

Advanced Statistics

Paolo Coletti – A.Y. 2010/11 – Free University of Bolzano Bozen

Table of Contents

| | | |
|-----------|--|-----------|
| 1. | Statistical inference | 2 |
| 1.1 | Population and sampling | 2 |
| 2. | Data organization | 4 |
| 2.1 | Variable's measure | 4 |
| 2.2 | SPSS | 4 |
| 2.3 | Data description | 5 |
| 3. | Statistical tests | 7 |
| 3.1 | Example | 7 |
| 3.2 | Null and alternative hypothesis..... | 11 |
| 3.3 | Type I and type II error | 11 |
| 3.4 | Significance..... | 12 |
| 3.5 | Accept and reject..... | 12 |
| 3.6 | Tails and critical regions | 13 |
| 3.7 | Parametric and non-parametric test | 15 |
| 3.8 | Prerequisites..... | 15 |
| 4. | Tests | 16 |
| 4.1 | Student's t test for one variable | 16 |
| 4.2 | Student's t test for two populations..... | 16 |
| 4.3 | Student's t test for paired data | 18 |
| 4.4 | F test | 19 |
| 4.5 | One-way analysis of variance (ANOVA)..... | 20 |
| 4.6 | Jarque-Bera test..... | 22 |
| 4.7 | Kolmogorov-Smirnov test..... | 22 |
| 4.8 | Sign test | 23 |
| 4.9 | Mann-Whitney (Wilcoxon rank sum) test | 26 |
| 4.10 | Wilcoxon signed rank test | 28 |
| 4.11 | Kruskal-Wallis test | 30 |
| 4.12 | Pearson's correlation coefficient..... | 32 |
| 4.13 | Spearman's rank correlation coefficient..... | 34 |
| 4.14 | Multinomial experiment..... | 36 |
| 5. | Which test to use? | 41 |
| 6. | Regression model | 43 |
| 6.1 | The least squares approach | 43 |
| 6.2 | Statistical inference | 46 |
| 6.3 | Multivariate and non linear regression model | 47 |
| 6.4 | Multivariate statistical inference..... | 48 |
| 6.5 | Qualitative independent variables | 49 |
| 6.6 | Qualitative dependent variable | 50 |
| 6.7 | Problems of regression models | 51 |

1. Statistical inference

Statistic is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information. A population is a set of units (usually people, objects, transactions, etc.) that we are interested in studying. A sample is a subset of units of a population, whose elements are called cases or, when dealing with people, subjects. A statistical inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

For example, we may introduce a variable X which models the temperature at midday in January. Clearly this is a random variable, since the temperature fluctuates randomly day by day and, moreover, temperatures of the future days cannot be even determined now. However, from this random variables we have data, measurements done in the past. In statistics people deal with observations or, in other words, realizations, x_1, x_2, \dots, x_n of a random variable X . That is, each of x_i is a random variable that has the same probability distribution as its originating random variable X . It characterizes the i th performance of the stochastic experiment determined by the random variable X . Given this information, we want to characterize the distribution of X or some of its characteristics, like the expected value. In the simplest cases, we can even establish via theoretical considerations the shape of the distribution and then try to estimate from the data its parameters.

In other words, statistical inference concerns the problem of inferring properties of an unknown distribution from data generated by that distribution. The most common type of inference involves approximating the unknown distribution by choosing a distribution from a restricted family of distributions. Generally the restricted family of distributions is specified parametrically. For the temperature example we can assume that X be normally distributed with a known variance σ^2 and an expected value to be determined. Among all normal distribution with this variance we want to find the one which is the most likely candidate for having produced the finite sequence x_1, x_2, \dots, x_n of temperature observed in the past days.

Making inference about parameters of a distribution, people deal with statistic or estimates. Any function $f_n(x_1, x_2, \dots, x_n)$ of the observations is called a statistic. For example, the sample mean $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ is a common statistic, typically used to estimate the expected value. The sample variance $\bar{s}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ is another useful estimate. Being a function of random variables, a statistic is a random variable itself. Consequently we may, and will, talk about its distribution.

1.1 Population and sampling

A statistical research can analyze data from the entire population or only on a sample. The population is the set of all objects for which we want to infer information or relations. In this case, data set is complete and statistical research simply describes the situation without going on to any other objective and without using any statistical test. When data are instead available only on a sample, a subset of the population, statistical research analyses whether information and relations found on the sample can be extended on the entire population from which the sample comes from or they are valid only for that particular sample choice.

Therefore, sample choice is a very important and delicate issue in statistical researches. Many

statistical methods let us extend results (estimates or tests' results) found on the sample to the population, provided that sample is a random sample, a sample whose elements are randomly extracted from the population without any influence from the researcher, from previously taken sample's elements or from other factors. Building such a sample, however, is a difficult task since a perfectly random selection is almost a utopia. For example, any random sampling on people will necessarily include people who are unwilling to give information, who have disappeared, and who lie; these people cannot be excluded nor replaced with others, because otherwise the sample would not be random anymore. A previously random sample with excluded elements can screw the estimates: in our example, problematic people are typically old and with low education, thus unbalancing our sample in favor of young and educated subjects.

A common strategy to build a sample which behaves like a random sample is the stratified sampling. With this method, the sample is chosen respecting the proportions of the variables which are believed to be important for the analysis and which are believed to be able to influence the analysis results. For example, if we analyze people we should take care to build a sample which reflects the sex proportions, age, education and income distribution, the residence (towns, suburbs, countryside) proportions, etc. In this way, the sample will reflect exactly the population at least for what the considered variables are concerned. Whenever a person is not available for answering, we replace him with another one with the same variables' values. Obviously these variables must be chosen with care and with a look at previous studies on the same topic, balancing their number since too few variables will create a badly stratified sample, while a too many will make sample creation and people's substitution very difficult.

Another aspect is the sample size. Obviously, the larger the sample the better. However, this relation is not direct, i.e. doubling the sample size does not yield doubly better results. The relation in many statistical tests goes approximately like \sqrt{n} , which means that we need to quadruple the sample size to get doubly better results. In any case, it is much more important to have a random or well stratified sample rather than a numerous sample. Quality is much better than quantity.

A common mistake related to sample size is supposing that it should be proportional to the population. This is, at least for all the test analyzed in this book, false: for large populations, test's *results depend only on the absolute size and not on the proportion*. Thus, a population of 1000 with a sample of 20 does not yield better results compared to a population of 5000 with a sample of 20.

2. Data organization

2.1 Variable's measure

In a statistical research we face basically three types of variables:

- scale variables are fully numerical variables with an intrinsic mathematical meaning. For example, a temperature or a length are scale variables since they are numeric and any mathematical operation on these variables makes sense. Also a count is a numerical variable, even though it has restrictions (cannot be negative and is integer), because it makes sense to perform mathematical operations on it. However, numerical codes such as phone numbers or identification codes are not scale variables even though they seem numeric, since no mathematical operation makes sense on them and the number is used only as a code;
- nominal variables represent categories such as sex, nationality, degree course, plant's type. These variables divide the population into groups. Variables such as identification number are nominal since they divide the sample into categories, even though each case is a single category;
- ordinal variables are a midway between nominal and scale variables. They represent categories which do not have a mathematical meaning (even though many times categories are identified by numbers, such as in a questionnaire's answers) but these categories have an ordinal meaning, i.e. can be put in order. Typical examples are questionnaire's answers such that "very bad", "bad", "good", "very good", or some time issues such as "first year", "second year", "third year".

Ordinal and nominal variables, often referred to as categorical, are used in SPSS in two ways: as variables by themselves, such as in multinomial experiments (see section 4.14) and, more often, as a way to split the sample into groups to perform tests on two or more populations, such as Student T test for two populations (see section 4.2), ANOVA (see section 4.5), Mann-Whitney (see section 4.9) and Kruskal-Wallis (see section 4.11).

2.1.1 Grouping

It is also a common procedure to degrade scale variable them to ordinal variables, arbitrarily fixing intervals or bins and grouping the cases into their appropriate bin. For example, an age variable expressed in years can be degraded to an ordinal variable dividing the subjects into "young", up to 25, "adult", from 26 to 50, "old", from 51 to 70, "very old", 71 and over. The new variables that we obtain are suited for different statistical tests which open up more possibilities. However, any *grouping procedure reduces the information that we have introducing arbitrary decisions in the data and possible biases*. For example, if our sample has a very large count for people of age 26, the previous arbitrary choice of 25 as a limit for "young" group has put many people, who are more similar to 25 years old people rather than to 50 years old people, into the "adult" group.

SPSS: Transform → Recode into Different Variables

2.2 SPSS

SPSS means Statistical Package for Social Sciences and it is a program to organize statistical data and perform statistical research. SPSS organizes data in a sheet called Data View which is a database table, more or less like Excel's tables. Each case is represented by an horizontal lines and identified very often by the first variable which is an ID number. Variables instead use vertical

columns. Unlike Excel and like database tables, SPSS data table is extremely well structured and each variable has a lot of features. These features are found in Variable View sheet:

- Name: feel free to use any meaningful name, but without special characters and without spaces. When data have many variables it is a good idea to indicate names as v_ followed by a number (it will be possible to indicate a human readable name later).
- Type: numeric is the most common type. String should be used only for completely free text, while categorical variables should be numeric with a number corresponding to each category (it will be possible to indicate a human readable name later); a common mistake is using a string variable for a categorical variable, which has the impact that SPSS will refuse to perform certain operations with that variable.
- Width and decimals
- Label: this is the variable's label which will appear in charts and tables instead of the variable's name.
- Values: this feature represents the association between values and categories. It is used for categorical variables, which, as said before, should use numbers for each category. In this field values' labels can be assigned and in charts and tables these labels will appear instead of numbers. Obviously scale variables should not receive values' labels.
- Missing: whenever a variable's value is unknown for a certain case a special numeric code should be used, traditionally a negative number (if the variable has only positive numbers) or the largest possible number such as 9999. If this number is inserted here among the missing values, SPSS will simply ignore that case whenever that variable is involved in any operation. It is also possible, in Data View, to clear the cell completely and SPSS will indicate it with a dot which is a system-missing number (same effect as missing value).
- Measure: variable's measure must be carefully indicated, since it will have implications on which operations may be done on the variable.

SPSS has four basic menus:

- Transform: this menu lets us build new variables or modify existing ones, usually working on a case by case base, thus performing only horizontal operations. Very useful are commands:
 - compute, which build a new variable, typically scale, using mathematical operations;
 - recode, which build a new variable, typically categorical, using recoding;
- Data: this menu lets you rearrange your data in a more global way. Very useful are the commands:
 - split, splits the file using a nominal or ordinal variable in such a way to be able to analyze it automatically in groups;
 - select, lets us filter out some temporarily undesired cases;
 - weight, lets us weight the cases using a variable whenever each case represents several cases with the same data (all the statistics will use a new sample size based on the weights);
- Analyze: this menu is the core of SPSS with all the statistical tests and models;
- Graphs: this is the menu to create charts.

2.3 Data description

SPSS offers a variety of numerical and graphical tools to quickly describe data. The choice of the tool depends on variable's measure:

- **SPSS: Analyze → Descriptive Statistics → Frequencies**
Frequencies is indicated as a description for a single categorical variable, while for a scale variable frequency table becomes too long and full of single cases. However, it is always a good idea to start any statistical research with frequencies for every variable, including scale ones, to spot out data entry mistakes which are very common in statistical data.
- **SPSS: Graphs → Chart Builder → Pie/Polar**
Pie chart is indicated as a graph for a single nominal and ordinal variable.
- **SPSS: Graphs → Chart Builder → Bar**
Pie charts are indicated as a graph for a single categorical variable. Using colors and three-dimensionality they work also for two or even three nominal and ordinal variables.
- **SPSS: Analyze → Descriptive Statistics → Descriptives**
Descriptive statistics (mean, median, standard deviation, minimum, maximum, range, skewness, kurtosis) is indicated as a description for a single scale variable and usually it does not make sense for categorical variables.
- **SPSS: Graphs → Chart Builder → Histogram**
Histogram is indicated as a graph for a single scale variable. Variable values are grouped into bins for the variable representation. The choice of binning influences the histogram.
- **SPSS: Graphs → Chart Builder → Boxplot**
Boxplot is indicated as a graph for a single scale variable. The central line represents the median and the box represents the central 50% of the variable's distribution on the sample. Boxplots may be used also to compare the values of a scale variable by groups of a categorical variable.
- **SPSS: Analyze → Descriptive Statistics → Crosstabs**
Contingency table (see section 4.14.2) is indicated as a description for two categorical variables.
- **SPSS: Analyze → Compare Means → Means**
Means comparison is a way to compare the means of a scale variable for groups of a categorical variable, usually followed by Student's T test or ANOVA (see sections 4.2 and 4.5).
- **SPSS: Analyze → Correlate → Bivariate**
Bivariate correlation (see sections 4.12 and 4.13) is a description for the linear relation between two scale variables.
- **SPSS: Graphs → Chart Builder → Scatter/Dot**
Scatterplot is indicated as a graph for two scale variables.

3. Statistical tests

Statistical tests are inference tools which are able to tell us the probability with which results obtained on the sample can be extended to the population.

