x_{i1}, …, x_{ip} are measured. The role of the regression model is to estimate the value of the coefficients β_0, β_1, …, β_p and present them in a table.

R-squared coefficient (R2) squared, coefficient of determination)

The output of a linear OLS regression model is usually the R2 value (coefficient of determination). It is a coefficient that measures the tightness (accuracy, quality) of the model.

The higher the coefficient value, the higher the tightness (accuracy, quality) of the model. Often, percentages (0 to 100 percent) are mentioned instead of decimal places. In general, a good model is one that has an $R^{2\ coefficient}$ higher than 0.5 (50 percent).

However, even lower $R^{2\ values}$ do not necessarily mean disaster. Low $R^{2\ values}$ may simply indicate that there are many unobserved independent variables that are related to the dependent variable. In such a case, it would be very inaccurate to predict future values of the dependent variable based on knowledge of the independent variables; precisely because we cannot determine the influence of unobserved factors. However, this model can still determine the relationship between the dependent variable and the observed independent variables well.

It is true that if we add another independent variable to the model, $R^{2\ will}$ always increase. We could therefore achieve a significant increase in R^{2} by adding thousands of independent variables to the model. The $R^{2}_{coefficient}$ would be higher even if all of these independent variables were insignificant. A model with thousands of insignificant variables would obviously not be very good. For this reason, it is more appropriate to use the so-called Akaike criterion to compare two models that may have a different number of independent variables.

Akaike criteria (Akaike criteria)

Another criterion that is often included in regression output in software is the so-called Akaike criterion. Unlike $R^{2,}$ this coefficient does not have a direct absolute interpretation. It can take on negative and positive numbers of any size. A lower Akaike criterion means a better model. This criterion is suitable for comparing the quality of two or more models.

Nonlinear regression

If the dependent variable can only take on a finite number of discrete values, it is necessary to use so-called nonlinear regressions. Nonlinear regression exists in two versions depending on the transformation function used: logit or probit. In the vast majority of cases, it does not matter which function we use - both give almost identical results, because the logit function is very similar to the probit function.

Depending on the type of dependent variable, one of the three types of logistic regression must be correctly determined and used (not arbitrarily as when deciding between logit and probit).

- If the dependent variable takes on only two values (e.g. low yield and high yield), it is necessary to use logistic binary regression.
- If the dependent variable takes on more than two values where the order matters (e.g. low yield, medium yield and high yield), it is necessary to use ordinal logit (ordered logit) regression.
- If the dependent variable takes on more than two values for which it makes no sense to consider the order (e.g. green color, red color, and blue color), it is necessary to use logistic multinomial regression.

McFadden's R^2 (McFadden R-squared)

Nonlinear regression does not use the classic coefficient R^2. This can only be calculated in the case of linear regression. Instead, most statistical programs offer the McFadden coefficient R^2. Unlike the classic R^2, this coefficient can also take values that do not fall within the interval (0, 1). However, here too, the higher the coefficient, the better the model.

Akaike criterion

The Akaike criterion can be calculated in both linear and nonlinear regression. In both cases, the lower the criterion value, the better the model. This criterion is suitable only for comparing two models, as it does not have a direct absolute interpretation.

Number (percentage) of correctly predicted cases

Logistic regression offers this data, which takes on either a certain natural number (count) or values from 0% to 100% (percentage). It is calculated as follows: the measured values for the independent variables (x_{i1}, …, x_{ip}) are inserted into the regression equation calculated from the measured values. In this way, the predicted value of the dependent variable is calculated. If this back prediction is correct (corresponds to the actual value of the dependent variable for the given observation), it is counted as a correctly predicted case. If it is not correct, it is a incorrectly predicted case. This is done for all observations. The criterion then shows the number (or percentage) of correctly predicted cases. This criterion therefore tests whether our calculated model works for our data. Of course, the higher the percentage of correctness, the better our model.