Every statistical test has these features:

- the null hypothesis H_0 and its contradictory hypothesis H_1 . It is very important that these hypotheses are built without looking at the sample;
- a sample of observations x_1, x_2, \dots, x_n and a population, to which we want to extend information and relations found on the sample;
- prerequisites, special assumptions which are necessary to perform the test. Among these assumptions there is always, even though we will not repeat it every time, that data must come from a random sample;
- the statistic $f_n(x_1, x_2, \dots, x_n)$, a function calculated on the data, whose value determines the result of the test;
- a statistic's distribution from which we can obtain the test's significance. When using statistical computer programs, significance is automatically provided by the program next to the statistic's value;
- significance, also called p-value, from which we can deduct whether accepting or rejecting null hypothesis.

3.1 Example

In order to show all the elements of a statistical test, we run through a very simple example and we will, later, analyze the theoretical aspects of all the test's steps.

We want to study the age of Internet users. Age is a random variable for which we do not have any idea of the distribution nor its parameters. However, we make the hypothesis that age is a continuous random variable with an expected value. We want to check whether the expected value is 35 years or not. We formulate the test's hypotheses:

- $H_0: E(\text{age}) = 35$
- $H_1: E(\text{age}) \neq 35$

Of this random variable the only thing we know are the observations on a random sample of 100 users, which are: 25; 26; 27; 28; 29; 30; 31; 30; 33; 34; 35; 36; 37; 38; 30; 30; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 20; 54; 55; 56; 57; 20; 20; 20; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35; 36; 37; 35.

Now we calculate the age average on the sample, $\overline{\text{age}}_{100} = 36.2$, which is an estimation for the expected value. We compare this result with the 35 of the H_0 hypothesis and we find a difference of +1.2. At this point, we ask ourselves whether this difference is large enough, implying that the expected value is not 35 and thus H_0 must be rejected, or is small and can be caused by an unlucky choice of the sample and therefore H_0 must be accepted.

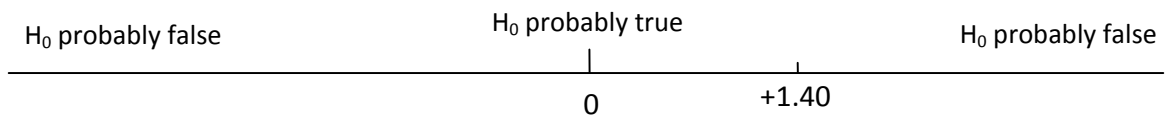
This conclusion in a statistical research cannot be drawn from a subjective decision whether the difference is large or small. It is taken using formal arguments and therefore we must rely on this

statistic function:

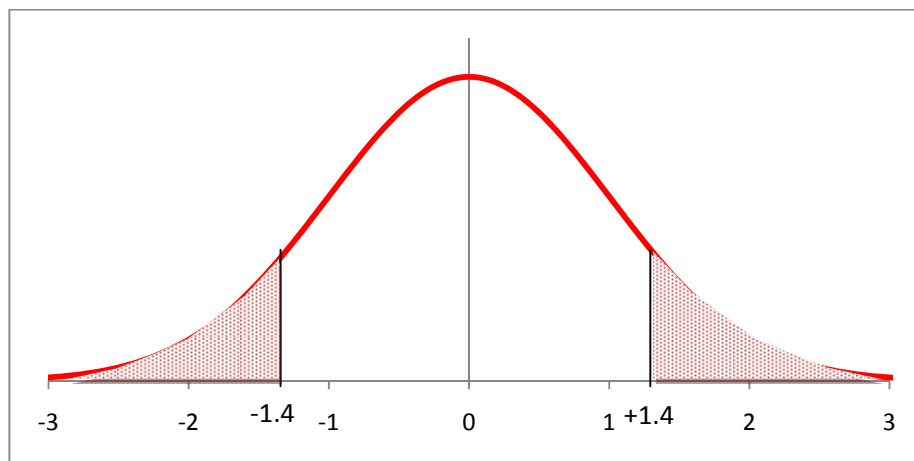
$$\frac{\bar{age}_n - \text{hypothesized expected value}}{\sqrt{\text{sample variance}/n}}$$

It is noteworthy to look at this statistic numerator. When the average of age is very close to the hypothesized expected value, the statistic will be close to 0. On the other hand, when the two quantities are very different, compared to the sample standard deviation, the statistic is very large. Statistic's value is also influenced by the sample's number of elements n : the larger is the sample, the larger the statistic.

Summing up, considering that in our case the sample standard deviation is 8.57, statistic is +1.40. The situation is therefore



At this point we ask ourselves where is exactly the point which separated the “H₀ true” zone from the “H₀ false” zone. To find it out, we calculate the probability to obtain an even worse result than the one we have got now. The meaning of “worse” in this situation is “worse for H₀”, therefore any result larger than +1.40 or smaller than -1.40. We use central limit theorem which guarantees us that, if n is large enough and if the hypothesized expected value is the real expected value of the distribution (i.e. H₀ is true), our statistic has a standard normal distribution. In fact, the only reason why we have built this statistic instead of using directly the difference at the numerator is because we know the statistic's distribution. Therefore we know that the probability of getting a value larger than +1.40 or smaller than -1.40 is¹ 16%. This value is called significance or p-value.

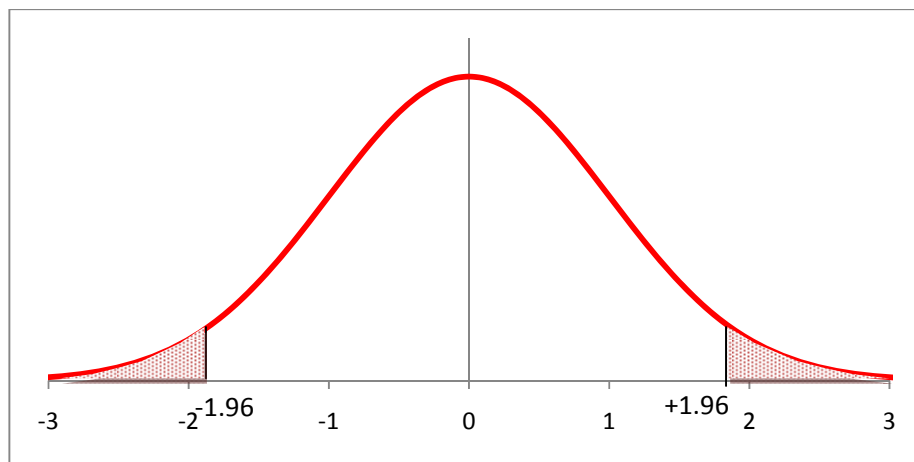


If significance is large it means that, supposing H₀ to be true and taking another random sample, the probability of obtaining a worse result is large and therefore the result that we have

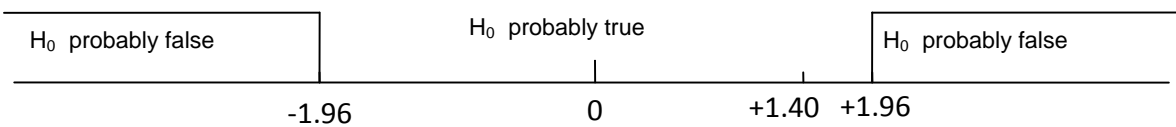
¹ This value can be calculated through normal distribution tables or using English Microsoft Excel function NORMDIST(-1.4;0;1;TRUE) which gives the area under the normal distribution on the left of -1.4, equal to 8%. Area on the right of +1.4 is obviously the same.

obtained can be considered to be really close to 0, something which pushes us to accept the idea that H_0 is true. When, instead, significance is small, it means that if we suppose that H_0 is true we have a small probability of getting such a bad result, something which pushes us to believe that H_0 be false. In the example's situation we have a significance of 16%, which usually is considered large (the chosen cut point is typically 5%) and therefore we accept H_0 .

A slightly different method, which yields to the same result, is fixing the cut point a priori, let's say 5%, and finding the corresponding critical value after which the statistic is in the rejection region. In our case, considering two areas of 2.5% on the left and on the right side, the critical value for a standard normal distribution is² 1.96.



At this point the situation is



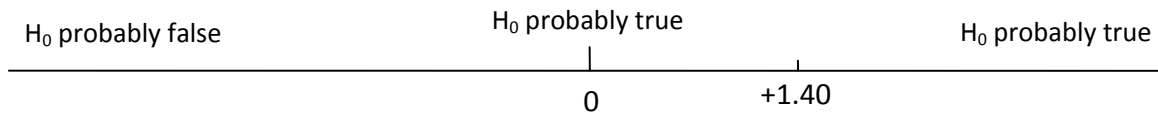
The first method gives us an immediate and straightforward answer and in fact is the one typically used by computer programs. The second method instead is more suited for one-tailed tests and is easier to apply if a computer is not available.

An example of a one-tailed test is the situation when we want to check whether the expected value of the age is smaller or larger than 35. We write the hypotheses in this way:

- $H_0: E(\text{age}) \geq 35$
- $H_1: E(\text{age}) < 35$

In this case, the difference of +1.2 between sample average and 35, since it is positive, leads us to strongly believe that H_0 be true. In fact, now the situation of the statistic is different from before, i.e.

² This value can be calculated through normal distribution tables or using English Microsoft Excel NORMINV(2.5%;0;1) which gives the critical value -1.96 for which the area under the normal distribution on the left of it is 2.5%. Due to symmetry of the distribution, critical value on the right is obviously +1.96.

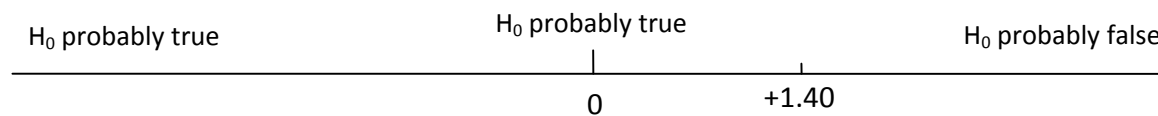


In fact here we do not have any doubt since the statistic value falls right in the middle of the “H₀ true” area.

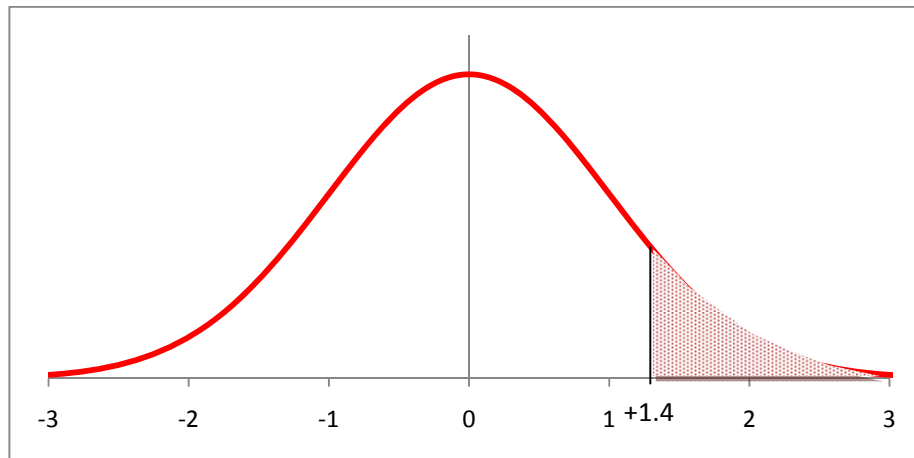
Writing however the hypotheses in this way:

- H₀: E(age) ≤ 35
- H₁: E(age) > 35

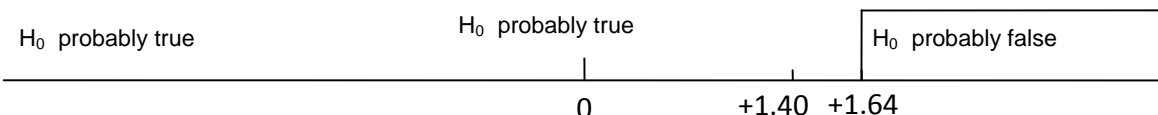
In this case, the situation of the statistic is



and here we have the same problem of determining whether +1.40 is close to 0 or far away from it. As usual, to determine it we have two methods. The first one calculates the probability of getting a worse result, where worse means “worse for H₀”. In this situation, however, a worse result is larger than +1.40, while results smaller than -1.40 are strongly in favor of H₀. The statistic is always distributed like a standard normal, under the hypothesis that H₀ be true,



and the area, thus the significance, is 8%. Using the second method the critical value is not 1.96 anymore, but³ 1.64. The critical region is larger than before, since now the 5% is all concentrated on the left part.



³ This value can be calculated through normal distribution tables or using English Microsoft Excel NORMINV(5%;0;1) which gives the critical value -1.64 for which the area under the normal distribution on the left of it is 5%. Due to symmetry of the distribution, critical value on the right is obviously +1.64.

3.2 Null and alternative hypothesis

The heart of a statistical test is null hypothesis H_0 , which represents the information that we are officially trying to extend from the sample to the population⁴. It is important that the null hypothesis gives us additional information, since we need to suppose it to be true and use its information to know the statistic's distribution. If, in the previous example, the null hypothesis had not given us the additional information that the real expected value be 35, we could not use the fact that that statistic function be normally distributed. Therefore, the null hypothesis must always contain an equality, while strict inequalities are reserved for H_1 . When the test is one-tailed, we write the null hypothesis in the form of a non-strict inequality such as $E(\text{age}) \geq 35$ for practical purposes, but theoretically we should write the equality $E(\text{age}) = 35$ and simply not take into account the $E(\text{age}) > 35$ possibility.

For example, usable hypotheses are " $E(\xi) = 35$ " or "distribution of ξ is exponential" or even " ξ and ζ are independent". On the other hand, hypotheses such as " $E(\xi) \neq 35$ " or "distribution of ξ is not exponential" are not acceptable. Also " ξ and ζ are dependent" is not acceptable, since it does not provide us with any information on how they are dependent.

Together with null hypothesis we always write alternative hypothesis H_1 , which is the logical contradiction of null hypothesis.

3.3 Type I and type II error

Once the statistic is calculated we must take a decision: accept H_0 or reject H_0 . When H_0 is rejected, we face one of the two following situations:

- null hypothesis is really false and we rejected it: very good;
- null hypothesis is really true and we rejected it: we committed a type I error.

If we accept H_0 , we face one of the two following situations:

- null hypothesis is really false and we accepted it: we committed a type II error;
- null hypothesis is really true and we rejected it: very good.

There are two different types of errors that we may commit when taking a decision after a statistical test and it would be wonderful if we could reduce at the same time the probability of committing both errors. Unfortunately, *the only method to reduce the probability to commit both errors is taking a large sample*, hopefully taking the entire population. This thing is clearly not feasible in many situations where gathering data is very expensive.

There is a method to reduce probability of committing a type I error: rejecting only in the situations where H_0 is evidently false. In this way a type I error will be very rare since we are rejecting in very few situations. Unfortunately, if we reject with parsimony, we will accept very often and this means committing a lot of type II errors. Same thing if, vice versa, we reject too much: we will commit very few type II errors but many type I errors.

Thus, we must decide which error is the more severe one and try to concentrate on reducing the probability of committing it. Every statistical research concentrates on type I errors, trying to

⁴ As we will see later, it is instead H_1 the information that we will be able to extend to the population, while, unfortunately, it is never possible to extend H_0 .

reduce the probability of committing them under a significance level usually 5% or 1%. Using an example drawn from a juridical situation:

- H_0 : suspect deserves 0 years of prison (suspect is innocent)
- H_1 : suspect deserves > 0 years of prison (suspect is guilty)

In this case, a type I error means condemning an innocent, while a type II error means an innocent verdict for a guilty. It is common belief that in this case a type I error should be avoided at all cost, while a type II error be acceptable.

The reason why statistical tests concentrate their attention on avoiding type I errors derived from the historical development of science which takes as correct the current theories (H_0) and tries to minimize the error to destroy, by mistake, a well-established theory in favor of new theories (H_1). It is therefore a conservative approach. For example:

- H_0 : hearth pumps blood
- H_1 : hearth does not pump blood

A type I error in this case would be a disaster since it would mean rejecting the correct hypothesis that blood is pumped by hearth, giving us no other clue since H_1 carries only a negative information.

3.4 Significance

Significance or p-value is the *probability of committing a type I error*. This probability is calculated assuming that H_0 be true and comparing the value of the statistic that we calculate on our sample's data with the statistic's distribution. A small significance means that if we reject we have only a small probability of committing a mistake, and therefore we will reject. A large significance means that if we reject we are facing a large probability of committing a mistake, and therefore we will accept H_0 .

Another equivalent definition for the significance is the *probability of obtaining, taking another random sample, an equal or worse statistic's value under the hypothesis that H_0 be true*. A small significance means that the statistic's value is really bad and therefore we will reject H_0 . A large significance means that the statistic's value is much better than what we expected and therefore we will accept H_0 .

Since we try to minimize type I errors, we will fix a very small significance level under which null hypothesis is rejected, usually 5% or 1%. In this way, probability of a type I error is low and when we reject we are almost sure that H_0 is really false.

Confidence is equal to 100% minus the significance.

3.5 Accept and reject

At the end of the statistical test we must decide whether accepting or rejecting:

- if significance is above the significance level (usually 5% or 1%), we accept H_0 ;
- if significance is below the significance level, we reject H_0 .

It is very important to underline the fact that when we reject we are almost sure that H_0 is false, since we are keeping type I errors under a small significance level. However, when we accept we may not say that H_0 be true, since we do not have any estimation on type II errors. Therefore, rejecting is a sure thing, while *accepting is a "no answer" and from it we are not allowed to draw any conclusion*.

This approach is called falsification, since *we are only able to falsify H_0 and never to prove it*. If we need to prove that H_0 be true, we must rewrite the hypotheses and put the information we want to extend to the population in the H_1 hypothesis instead, perform the test again and hope to reject.

Another important effect that we must underline is the sample size. When sample size is extremely small, data are almost random and probability of committing type I error is very large. Therefore significance is very large and, using the traditional small significance levels, we will accept. Therefore *a statistical test with few data automatically accepts everything*, since it does not have enough data to prove that H_0 be false. Again, accepting must never imply that H_0 be true.

3.5.1 Paradox

Using the falsification approach we can, through a smart choice of null hypotheses, accept two contradictory null hypotheses. Using as sample the one of the previous example and formulating the hypotheses

- $H_0: E(\text{age}) = 35$
- $H_1: E(\text{age}) \neq 35$

we accept $E(\text{age}) = 35$ with a significance level of 5%. Using instead these hypotheses

- $H_0: E(\text{age}) = 36$
- $H_1: E(\text{age}) \neq 36$

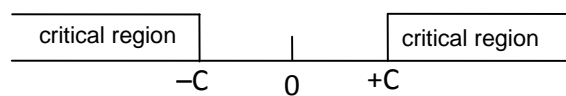
we accept $E(\text{age}) = 36$ with a significance level of 5%. We have thus accepted two hypotheses which say different and contradictory things. This is only an apparent paradox, since accepting does not mean that they are true but only that they might be true. Therefore, for the population from which our sample is extracted, the expected value might be 35 or 36 (or many other close values, such as 35.3, 36.5, 37, etc.). This is due to a relatively small size of the sample; if we increase the sample size, the interval of values for which we accept would decrease.

3.6 Tails and critical regions

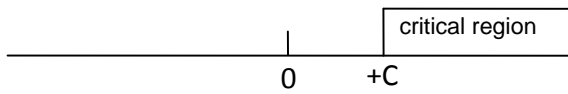
Statistical tests where the null hypothesis contains an equality and alternative hypothesis a not equality are two-tailed tests. Statistical tests where the null hypothesis contains a non-strict inequality and alternative hypothesis a strict inequality are one-tailed tests, such as

- $H_0: E(\text{age}) \geq 35$
- $H_1: E(\text{age}) < 35$

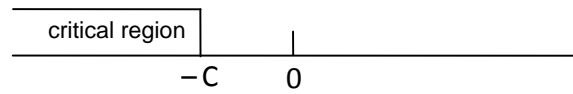
The name of these tests comes from the number of critical regions. A critical region is an area for which null hypothesis is rejected when the statistic's value falls in that area, according to the second method that we have seen in the example 3.1. The number of critical regions, which usually are far away from the center of the distribution and therefore are called tails, determines the name of the test two-tailed or one-tailed.



two-tailed test



one-tailed test with critical region on the right



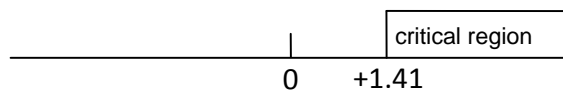
one-tailed test with critical region on the left

The point where the critical region starts is called critical value and is usually calculated from tables of the statistic's distribution. In the two-tailed test the two regions are always symmetric, while for one-tailed test we face the problem of determining on which side is the rejection region.

In order to find where the critical region is in one-tailed tests, we try to see what happens if we have an extremely large positive value for the statistic. If such an extremely large positive value (which, being very large, is for sure in the right tail) is not in favor of null hypothesis, it means that the right tail is not in favor of null hypothesis and therefore it is the rejection region. Otherwise, if this extremely large value of the statistic is in favor of the null hypothesis, the right region is not a rejection region and the critical region is on the left. For example, we consider example 3.1

- $H_0: E(\text{age}) \leq 35$
- $H_1: E(\text{age}) > 35$

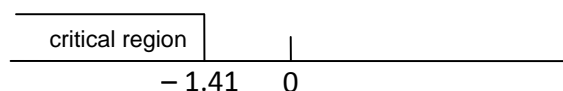
and we use the same statistic $\frac{\bar{age}_n - \text{hypothesized expected value}}{\sqrt{\text{sample variance}/n}}$. When this statistic's value is positive and extremely large, it means that the average of age is much more than the hypothesized expected value and this is a clear indication that the real expected value is much larger than 35. This is in contradiction with null hypothesis which says that expected value must be smaller or equal to 35. Therefore a positive value of the statistic, on the right tail, is contradicting null hypothesis and this means that right tail is a critical region.



Considering instead hypotheses

- $H_0: E(\text{age}) \geq 35$
- $H_1: E(\text{age}) < 35$,

when the statistic's value is positive and extremely large, it means that the average of age is much more than the hypothesized expected value and this is a clear indication that the real expected value is much larger than 35. This is exactly what the null hypothesis says. Therefore a positive value of the statistic, on the right tail, is in favor of the null hypothesis and this means that right tail is not critical region. Therefore the critical region is on the left.



Some important features to note on critical values:

- decreasing significance level implies that critical value goes away from 0. This is evident if we consider the fact that decreasing the significance level we are even more afraid of type I errors

and therefore we reject with much more care, thus reducing the rejection zone;

- the critical value of a one-tailed test is always closer to 0 than the critical value of two-tailed tests. This is because the critical tail of a one-tailed test must contain the probability that for a two-tailed test is split in two regions and therefore the zone must be larger;
- for each two-tailed test there are two corresponding one-tailed tests. One of them has the statistic's value completely on the other side of the rejection region, therefore for this one we always accept. This is the reason why using the significance method to determine whether accepting or rejecting can be misleading for one-tailed tests, since it is not evident whether the test has an obvious accept verdict or not.

3.7 Parametric and non-parametric test

There are parametric and non-parametric statistical tests. A parametric test implies that the distribution in question is known up to a parameter or several parameters. For example, it is believed that many natural phenomena are normally distributed. Estimating μ and σ^2 of the phenomenon is a parametric statistical problem, because the shape of the distribution, a normal one, is known up to these two parameters. On the other hand, non-parametric tests do not rely on any underlying assumptions about the probability distribution of the sampled population. For example, we may deal with continuous distribution without specifying its shape.

Non-parametric tests are also appropriate when the data are non-numerical in nature but can be ranked, thus becoming rank tests. For example, taste-testing foods we can say we like product A better than product B, and B better than C, but we cannot obtain exact quantitative values for the respective measurements. Other examples are tests where the statistic is not calculated on sample's values but on the relative positions of the values in their set.

3.8 Prerequisites

Each test, especially parametric ones, may have prerequisites which are necessary for the statistic to be distributed in a known way (and thus for us to calculate its significance).

A typical prerequisite for many parametric tests is that the sample comes from a certain distribution. To verify it:

- if data are not individual measures but are averages of many data, the central limit theorem guarantees us that they are approximately normally distributed;
- if data are measures of a natural phenomena, they are often affected by random errors which are normally distributed;
- we can hypothesize that data comes from a certain distribution if we have theoretical reasons to do it;
- we can plot the histogram of the data to have a hint on the original population's distribution, if the sample size is large enough;
- we can perform specific statistical tests to check the population's distribution, such as Kolmogorov-Smirnov or Jarque-Bera tests for normality.

Every test has as a prerequisite that the sample be a random sample, even though we will not indicate it.

4. Tests

4.1 Student's t test for one variable

Prerequisites: variable normally distributed (if sample variance is used).

H₀: expected value = m

Statistic:
$$\frac{\text{sample average} - m}{\sqrt{\text{population or sample variance} / n}}$$

Statistic's distribution: Student's t with $n - 1$ degrees of freedom; when $n > 30$ standard normal.

SPSS: Analyze → Compare Means → One-Sample T Test



William "Student"
Gosset
(1886-1937)

Student's t test is the one we have already seen in the example in its large sample version. It is a test which involves a single random variable and checks whether its expected value is m or not.

For example, taking $m = 32$ and a sample of 10 elements: 25; 26; 27; 28; 29; 30; 30; 31; 33; 34

- H₀: E = 32
- H₁: E ≠ 32

Sample average is 29.3 and sample standard deviation is 2.91. Statistic is therefore -2.94 and its significance is⁵ 1.7%. H₀ is rejected since 1.7% is below significance level; this means that extracting another sample of 10 elements from a distribution with an expected value equal to 32, we have a very small probability of getting such bad results. We can thus say that expected value is not 32.

As we can easily see, Student's t test for one variable is exactly the test version of the average confidence interval.

4.2 Student's t test for two populations

Prerequisites: two populations A and B and the variable must be distributed normally on the two populations

H₀: expected value on population A = expected value on population B

Statistic:
$$\frac{\text{sample A average} - \text{sample B average}}{\sqrt{\frac{(n_A - 1) \text{ sample or population A variance} + (n_B - 1) \text{ sample or population B variance}}{n - 2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}}$$

Statistic's distribution: Student's t with $n - 2$ degrees of freedom; when $n > 31$ standard normal.

SPSS: Analyze → Compare Means → Means

SPSS: Analyze → Compare Means → Independent-Samples T Test

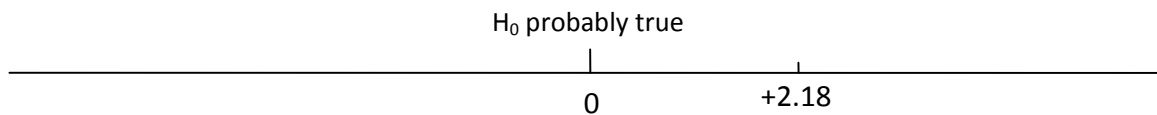
⁵ Significance can be calculated in two ways. (1) Using Student's t distribution table. (2) Using English Microsoft Excel function TDIST(2.94;9;2) which gives us the sum of the two tails areas, those on the left of -2.94 and on the right of +2.94.

This test is used whenever we have two populations and one variable calculated on this population and we want to check whether the expected value of the variable changes on the populations.

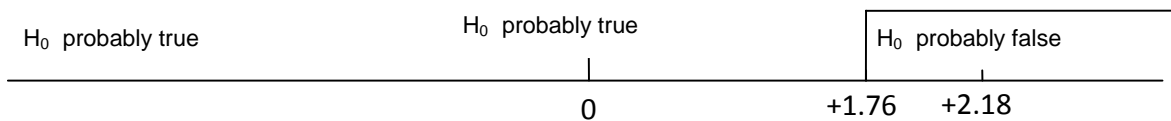
For example, we want to test

- H_0 : $E(\text{height})$ for male $\leq E(\text{height})$ for female
- H_1 : $E(\text{height})$ for male $> E(\text{height})$ for female

We take a sample of 10 males (180; 175; 160; 180; 175; 165; 185; 180; 185; 190) e 8 female (170; 175; 160; 160; 175; 165; 165; 180). We suppose that male's and female's heights are normally distributed with the same variance. Male's sample average is 177.5 while for female it is 168.75. Statistic's value is +2.18. Since it is one-tailed test we draw the graph to have a clear idea where does the statistic fall.



If the statistic were extremely large, this would be strongly in contradiction with H_0 and therefore rejection region in on the right.



Critical value for one-tailed test is⁶ 1.76 and therefore we reject. Using instead the significance method, after having checked that statistic does not fall on the “ H_0 true” area, we get⁷ a significance of 2.2% and therefore we reject, meaning that male population has an expected height significantly larger than female population.

⁶ Critical value can be calculated in two ways. (1) Using English Microsoft Excel function TINV(5%;16), which gives us the critical value for the two-tailed test, therefore probability split into 2.5% and 2.5%. For one-tailed test probability must be doubled, TINV(10%;16), since in this way it would be split into 5% and 5%. (2) Using Student's t distribution table.

⁷ Significance can be calculated in four ways. (1) Using one of the statistical t tests (Zweistichproben t test) in the Data Analysis toolpak in Microsoft Excel, choosing among known variances (in this case populations' variances have to be indicated explicitly), equal and unknown, different and unknown (in these latter two cases populations' variances are estimated from sample data automatically by Excel), which gives us statistic's value and its significance. (2) Using English Microsoft Excel function TTEST which gives us the significance directly from the data, choosing type=2 if we suppose equal variances or type=3 if we suppose different variances. (3) Using English Microsoft Excel function TDIST(2.18;16;1) which gives us the area of one of the two tails. (4) Using Student's t distribution table.

4.3 Student's t test for paired data

Prerequisites: two variables ζ and ξ on the same population and $\zeta - \xi$ must be normally distributed

H₀: E(ζ) = E(ξ), which means E(ζ) - E(ξ) = 0

The test can also be performed with null hypothesis: **H₀: E(ζ) - E(ξ) = m**

Statistic: we use $\zeta - \xi$ as variable and we perform Student's t test for one variable

Statistic's distribution: same as Student's t test for one variable

SPSS: Analyze → Compare Means → Paired-Samples T Test

This test is used whenever we have a single population and two variables calculated on this population and we want to check whether the expected value of these two variables is different.

For example, we want to test whether population's income in a country has changed. We take a sample of 10 people's income and then we take the same 10 subjects' income the next year

| Income 2010 (thousands €) | Income 2011 (thousands €) | Difference 2010 - 2011 |
|------------------------------|------------------------------|---------------------------|
| 20 | 21 | -1 |
| 23 | 23 | 0 |
| 34 | 36 | -2 |
| 53 | 50 | +3 |
| 43 | 40 | +3 |
| 45 | 44 | +1 |
| 36 | 12 | +24 |
| 76 | 80 | -4 |
| 44 | 45 | -1 |
| 12 | 15 | -3 |

Two things are very important here. The subjects must be exactly the same, no replacement is clearly possible. When calculating the difference the sign is important, so it is a good idea to clearly write what is subtracted from what, especially for one-tailed tests.

Hypotheses are:

- H₀: E(income) for 2010 - E(income) for 2011 = 0
- H₁: E(income) for 2010 - E(income) for 2011 ≠ 0

Sample average for the difference is +2.0 and sample standard deviation is 8.07. Statistic is +0.78 with⁸ a significance of 45.3%. H₀ is thus accepted. This does not mean that income has remained the same, but simply that our data are not able to prove that it has changed.

⁸ Significance can be calculated in four ways. (1) With the Student's t test for one variable formula using $m=0$. (2) Using English Microsoft Excel function TTEST which gives us the significance directly from the data, choosing type=1. (3) Using the statistical t test (Zweistichproben t test bei abhängig Stichproben) in the Data Analysis toolpak in Microsoft Excel, which gives us statistic's value and its significance. (4) Using Student's t distribution table.

4.4 F test

Prerequisites: two populations A and B and the variable must be distributed normally on the two populations

H₀: Var on population A = Var on population B

Statistic: sample A variance/sample B variance

Statistic distribution: Fisher's F distribution with $n_A - 1$ and $n_B - 1$ degrees of freedom



George Waddel Snedecor (1881-1974)



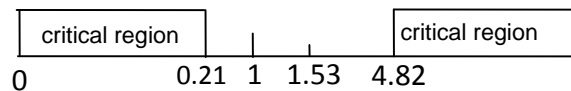
Ronald Fisher (1890-1962)

The name of this test was coined by Snedecor in honor of Fisher. It checks the variances of two populations. It is interesting to note that, unlike all the other tests, statistic's best value for H₀ is 1 and not 0. Since F distribution is only positive and not symmetric, special care must be taken into account on the statistic's position when calculating the significance since it can be misleading. In particular, the opposing statistic's value is not the opposite but the reciprocal.

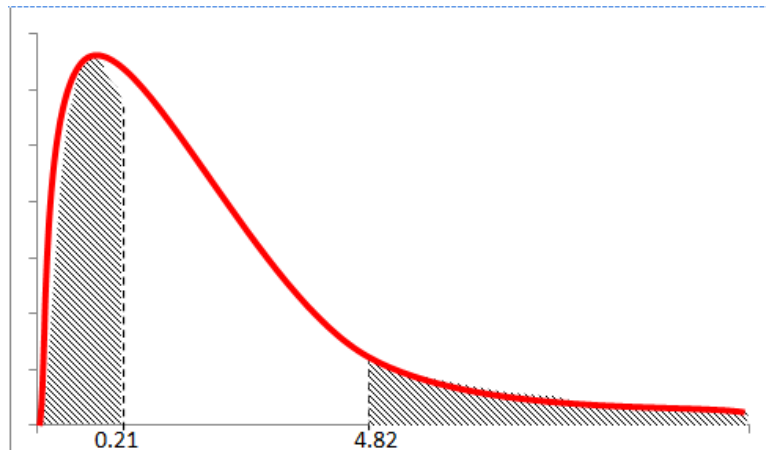
For example, supposing that height for male and female is normally distributed, we test

- H₀: Var(height)for male = Var(height) for female
- H₁: Var(height) for male ≠ Var(height) for female.

We use the previous sample and we get a sample variance of 84.7 for male and 55.4 for female. Statistic is thus 1.53. Degrees of freedom are 9 and 7. The two critical values are⁹ 4.82 and $\frac{1}{4.82} = 0.21$ and therefore we accept H₀. Using the significance method, after having checked that the statistic is on the right of 1, we get an area of 29% for the right part and therefore significance is 58%.



⁹ Calculation of critical values or significance can be done in different ways. (1) Using the statistical F test (Zwei-Stichproben F-Test) in the Data Analysis toolpak in Microsoft Excel, which gives us statistic's value and its significance. (2) Using English Microsoft Excel function FTEST which gives us the significance directly from the data. This method can be misleading when statistic is on the left of 1. (3) Using English Microsoft Excel function FDIST(1.53;9;7) which gives us the area of the right tail. (4) Using English Microsoft Excel function FINV(2.5%;9;7) and 1/FINV(2.5%;9;7) to get the two critical values. Pay attention to the inverted degrees of freedom for the second calculation. (5) Using F distribution table, which however usually provides only the critical values.



4.5 One-way analysis of variance (ANOVA)

Prerequisites: k populations, variable is normally distributed on every population with the same variance

H₀: expected value of the variable is the same on all populations

$$\text{Statistic: } F = \frac{\text{Variance Between}}{\text{Variance Within}} = \frac{1}{k-1} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2 / \frac{1}{n-k} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2$$

Statistic distribution: Fisher's F distribution with degrees of freedom equal to $k-1$ and $n-k$

SPSS: Analyze → Compare Means → Means

SPSS: Analyze → Compare Means → One-Way ANOVA

This test is the equivalent of Student's t test for two unpaired populations when the populations are more than two. We note that if only one population has an expected value different from the other, the test rejects. Therefore, a rejection guarantees us that populations do not have the same expected value but does not tell us which populations are different and how. Optimal statistic value for H₀ is 0 and, since F distribution has only positive values, this test has only the right tail.

For example, we have heights for young (180; 170; 150; 160; 170), adults (170; 160; 165) and old (155; 160; 160; 165; 175; 165) and we want to check

- H₀: E(height) for young = E(height) for adults = E(height) for old
- H₁: at least one of the E(height) is different from the others

We suppose heights are normally distributed with the same variance. From data we get a sample average of 166 for young, 165 for adults and 163.3 for old. Now we ask ourselves whether these differences are large enough to say that there are differences among populations' expected values or not.

The origins of the analysis of variance lie in the splitting of sample's variance in this way¹⁰:

¹⁰ Variance = $\frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n)^2 = \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_n)^2 = \frac{1}{n} \sum_{g=1}^k \left\{ \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g} + \bar{x}_{n_g} - \bar{x}_n)^2 \right\} = \frac{1}{n} \sum_{g=1}^k \left\{ \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2 + (\bar{x}_{n_g} - \bar{x}_n)^2 + 2(x_j - \bar{x}_{n_g})(\bar{x}_{n_g} - \bar{x}_n) \right\} = \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2 + \frac{1}{n} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2 + \frac{2}{n} \sum_{g=1}^k \left\{ (\bar{x}_{n_g} - \bar{x}_n) \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g}) \right\}$

$$\text{Variance} = \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2 + \frac{1}{n} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2$$

We now define the sample's variance between groups as a measure of the averages variations between values of different groups

$$\text{variance between} = \frac{1}{k-1} \frac{1}{n} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2$$

and the sample's variance within group as a measure of the variations among values of the same group

$$\text{variance within} = \frac{1}{n-k} \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2$$

The idea behind the test is to compare these two measures: if the variance between is much larger than the variance within, it means that at least one population is significantly different from the others, while if the variance between is not large compared to the variance within it means that variations due to a change in the population have the same size as variations due to other effects and can thus be considered negligible. Simplifying the $1/n$ the statistic is

$$F = \frac{\text{variance between}}{\text{variance within}} = \frac{1}{k-1} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2 \bigg/ \frac{1}{n-k} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2$$

which is distributed as a Fisher's F distribution with $k-1$ and $n-k$ degrees of freedom. Rejection region is clearly on the right, since that area is the one where Variance Between is much larger than Variance Within.

Going back to our example, statistic's value is $F = \frac{19.88/2}{803.33/11} = 0.136$ with degrees of freedom 2 and 11 and a significance¹¹ of 87.4% and therefore we accept.

$$\bar{x}_n) \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g}) \} = \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2 + \frac{1}{n} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2 + \frac{2}{n} \sum_{g=1}^k (\bar{x}_{n_g} - \bar{x}_n) (n_g \bar{x}_{n_g} - n_g \bar{x}_{n_g}) = \frac{1}{n} \sum_{g=1}^k \sum_{j=1}^{n_g} (x_j - \bar{x}_{n_g})^2 + \frac{1}{n} \sum_{g=1}^k n_g (\bar{x}_{n_g} - \bar{x}_n)^2 + 0$$

¹¹ Significance can be calculated in different ways. (1) Using the one-way ANOVA (ANOVA: Einfaktorielle Varianzanalyse) in the Data Analysis toolpak in Microsoft Excel, which gives us statistic's value and its significance. (2) Using English Microsoft Excel function FDIST(0.136;2;12) which gives us the area of the right tail. (3) Using F distribution table, which usually provides the right side critical values.

4.6 Jarque-Bera test

Prerequisites: none.

H₀: variable follows a normal distribution

Statistic: $\frac{n}{6} \left(\text{sample skewness}^2 + \frac{\text{sample Kurtosis}^2}{4} \right)$

Statistic distribution: Jarque-Bera distribution. When $n > 2000$, chi square distribution with 2 degrees of freedom.



Carlos Jarque



Anil Bera

This test checks whether a variable is distributed, on the population, according to a normal distribution. It uses the fact that a normal distribution has always a skewness and a Kurtosis of 0. Its statistic is clearly equal to 0 if the sample's data have a skewness and Kurtosis of 0 and increases if these measures are different from 0. The statistic is multiplied by n , meaning that if we have many data they must have display very small skewness and Kurtosis to get a low statistic's value.

Sample's skewness and sample's Kurtosis are calculated as

$$S = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2} \right)^3} \quad K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^4}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2} \right)^4} - 3.$$

4.7 Kolmogorov-Smirnov test

Prerequisites: none.

H₀: variable follows a known distribution

Statistic: $\sup_x \left| \frac{\text{number of sample data} < x}{n} - F(x) \right|$, where F is the cumulative distribution of the known r.v.

Statistic distribution: Kolmogorov distribution

SPSS: Analyze → Nonparametric Tests → One Sample



Andrey Kolmogorov
(1903-1987)



Vladimir Ivanovich
Smirnov
(1887-1974)

This is a rank test which checks whether a variable is distributed, on the population, according to a known distribution specified by the researcher. The test for each x calculates the the difference between the percentage of sample's data smaller than this x and the probability of getting a value smaller than x from the known distribution. Clearly, if sample's data are distributed according to the known distribution, these differences are very small for every x since the percentage of smaller values reflects exactly the probability of finding smaller values. The statistic is defined as the maximum, for all the x , of these differences.

For example, we want to check whether data 3; 4; 5; 8; 9; 10; 11; 11; 13; 14 come from a $N(9; 25)$ distribution. For $x = 2$, $\left| \frac{\text{number of sample data} < 2}{10} - F_{N(9;25)}(2) \right| \approx |0 - 0.05| = 0.05$; for $x = 3$, $\left| \frac{\text{number of sample data} < 3}{10} - F_{N(9;25)}(3) \right| \approx |0 - 0.08| = 0.08$; for $x = 4$, $\left| \frac{\text{number of sample data} < 4}{10} - F_{N(9;25)}(4) \right| \approx |0.1 - 0.12| = 0.02$; for $x = 5$, $\left| \frac{\text{number of sample data} < 5}{10} - F_{N(9;25)}(5) \right| \approx |0.2 - 0.16| =$

0.04 and so on. Obviously, this calculation is not done only for integer values but for all values and doing it manually is, in many cases, a very hard task. In this case, the maximum is 0.21 obtained for a value of x immediately after 11. Its significance is much larger than 5% and therefore we accept.

4.8 Sign test

Prerequisites: continuous distribution.

H₀: median is m

Statistic: outcomes on the left or on the right of m

Statistic distribution: $B(n; 50\%)$; for $n \geq 10$ $\frac{\text{statistic} \pm 0.5 - 0.5 \cdot n}{\sqrt{0.25 \cdot n}} \sim N(0; 1)$

SPSS: Analyze → Nonparametric Tests → One Sample

Sign test is a rank test which tests the central tendency of a probability distribution. It is used to decide on whether the population median equals or not the hypothesized value.

Consider the example when 8 independent observations of a random variable X having a continuous distribution are 0.78, 0.51, 3.79, 0.23, 0.77, 0.98, 0.96, 0.89. We have to decide whether the distribution median η_X is equal to 1.00. We formulate the two hypotheses:

- $H_0: \eta_X = 1.00$
- $H_1: \eta_X \neq 1.00$

If the null hypothesis is true, we expect approximately half of the measurements to fall on each side of the hypothesized median. If the alternative is true, there will be significantly more than half on one of the sides. Thus, our test statistic will be either S_- or S_+ . These two quantities denote the number of observations falling below and above 1.00. Since X was assumed to have a continuous distribution, $P(X = 1.00) = 0$. In other words, every observation falls either below or above 1.00, never hitting this value itself. Consequently, $S_- + S_+ = 8$. In practice it can be that an observation is exactly 1.00. In this situation, since this observation is strongly in favor of H_0 hypothesis, we will consider it to belong to S_- when S_+ is larger and to S_+ when S_- is larger.

Note that this choice of test statistic does not require having exact values of the observations. In fact, it is enough to know whether each observation is larger or smaller than 1.00. To the contrary, the corresponding small sample parametric test (which is the Student's t test for one variable) requires exact values in order to calculate the sample's average and variance.

Now we take S_- and consider the significance of this test. This is the probability (assuming that H_0 is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and thus supportive to the alternative hypothesis, as the actual one computed from the sample data. In our case $S_- = 7$. There are two more contradictory outcomes of the experiment: when $S_- = 8$, the case when all observations have fallen on the same side of the hypothesized median, and when $S_- = 0$. And there is a result which is as contradictory as the one we have, $S_- = 1$. Thus significance equals $P(S_- = 7) + P(S_- = 8) + P(S_- = 1) + P(S_- = 0)$.

Note that the distribution of S_- has a binomial distribution $B(8; 0.5)$. Indeed, if we suppose that H_0 is correct, having an outcome on the left of 1.00 is an event with probability 50%. And having s_- outcomes on the left of 1.00 on a total of 8 independent observations is a binomial with $p = 50\%$ and $n = 8$. Therefore, remembering that

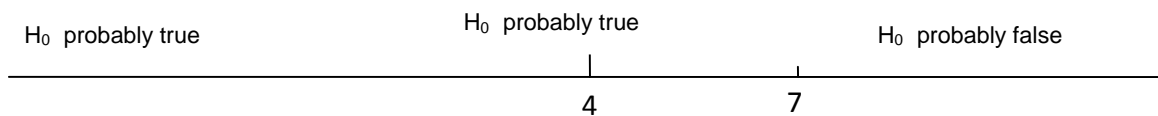
$$P(B(n; p) = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

we can calculate¹² $P\{S_- = 7\} + P\{S_- = 8\} = 0.035$. Remembering that the binomial distribution in the particular case of $p = 50\%$ is symmetric and therefore $P(S_- = 1) + P(S_- = 0) = P(S_- = 7) + P(S_- = 8)$, we get that significance is 7%. Setting a significance level of 5%, we accept null hypothesis meaning that our data are not able to support the hypothesis that median is not 1.00.

The corresponding one-tailed test is used to decide on whether the distribution median equals to the hypothesized value or falls below/exceeds it. Referring to the set of data considered above, the corresponding two mutually exclusive hypotheses read, for example:

- $H_0: \eta_X \geq 1.00$
- $H_1: \eta_X < 1.00$

As test statistic we choose S_- . In order to find out where is the rejection region, we note that when our statistic is huge the observations falling below 1 will be more numerous than the ones exceeding 1 and this is in favor with the alternative hypothesis. Thus the zone on the right is the rejection region, while the zone on the left, where S_- is small, is not a rejection region. Because $S_- = 7$, there is only one more contradictory to H_0 outcome is $S_- = 8$. Thus the significance equals $P(S_- = 7) + P(S_- = 8)$. The random variable S_- has always a binomial distribution whose probability of a success is 1/2 and we conclude that the significance is $P(B(8; 50\%) = 7) + P(B(8; 50\%) = 8) = 3.5\%$.



Thus, when H_0 is true, the probability to face an outcome as contradictory as the actually observed one or an outcome more contradictory to H_0 , equals 3.5%. Consequently, the sample data suggest that if we reject H_0 we may be wrong in only 3.5% of the cases.

Note that, as compared with the two-tailed test, now the probability of type I error is two times smaller although the sample information remains the same. This is not surprising because the one-tailed test starts from a more precise guess, it starts with the implicit hypothesis that η_X can never be larger than 0.

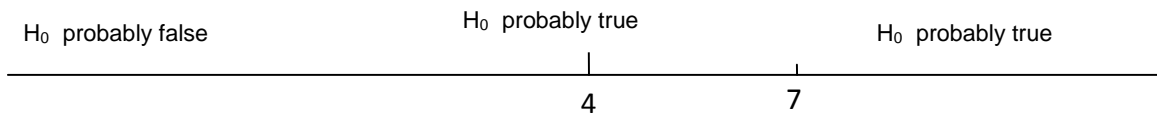
If we make the other one-tailed test instead:

- $H_0: \eta_X \leq 1.00$
- $H_1: \eta_X > 1.00$,

if we take S_- as statistic, in order to find out where is the rejection region, we note that when our statistic is huge the observations falling below 1 will be more numerous than the ones exceeding 1 and this is in favor with the null hypothesis. Therefore larger values of the statistic are all in favor of

¹² These quantities can be much easily calculated in two different ways: (1) using binomial distribution cumulative tables, which give directly $P(B(n; p) \leq x)$ and in our case $P(B(8; 50\%)=7) + P(B(8; 50\%)=8) = 100\% - P(B(8; 50\%) \leq 6)$; (2) using English Microsoft Excel function $100\% - \text{BINOMDIST}(6; 8; 50\%; \text{TRUE})$ which gives us $100\% - P(B(8; 50\%) \leq 6)$.

H_0 . Therefore the rejection region is now for small values of the statistic



Without even calculating the significance, it is evident that we must accept H_0 . In any case, the worse cases are $S_- = 6$, $S_- = 5$, $S_- = 4$, $S_- = 3$, $S_- = 2$, $S_- = 1$ and $S_- = 0$. Therefore, $P(S_- \leq 7) = P(B(8; 50\%) \leq 7) = 0.996$.

Recall that the normal distribution provides a good approximation for the binomial distribution when the sample size is large (usually $n \geq 10$). Thus, using the central limit theorem, we may use $N(0.5n; 0.25n)$ to approximate the distribution of our statistic. Using standardization

$$\frac{S - 0.5 \cdot n}{\sqrt{0.25 \cdot n}} \sim N(0; 1),$$

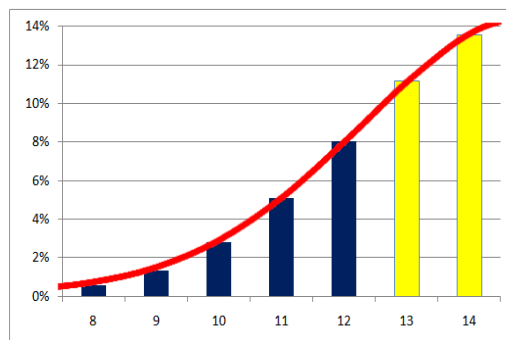
where S is our statistic S_- or S_+ . Due to technical reasons¹³ a correction of ± 0.5 is applied to the formula

$$\frac{S \pm 0.5 - 0.5 \cdot n}{\sqrt{0.25 \cdot n}} \sim N(0; 1),$$

For example, we have a sample of 30 elements with 18 elements on the left of 2.00 and 12 elements on the right of 2.00 and we want to test

- H_0 : median ≥ 2.00
- H_1 : median < 2.00 .

¹³ A technical problem which arises whenever we try to approximate a discrete distribution ($B(n; 50\%)$ in our case) with a continuous one ($N(0.5n; 0.25n)$ in our case). Discrete probability distribution does not have any probability for non integer values, while continuous one does.



Therefore we have to decide what to do with the values between 12 and 13, where the binomial distribution does not exist, however the normal distribution has a consistent probability. We take a compromise, taking for the normal approximations all the values up to 12.5. Therefore we add a $+0.5$ to the previous formula. It is always an addition whenever we are on the left tail, while it is clearly a subtraction whenever we are on the right tail and have thus a \geq sign:

$$\frac{S \pm 0.5 - 0.5 \cdot n}{\sqrt{0.25 \cdot n}} \sim N(0; 1)$$

We take as statistic S_+ . Since it is a one-tailed test we have to see where the rejection region is. Supposing a very large value for the statistic, i.e. $S_+ = 30$, this means that probably the median is much larger than the hypothesized value and this is in favor of H_0 . Therefore, rejection region is not for large statistic's value and it is on the other side, the left one. Values more or equal contradictory to H_0 are thus $S_+ \leq 12$. Using the exact calculation yields to $P(B(30; 50\%) \leq 12) \approx 18.07\%$, while using approximated calculation¹⁴ we have

$$P\left(N(0; 1) \leq \frac{12 + 0.5 - 0.5 \cdot 30}{\sqrt{0.25 \cdot 30}}\right) \approx P(N(0; 1) \leq -0.9129) \approx 18.06\%.$$

In both cases we accept, meaning that our sample data are not able to prove that H_0 be wrong.

4.9 Mann-Whitney (Wilcoxon rank sum) test

| |
|--|
| Prerequisites: the two probability distributions are continuous |
| H_0: position of distribution for population A = position of distribution for population B |
| Statistic: sum of ranks of the smaller group |
| Statistic distribution: Wilcoxon rank sum table or $\frac{t_1 - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \sim N(0; 1)$ when sample is large and tables are not available |
| Alternative statistic: $U =$ sum of ranks of the smaller group minus $n_1(n_1 + 1)/2$, where n_1 is the size of the smaller group |
| Alternative statistic distribution: Mann-Whitney table or $\frac{t_1 - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \sim N(0; 1)$ when sample is large and tables are not available |
| SPSS: Analyze → Nonparametric Tests → Independent Samples |

Suppose two independent random samples are to be used to compare two populations and we are unwilling to make assumptions about the form of the underlying population probability distributions (and therefore we cannot perform Student's t test for two populations) or we may be unable to obtain exact values of the sample measurements. If the data can be ranked in order of magnitude, the Mann-Whitney test (also called Wilcoxon rank sum test) can be used to test the hypothesis that the probabilities distributions associated with the two populations are identical.

For example, suppose six economists who work for the government and seven who work for universities are randomly selected, and each one is asked to predict next year's inflation. The objective of the study is to compare the government economists' predictions to those of the university economists. Assume the government economists have given: 3.1, 4.8, 2.3, 5.6, 0.0, 2.9. The university economists have suggested instead the following values: 4.4, 5.8, 3.9, 8.7, 6.3, 10.5, 10.8. That is, there is a random variable X equal to the next year's inflation given by a governmental economist. Asking governmental economists about their prediction, we observe independent outcomes, X_i , of X . As well, there is another random variable Y equal to the next year's inflation given by a university economist. Approaching a university economist concerning his forecast of the

¹⁴ The probability of a normal distribution can be calculated in two ways: (1) looking into a standard normal distribution table; (2) using English Microsoft Excel function NORMDIST(-2.5/SQRT(0.25*30);0;1;TRUE).

inflation rate, we observe an independent outcomes, Y_i , of this random variable. We have to decide whether X and Y have the same distributions or not, basing our decision only on the sample observations, which is the only information we have.

- H_0 : the probability distribution corresponding to the government economists' predictions of inflation rate is in the same position as the university's economists' one
- H_1 : the probability distribution corresponding to the government economists' predictions of inflation rate is in a different position as the university's economists' one

To solve this problem, we first rank all available sample observations, from the smallest (a rank of 1) to the largest (a rank of 13): 1 (0.0), 2 (2.3), 3 (2.9), 4 (3.1), 5 (3.9), 6 (4.4), 7 (4.8), 8 (5.6), 9 (5.8), 10 (6.3), 11 (8.7), 12 (10.5), 13 (10.8). The test statistic for the Mann-Whitney test is based on the totals of the ranks for each of the two samples – that is, on rank sums. If the two rank sums are nearly equal, the implication is that there is no evidence that the probability distributions from which the samples were drawn are different. On the other hand, when the two rank sums differ substantially, it suggests that the two samples may have come from different distributions. We denote the rank sum for governmental economists by t_g and that for university economists by t_u . Then $t_g = 4 + 7 + 2 + 8 + 1 + 3 = 25$ and $t_u = 6 + 9 + 5 + 11 + 10 + 12 + 13 = 66$. The sum of t_g and t_u will always equal $n(n + 1)/2$, that is the sum of all integers from 1 through n . In the particular case in hands, $n_g = 6$, $n_u = 7$, $n = 13$, and $t_g + t_u = 13(13 + 1)/2 = 91$. Since $t_g + t_u$ is fixed, a small value for t_g implies a large value for t_u (and vice versa) and a large difference between t_g and t_u . Therefore, the smaller the value of one of the rank sums, the greater the evidence to indicate that the samples were selected from different distributions. However, when comparing these two values, we must also take into account the fact that a t may be small due to the fact that the corresponding n is small; in our case, t_g may be smaller because the governmental sample has less subjects. The test's statistic is any of the two rank sums. Critical values for this statistic are given in appropriate Wilcoxon rank sum tables. We take t_g and looking at the table for $n_g = 6$ and $n_u = 7$ we get, for a significance level of 5%, critical values of 28 and 56.



Since our statistic is in the critical region, we reject, meaning that our data confirm that the two distributions are different.

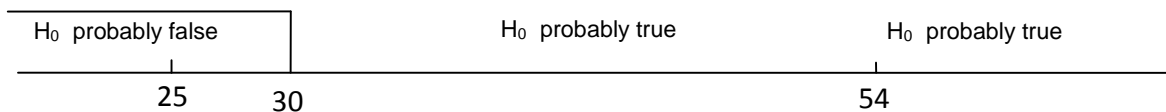
Note that the assumptions necessary for the validity of the Mann-Whitney test do not specify the shape of probability distribution. However, the distributions are assumed to be continuous so that the probability of tied measurements is zero, and, consequently, to each measurement can be assigned a unique rank. In practice, however, rounding of continuous measurements may sometimes produce ties. As long as the number of ties is small relative to the sample sizes, the Mann-Whitney test procedure is applicable. On the other hand, the test is not recommended to compare discrete distributions for which many ties are expected. Ties may be treated in the following way: assign tied measurements the average of the ranks they would receive if they were unequal. For example, if the third-ranked and fourth-ranked measurements are tied, we assign to each one a rank of $\frac{3+4}{2} = 3.5$. If the third-ranked, fourth-ranked and fifth-ranked measurements are tied, we assign to each one a rank

of $\frac{3+4+5}{3} = 4$.

Returning to our example, we may formulate the question more exactly: is it true that the university economists' predictions tend to be higher than the predictions of the governmental economists? In other words, is the density f_Y shifted to the right with respect to density f_X ? Conceptually this shift equals the systematic component in the difference between the predictions of a generic university economist and a generic government economist. That is:

- H_0 : the probability distribution corresponding to the government economists' predictions of inflation rate is in the same position or shifted to the right with respect to the university's economists' one
- H_1 : the probability distribution corresponding to the government economists' predictions of inflation rate is shifted to the left with respect to the university's economists' one

We have to find out the rejection region. We take t_g as statistic and suppose that its value is very large. This means that governmental economists make predictions with larger ranks and thus with higher values than university's economists. This is strongly in favor of H_0 and therefore rejection region is on the other side, the left one. Critical values are different and they are, for a significance level of 5%, 30 and 54. Statistic falls in the rejection region and thus our data confirms that governmental predictions are shifted to the left.



When sample size, n_1 or n_2 , is larger than 10, tables do not provide us with critical values anymore. In these cases statistic distribution can be approximated with a normal distribution

$$\frac{t_1 - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \sim N(0; 1).$$

4.10 Wilcoxon signed rank test

Prerequisites: the difference is a random variable having a continuous probability distribution.

H_0 : position of distribution for variable A = position of distribution for variable B

Statistic: sum of ranks of differences

Statistic distribution: Wilcoxon signed rank table or $\frac{t - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0; 1)$ when sample is large and tables are not available

SPSS: Analyze → Nonparametric Tests → Related Samples



Frank Wilcoxon (1892-1965)

Rank tests can also be employed to compare two probability distributions when a paired difference design is used. For example, consumer preferences for two competing products are often compared by analyzing the responses in a random sample of consumers who are asked to rate both

products. Thus, the ratings have been paired on each consumer. Consider for example a situation when 10 students have been asked to compare the teaching ability of two professors, say (1) and (2) . Each of the students grades the teaching ability on a scale from 1 to 10, with higher grades implying better teaching. The results of the experiment are as follows:

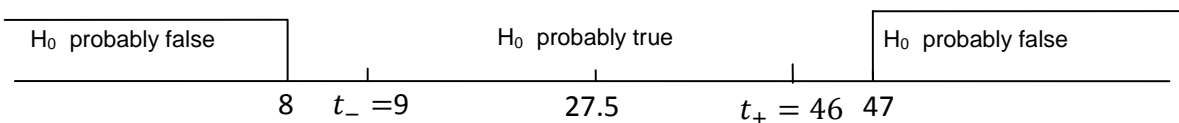
| student | $g^{(1)}$ | $g^{(2)}$ | $g^{(1)} - g^{(2)}$ | $ g^{(1)} - g^{(2)} $ | sign of $g^{(1)} - g^{(2)}$ | rank of $ g^{(1)} - g^{(2)} $ |
|---------|-----------|-----------|---------------------|-----------------------|-----------------------------|-------------------------------|
| 1 | 6 | 4 | 2 | 2 | + | 5 |
| 2 | 8 | 5 | 3 | 3 | + | 7.5 |
| 3 | 4 | 5 | -1 | 1 | - | 2 |
| 4 | 9 | 8 | 1 | 1 | + | 2 |
| 5 | 4 | 1 | 3 | 3 | + | 7.5 |
| 6 | 7 | 9 | -2 | 2 | - | 5 |
| 7 | 6 | 2 | 4 | 4 | + | 9 |
| 8 | 5 | 3 | 2 | 2 | + | 5 |
| 9 | 6 | 7 | -1 | 1 | - | 2 |
| 10 | 8 | 2 | 6 | 6 | + | 10 |

Here $g^{(1)}$ and $g^{(2)}$ are the grades assigned by each Student's to professor (1) and (2) . Since this is a paired difference experiment, we analyze the differences between the measurements. Examining the differences allows removing a possible common causality behind these ratings. In fact, the fourth and the sixth students seem to have given higher than other students' ratings to both professors.

This rank test requires that we calculate the ranks of the absolute values of the differences between the measurements. Since there are ties, the tied absolute differences are assigned the average of the ranks they would receive if they were unequal but successive measurements. For example, the absolute value 3 appears two times. If these were unequal measurements, their ranks would have been 8 and 7. Thus the rank for 3 equals $\frac{8+7}{2} = 7.5$. In the same way, the rank for 2 equals $\frac{6+5+4}{3} = 5$, the rank for 1 is $\frac{3+2+1}{3} = 2$. After the absolute differences are ranked, the sum of the ranks of the positive differences of the original measurements, t_+ , and the sum of the ranks of the negative measurements, t_- , are computed. In our case: $t_+ = 5 + 7.5 + 2 + 7.5 + 9 + 5 + 10 = 46$ and $t_- = 2 + 5 + 2 = 9$. Now we are ready to test the non-parametric hypotheses:

- H_0 : the probability distributions of the ratings for professor (1) is in the same position as the one for professor (2) , $(1) \sim (2)$
- H_1 : the probability distributions of the ratings for professor (1) is in a different position as the one for professor (2) , $(1) \neq (2)$

As the test statistic we use any t . The more the difference between t_- and t_+ , the greater the evidence to indicate that the two probability distributions differ in location. Note that also for this test the sum of $t_- + t_+$ is fixed and equal to $n(n + 1)/2$. Left critical value is tabulated, while right critical value can be found for symmetry. In our case, we take for example t_- which is 9. The left critical value, for a significance level of 5%, is 8. The other critical value is $n(n + 1)/2 - 8 = 55 - 8 = 47$.



As it can be seen in the schema, this test is perfectly symmetric and when one t falls into the central region, the other automatically does the same. Vice versa, when one t falls into a rejection region, the other falls into the other rejection region. In our example we accept and therefore our data are not able to prove that the two distributions are different.

Obviously, also for this test we have one-tailed versions. This is performed in the usual way, taking care to choose one statistic and decide which the rejection region for that statistic is.

Since we have assumed that the distribution of a difference is continuous, there may not be differences which are exactly 0. However, in practice, they may occur due to rounding: in such cases, we must decide whether assigning their rank to t_- or to t_+ . For the two-tailed test there is no solution. Since a difference of 0 is in favor of H_0 hypothesis, assigning it to either statistic can unbalance the situation and push in favor of H_1 . Moreover, a difference of 0 is strongly in favor of H_0 , but it would have the smaller rank. So, the two-tailed test cannot be performed at all if we have any 0 difference. However, the one-tailed test can be performed. For example:

- H_0 : the probability distributions of the ratings for professor ⁽¹⁾ is in the same position or shifted to the left with respect to the one for professor ⁽²⁾, $(1) \lesssim (2)$, $(1) - (2) \lesssim 0$
- H_1 : the probability distributions of the ratings for professor ⁽¹⁾ is shifted to the right with respect to the one for professor ⁽²⁾, $(1) \succ (2)$, $(1) - (2) \succ 0$

A difference of 0 is in favor of H_0 hypothesis which includes also all the negative differences. Therefore, any 0 difference's rank is assigned, with these hypotheses, to t_- .

When $n \geq 25$ statistic's tables are not available anymore. Statistic's distribution can be approximated with:

$$\frac{t - n(n + 1)/4}{\sqrt{n(n + 1)(2n + 1)/24}} \sim N(0; 1),$$

where it is better to take as statistic the smaller between t_- and t_+ , since usually standard normal distribution tables provide the area on the left.

4.11 Kruskal-Wallis test

Prerequisites: there are 5 or more measurements in each sample; the k probability distributions from which the samples are drawn are continuous

H_0 : position of distribution of populations is the same

Statistic: $H = \frac{12}{n(n+1)} \left[\sum_{j=1}^k \frac{r_j^2}{n_j} \right] - 3(n + 1)$

Statistic distribution: chi square distribution with $k - 1$ degrees of freedom

SPSS: Analyze → Nonparametric Tests → Independent Samples



William Henry Kruskal (1919-2005)



Wilson Allen Wallis (1912-1998)

The Kruskal-Wallis test is the Mann-Whitney test when more than two populations are involved. Its corresponding parametric test is the Analysis of Variance.

For example, a health administrator wants to compare the unoccupied bed space for three

hospitals. She randomly selects 10 different days from the records of each hospital and lists the number of unoccupied beds for each day. Just as with two independent samples, we base our comparison on the rank sums for these three sets of data. Ties are treated as in the Mann-Whitney test by assigning the average value of the ranks to each of the tied observations:

| Hospital 1 | | Hospital 2 | | Hospital 3 | |
|-------------|------|---------------|------|---------------|------|
| Beds | Rank | Beds | Rank | Beds | Rank |
| 6 | 5 | 34 | 25 | 13 | 9.5 |
| 38 | 27 | 28 | 19 | 35 | 26 |
| 3 | 2 | 42 | 30 | 19 | 15 |
| 17 | 13 | 13 | 9.5 | 4 | 3 |
| 11 | 8 | 40 | 29 | 29 | 20 |
| 30 | 21 | 31 | 22 | 0 | 1 |
| 15 | 11 | 9 | 7 | 7 | 6 |
| 16 | 12 | 32 | 23 | 33 | 24 |
| 25 | 17 | 39 | 28 | 18 | 14 |
| 5 | 4 | 27 | 18 | 24 | 16 |
| $r_1 = 120$ | | $r_2 = 210.5$ | | $r_3 = 134.5$ | |

We test

- H_0 : the probability distributions of the number of unoccupied beds have the same position for all three hospitals
- H_1 : at least one of the hospitals has probability position different with respect to the others.

The test statistic, called H , is $\frac{12}{n(n+1)} \sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2$, where k denotes the number of distributions involved, n_j is the number of measurements available for the j th distribution, r_j is the corresponding rank sum, $\bar{r}_j = r_j/n_j$ is the mean rank for population j and $\bar{r} = (r_1 + r_2 + \dots + r_k)/n = \left[\frac{n(n+1)}{2} \right] / n = \frac{n+1}{2}$ (remembering that the sum of ranks is fixed, as for Mann-Whitney and Wilcoxon tests) is the mean rank for the whole population. As it can be seen from the formula, this statistic measures the extent to which the k ranks differ with respect to the average rank. Note that H statistic is always non-negative. It takes on the value zero if and only if all samples have the same mean rank, that is $\bar{r}_j = \bar{r}$ for all j . This statistic becomes increasingly large as the distance between a sample mean rank \bar{r}_j and the mean rank for the whole population grows.

However, the formula that is used for practical calculations is an easier one¹⁵:

$$H = \frac{12}{n(n+1)} \left[\sum_{j=1}^k \frac{r_j^2}{n_j} \right] - 3(n+1)$$

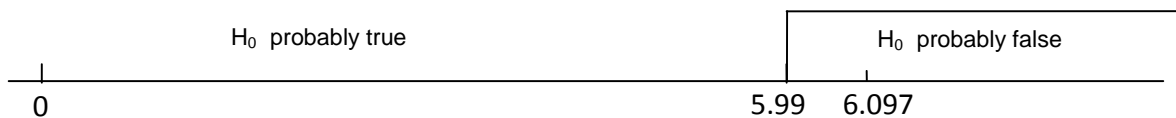
In our case $k = 3$, $n_1 = n_2 = n_3 = 10$ and $n = 30$. H is $\frac{12}{30 \cdot 31} \left[\frac{120^2}{10} + \frac{210.5^2}{10} + \frac{134.5^2}{10} \right] - 3 \cdot 31 =$

¹⁵
$$H = \frac{12}{n(n+1)} \sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2 = \frac{12}{n(n+1)} \sum_{j=1}^k n_j (\bar{r}_j^2 - 2\bar{r}_j\bar{r} + \bar{r}^2) = \frac{12}{n(n+1)} \left\{ \sum_{j=1}^k n_j \left(\frac{r_j^2}{n_j^2} - 2 \frac{r_j}{n_j} \frac{n+1}{2} + \frac{(n+1)^2}{4} \right) \right\} =$$

$$\frac{12}{n(n+1)} \left\{ \sum_{j=1}^k \frac{r_j^2}{n_j} - (n+1) \sum_{j=1}^k r_j + \frac{(n+1)^2}{4} \sum_{j=1}^k n_j \right\} = \frac{12}{n(n+1)} \left\{ \left[\sum_{j=1}^k \frac{r_j^2}{n_j} \right] - (n+1) \frac{n(n+1)}{2} + \frac{(n+1)^2}{4} n \right\} = \frac{12}{n(n+1)} \left\{ \left[\sum_{j=1}^k \frac{r_j^2}{n_j} \right] - n(n+1) \frac{2n+1}{2} + \frac{n(n+1)^2}{4} \right\} =$$

6.097.

The H statistic's distribution is, under the hypothesis that the null hypothesis is true, approximately a chi square distribution with $k - 1$ degrees of freedom. This approximation is adequate as long as each of the k sample sizes is at least 5. Chi square distribution has only one tail on the right and thus the rejection region for the test is located in the right tail. In our case $k = 3$, so we are dealing with a chi square distribution with 2 degrees of freedom. Using the significance method, we find¹⁶ a significance of 4.74% which means that we reject. Using the critical region method with 5% significance level, we get a critical value of 5.99.



4.12 Pearson's correlation coefficient

Prerequisites: coupled data

$H_0: \text{Corr}(X, Y) = 0$

Statistic: $\frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$

Statistic distribution: Student's t with $n - 2$ degrees of freedom

SPSS: Analyze → Correlate → Bivariate



Karl Pearson (1857-1936)

Consider two random variables, X and Y , of which we have only n couples of outcomes, $(x_i; y_i)$. It is important that the outcomes that we have are in couples, since we are interesting in estimating the correlation between the two variables. We use as estimator the Pearson's correlation coefficient which is defined, through the introduction of the SS (sum of squares) quantity, as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2 \cdot \sum_{i=1}^n (y_i - \bar{y}_n)^2}} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n}{\sqrt{\{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2\} \cdot \{\sum_{i=1}^n y_i^2 - n \bar{y}_n^2\}}}$$

As it can be seen from the formulas, quantities SS have two equivalent definitions, of which the latter is easier to use in practical calculations while the former is more useful for theoretical considerations. In particular, we can immediately observe from the second definition that SS_{xx} and SS_{yy} are strictly positive and therefore the square root and the denominator are well defined. In the particular case when all the x_i or all the y_i have the same value, the corresponding SS quantity becomes 0 and the Pearson correlation coefficient is no more defined. This is a very rare case and corresponds to the situation when there are only constant outcomes for random variable X or Y ; clearly, from constant outcomes we can not estimate anything concerning the behavior of random variables.

SS_{xx}/n is the estimation of the variance of random variable X , while SS_{yy}/n is the

¹⁶ Significance can be calculated in two different ways. (1) Using English Microsoft Excel function CHIDIST(6.097;2) which gives us the area of the left tail. (2) Using chi square distribution table, which usually provides the left side critical values.

estimation of the variance of random variable Y and SS_{xy}/n is the estimation for $\text{Cov}(X, Y)$. Since the correlation is exactly $\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$, Pearson's correlation coefficient is the estimation for the correlation.

The sign of r is determined only by the sign of SS_{xy} . It can moreover be easily demonstrated¹⁷ that $|SS_{xy}| \leq \sqrt{SS_{xx} \cdot SS_{yy}}$ and therefore the value of r must lie between -1 and $+1$, independently from how large or small are the numbers x_i and y_i . In other words, r is a scaleless variable. A value of r near or equal to zero is interpreted as little or no correlation between X and Y . In contrast, the closer r comes to -1 or $+1$, the stronger is the correlation of these variables. Positive values of r imply a positive correlation between X and Y . That is, if one increases, the other one increases as well. Negative values of r imply a negative correlation. In fact, X and Y move in the opposite directions: when X increases, Y decreases and vice versa. In sum, this coefficient of correlation reveals whether there is a common tendency in moves of X and Y .

We have a test to check whether $\text{Corr}(X, Y)$ is different from 0 , meaning that there is a linear relation between random variables X and Y . This test uses the fact that statistic

$$\frac{r \cdot \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

is distributed like a Student's t distribution with $n - 2$ degrees of freedom. We remind the fact that *independence implies zero correlation but not vice versa*: therefore, when the correlation is different from 0 , we are sure that the two random variables are dependent.

For example, suppose we have these 11 couples of data

| | | | | | | | | | | | |
|-----|---|---|---|---|---|---|----|---|---|---|----|
| x | 2 | 3 | 4 | 3 | 5 | 6 | 7 | 3 | 1 | 3 | 4 |
| y | 5 | 5 | 7 | 5 | 7 | 7 | 14 | 5 | 3 | 1 | 12 |

we get $SS_{xx} = 4 + 9 + 16 + 9 + 25 + 36 + 49 + 9 + 1 + 9 + 16 - 11 \cdot 3.7^2 = 30.18$, $SS_{xy} = 10 + 15 + 28 + 15 + 35 + 42 + 98 + 15 + 3 + 3 + 48 - 11 \cdot 3.7 \cdot 6.5 = 47.36$ and $SS_{yy} = 138.73$. Therefore $r = \frac{47.36}{\sqrt{30.18 \cdot 138.73}} = 0.732$ with 11 couples of data and the value of our statistic is $\frac{0.732 \cdot \sqrt{9}}{\sqrt{0.464}} \approx 3.223$ and a significance, for the two-tailed test, of 1.04%. Therefore, taking a significance level of 5%, 11 couples of data with Pearson's correlation coefficient of 0.732 are enough to prove that the correlation is different from 0 and therefore the two variables are not independent.

¹⁷ This fact obtains by applying the Cauchy-Schwarz inequality, $|\tilde{a} \cdot \tilde{b}| \leq \|\tilde{a}\| \times \|\tilde{b}\|$, to the n -vectors \tilde{a} and \tilde{b} with $a_i = y_i - \bar{y}_n$ and $b_i = x_i - \bar{x}_n$.

4.13 Spearman's rank correlation coefficient

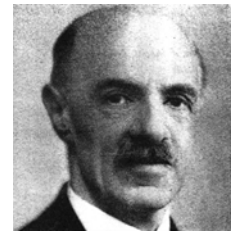
Prerequisites: coupled ranked data or coupled data from continuous distributions

H₀: ranks are uncorrelated

Statistic: Spearman's rank correlation coefficient

Statistic distribution: Spearman table

SPSS: Analyze → Correlate → Bivariate



Charles Spearman
(1863-1945)

The Spearman's rank correlation coefficient is the non parametric version of the Pearson's correlation coefficient.

Taking the same data of the previous example,

| | | | | | | | | | | | |
|-----|---|---|---|---|---|---|----|---|---|---|----|
| (1) | 2 | 3 | 4 | 3 | 5 | 6 | 7 | 3 | 1 | 3 | 4 |
| (2) | 5 | 5 | 7 | 5 | 7 | 7 | 14 | 5 | 3 | 1 | 12 |

this time instead of taking the values, we assign ranks. It is important that ranks be assigned independently for *x* and *y*, yet maintaining the coupled position of the data:

| | | | | | | | | | | | |
|-------------------------|-----|-----|-----|-----|---|----|----|-----|---|-----|-----|
| <i>r</i> ⁽¹⁾ | 2 | 4.5 | 7.5 | 4.5 | 9 | 10 | 11 | 4.5 | 1 | 4.5 | 7.5 |
| <i>r</i> ⁽²⁾ | 4.5 | 4.5 | 8 | 4.5 | 8 | 8 | 11 | 4.5 | 2 | 1 | 10 |

The Spearman's rank correlation coefficient, *r_S*, is calculated exactly as Pearson's correlation coefficient:

$$r_S = \frac{SS_{r^{(1)}r^{(2)}}}{\sqrt{SS_{r^{(1)}r^{(1)}}SS_{r^{(2)}r^{(2)}}}$$

Where, exactly as for Pearson's correlation coefficient,

$$SS_{r^{(1)}r^{(2)}} = \sum_{i=1}^n (r_i^{(1)} - \bar{r}_n^{(1)})(r_i^{(2)} - \bar{r}_n^{(2)}) = \left\{ \sum_{i=1}^n r_i^{(1)}r_i^{(2)} \right\} - n \cdot \bar{r}_n^{(1)} \cdot \bar{r}_n^{(2)},$$

$$SS_{r^{(j)}r^{(j)}} = \sum_{i=1}^n (r_i^{(j)} - \bar{r}_n^{(j)})^2 = \left\{ \sum_{i=1}^n (r_i^{(j)})^2 \right\} - n(\bar{r}_n^{(j)})^2$$

$$\bar{r}_n^{(j)} = \frac{1}{n} \sum_{i=1}^n r_i^{(j)}$$

The value of *r_S* always falls between -1 and +1, with +1 indicating perfect positive correlation and -1 for perfect negative correlation. The closer falls *r_S* to -1 or +1, the greater the correlation between the ranks. Conversely, the nearer *r_S* is to 0, the less the correlation.

For Spearman's rank correlation we have, however, additional information since the values used in the calculation must be integer numbers between 1 and *n*. Therefore, through mathematical calculations, we can derive¹⁸ an alternative formula valid only when there are not tied

¹⁸ Starting from the consideration that $\sum_{i=1}^n r_i^{(j)} = 1 + 2 + 3 + \dots + (n-1) + n = \frac{n(n+1)}{2}$ we can obtain a simplification $SS_{r^{(1)}r^{(2)}} = \left\{ \sum_{i=1}^n r_i^{(1)}r_i^{(2)} \right\} - \frac{1}{n} \left[\frac{n(n+1)}{2} \right]^2 = \left\{ \sum_{i=1}^n r_i^{(1)}r_i^{(2)} \right\} - \frac{n(n+1)^2}{4}$ and $SS_{r^{(j)}r^{(j)}} = \left\{ \sum_{i=1}^n [r_i^{(j)}]^2 \right\} - \frac{1}{n} \left[\frac{n(n+1)}{2} \right]^2 = \left\{ \sum_{i=1}^n [r_i^{(j)}]^2 \right\} - \frac{n(n+1)^2}{4}$. Moreover, since $\sum_{i=1}^n [r_i^{(j)}]^2 = 1^2 + 2^2 + 3^2 + \dots + (n-1)^2 + n^2 =$

ranks:

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n d_i^2.$$

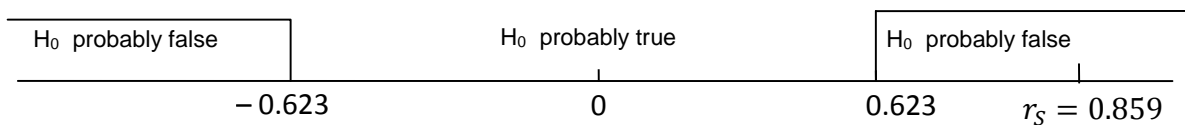
where $d_i = r_i^{(1)} - r_i^{(2)}$, the difference between the rank of the i th measurement in the first set and the rank of the i th measurement in the second set. We can see that if all ranks are identical, that is, $r_i^{(1)} = r_i^{(2)}$ for every i , then $r_s = 1$. We must take care to remember that this formula is valid only when there are no tied ranks.

Returning to our example, we see that

$$\sum_{i=1}^{10} d_i^2 = 2.5^2 + 0 + 0.5^2 + 0 + 1^2 + 2^2 + 0 + 0 + 1^2 + 3.5^2 + 2.5^2 = 31,$$

consequently, $r_s = 1 - 6 \cdot 31 / [11 \cdot (11^2 - 1)] \approx 0.859$. The fact that r_s is close to 1 indicates that the rankings given by the two magazines tend to agree, but the agreement is not perfect.

If the sets of ranks are formed by values taken by n independent realizations of random variables $R^{(1)}$ and $R^{(2)}$, the Spearman's rank correlation coefficient may be used for testing whether the value of $\text{Corr}(R^{(1)}, R^{(2)})$ is different from 0. The statistic is the coefficient itself. In the previous example, with $n = 11$ and a significance level of 5% we have a critical value of 0.623,



and therefore we reject, meaning that the ranks are correlated and there is a relation between the order of the two variables.

Spearman's rank correlation coefficient can be used, as every other rank test, in all the situations where effective measures are not available and only ranks are provided. Suppose ten new car models are evaluated by two consumer magazines and each magazine ranks the braking system of the cars from 1 (best) to 10 (worst). We want to determine whether the magazines' ranks are related. If they are, we may conclude that these rankings contain useful information about the braking system. Otherwise, if the rankings given by the two magazines are not related, we should not regard these ranking as containing useful information since they are contradictory and we do not know which one to use. Let the ranks given by the two magazines be as follows:

| | | | | | | | | | | |
|---------------|---|---|---|---|---|---|---|---|---|----|
| Car model i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------------|---|---|---|---|---|---|---|---|---|----|

$\frac{n(n+1)(2n+1)}{6}$ we see that $SS_{r^{(j)}, r^{(j)}} = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}$. Finally, taking into account that $r_i^{(1)} r_i^{(2)} = \frac{1}{2} \{ [r_i^{(1)}]^2 + [r_i^{(2)}]^2 - [r_i^{(1)} - r_i^{(2)}]^2 \}$, we obtain $SS_{r^{(1)}, r^{(2)}} = \sum_{i=1}^n r_i^{(1)} r_i^{(2)} - \frac{n(n+1)^2}{4} = -\frac{1}{2} \sum_{i=1}^n [r_i^{(1)} - r_i^{(2)}]^2 + \frac{1}{2} \left\{ \sum_{i=1}^n [r_i^{(1)}]^2 - \frac{n(n+1)^2}{4} \right\} + \frac{1}{2} \left\{ \sum_{i=1}^n [r_i^{(2)}]^2 - \frac{n(n+1)^2}{4} \right\} = -\frac{1}{2} \sum_{i=1}^n [r_i^{(1)} - r_i^{(2)}]^2 + \frac{1}{2} SS_{r^{(1)}, r^{(1)}} + \frac{1}{2} SS_{r^{(2)}, r^{(2)}} = -\frac{1}{2} \sum_{i=1}^n [r_i^{(1)} - r_i^{(2)}]^2 + \frac{n(n^2-1)}{12}$. Consequently, $r_s = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2} \sum_{i=1}^n [r_i^{(1)} - r_i^{(2)}]^2}{\sqrt{\frac{n(n^2-1)}{12} \frac{n(n^2-1)}{12}}} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n d_i^2$.

| | | | | | | | | | | |
|--------------------------------------|---|---|----|---|---|----|---|---|---|---|
| Rank given by magazine 1 $r_i^{(1)}$ | 4 | 1 | 9 | 5 | 2 | 10 | 7 | 3 | 6 | 8 |
| Rank given by magazine 2 $r_i^{(2)}$ | 5 | 2 | 10 | 6 | 1 | 9 | 7 | 3 | 4 | 8 |

In this case data are already ranked and the coefficient can be calculated directly.

4.14 Multinomial experiment

Many business analyses consist of enumerating the number of occurrences of some event. For example, we may count the number of consumers who choose each of the three brands of coffee, or the number of sales made by each of five automobile salespeople during a month. When there is a single scale to classify data, as in all examples above, we have a one dimensional classification. In some cases we may collect the count data characterizing several factors. For example, we may be interested in investigating whether the color of automobile purchased is related to the sex of the buyer. In this case we are dealing with a two dimensional classification. The corresponding data constitute a contingency table. Count data are traditionally analyzed using tables.

4.14.1 One dimensional classification

Prerequisites: $n \cdot \tilde{p}_i \geq 5$ for all i

H₀: $p_i = \tilde{p}_i$ for all i or, equivalently, $n_i = n \cdot \tilde{p}_i$ for all i

Statistic: table's chi square $\chi^2 = \sum_{i=1}^k \frac{(n_i - n \cdot \tilde{p}_i)^2}{n \cdot \tilde{p}_i}$,

Statistic distribution: chi square with $k - 1$ degrees of freedom

SPSS: Analyze → Nonparametric tests → One Sample

The properties of the one dimensional multinomial experiment are as follows:

- the experiment consists of n identical trials;
- the trials are independent;
- there are k possible outcomes to each trial;
- the probabilities of the k outcomes, denoted by p_1, p_2, \dots, p_k , remain the same from trial to trial, where $p_1 + p_2 + \dots + p_k = 1$ (therefore there is no other possible outcome outside the ones we are considering);
- the random variables of interest are the counts N_1, N_2, \dots, N_k in each of the k cells.

For example, suppose a large supermarket chain conducts a consumer preference survey by recording the brand of bread purchased by customers in its stores. Assume the chain carries three brands of bread, A, B and C. The brand preferences of a random sample of 150 consumers are observed, and the resulting count data are as follows: A: 61, B: 53, C: 36. Do these data indicate that a preference exists for any of these brands?

Our consumer preference survey satisfies the properties of a multinomial experiment. The experiment consists in randomly sampling $n = 150$ buyers from a large population of consumers containing an unknown proportion p_1 who prefer brand A, a proportion p_2 who prefer brand B, and a proportion p_3 who prefer the store brand, C. Approaching a buyer concerning his preference, we perform a single trial that can result in one of three outcomes: the consumer prefers brand A, B or C. Probabilities of these outcomes are p_1, p_2 , and p_3 , respectively. The buyer's preference of any single consumer in the sample does not affect the preference of another. Consequently, the trials are

independent. The recorded data are the numbers of buyers in each of the consumer preference categories. Thus, the consumer preference survey satisfies the five properties of a multinomial experiment.

Note that we may talk about the proportions p_1, p_2, p_3 as probabilities because in a population consisting totally of n agents, np_1 prefer brand A, np_2 opt for brand B, and np_3 for brand C. Consequently, the probability to choose randomly a customer who buys A, B, C will be correspondingly $np_1/n = p_1, np_2/n = p_2$ and $np_3/n = p_3$. That is why we may talk about p_i as a proportion as well as about a probability. The three probabilities p_1, p_2, p_3 are unknown and we want to use the survey data to make inferences about their size.

The general form for a test of a hypothesis concerning multinomial probabilities is as follows:

- $H_0: p_1 = \tilde{p}_1, p_2 = \tilde{p}_2, \dots, p_k = \tilde{p}_k$, where $\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_k$ represent the hypothesized values of the multinomial probabilities ($\tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3 = 1/3$ in the above example with three types of bread)
- H_1 : at least one of the multinomial probabilities does not equal its hypothesized value, in other words, there is an i such that the corresponding actual probability p_i does not coincide with its hypothesized value $\tilde{p}_i, p_i \neq \tilde{p}_i$.

We build a table of observed counts and a table of predicted counts $n \cdot \tilde{p}_j$

| | | | | | | |
|-----------------|----|----|--|--|----|----|
| A | B | C | | A | B | C |
| 61 | 53 | 36 | | 50 | 50 | 50 |
| observed counts | | | | predicted counts under H_0 hypothesis | | |

The test statistic is the table's chi square, a measure calculated as

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n \cdot \tilde{p}_i)^2}{n \cdot \tilde{p}_i},$$

where n_i are called observed counts, $n = n_1 + n_2 + \dots + n_k$ is the total sample size. This statistic is distributed as a chi square distribution with $k - 1$ degrees of freedom. Observing the chi square statistic, it is evident that when the observed numbers are very different from the predicted counts, $n \cdot \tilde{p}_i$, the value of the statistic is very large, while when the observed numbers coincides with the predicted ones the statistic is zero. Therefore, rejection region is only on the right. This test works only if the predicted counts are all $n \cdot \tilde{p}_i \geq 5$, while it is not important that the observed ones be at least 5.

In our particular example, $\chi^2 = \frac{(61-150 \cdot 1/3)^2}{150 \cdot 1/3} + \frac{(53-150 \cdot 1/3)^2}{150 \cdot 1/3} + \frac{(36-150 \cdot 1/3)^2}{150 \cdot 1/3} \approx 6.52$. Since here $k = 3$, we are dealing with a chi square distribution with 2 degrees of freedom. Statistic's significance is¹⁹ 3.84% and therefore we reject, meaning that consumers' preferences are not

¹⁹ Significance can be calculated in different ways: (1) using English Microsoft Excel function CHIDIST(6.52;2) which gives us the probability of the right tail of chi square distribution; (2) looking into chi square tables which usually provide critical values for different significance levels; (3) using English Microsoft Excel function CHIINV(5%;2) which gives us the critical value corresponding to 5% significance level; (4) test can be performed also using English Microsoft Excel function CHITEST which, given the observed table and the predicted table, gives us the value of chi square statistic and then using CHIDIST significance can be found.

uniform and that there is at least one type of bread that has a probability different from $1/3$. If we want to use the critical regions method, critical value for 5% is 5.99 and therefore 6.52 is in the rejection region which for this test is always on the right.

As another example, using the same data we want to check whether the bread's probabilities follow a 40%, 40%, 20% distribution:

- $H_0: p_1 = 40\%$ and $p_2 = 40\%$ and $p_3 = 20\%$
- $H_1: p_1 \neq 40\%$ or $p_2 \neq 40\%$ or $p_3 \neq 20\%$.

The observed and predicted tables are

| | | |
|----------|----------|----------|
| A | B | C |
| 61 | 53 | 36 |

observed counts

| | | |
|----------|----------|----------|
| A | B | C |
| 60 | 60 | 30 |

predicted counts
under H_0 hypothesis

and $\chi^2 = \frac{(61-60)^2}{60} + \frac{(53-60)^2}{60} + \frac{(36-30)^2}{30} \approx 1.43$ with a significance of 48.8%. Therefore we accept, meaning that our sample is not able to prove that consumers' preference is not 40%, 40%, 20%.

4.14.2 Two dimensional contingency table

Prerequisites: $r_i \cdot c_j/n \geq 5$ for all i, j

H_0 : classifications are independent

Statistic: table's chi square $\chi^2 = \sum_{j=1}^{k_c} \sum_{i=1}^{k_r} \frac{(n_{ij} - r_i \cdot c_j/n)^2}{r_i \cdot c_j/n}$,

Statistic distribution: chi square with $(k_r - 1)(k_c - 1)$ degrees of freedom

SPSS: Analyze → Descriptive Statistics → Crosstabs → Statistics → Chi-square

Suppose, for example, that an automobile magazine is interested in determining the relationship between the size and manufacturer of newly purchased automobiles. One thousand recent buyers of cars made in Germany are randomly sampled, and each purchase is classified with respect to the size (small, intermediate, and large) and manufacturer of the automobile (Volkswagen, BMW, Opel, Mercedes). The data are summarized in the two-way table:

| Size\Manufacturer | VW | BMW | Opel | Mercedes | Totals |
|---------------------|-----|-----|------|----------|--------|
| Small | 157 | 65 | 181 | 10 | 413 |
| Intermediate | 126 | 82 | 142 | 46 | 396 |
| Large | 58 | 45 | 60 | 28 | 191 |
| Totals | 341 | 192 | 383 | 84 | 1000 |

This table is called a contingency table.

SPSS: Analyze → Descriptive Statistics → Crosstabs

It presents multinomial count data classified in two dimensions, namely automobile size and manufacturer. Each count is indicated with $n_{i,j}$, where the first index is referred to the row, the size, and the second index to the column, the manufacturer. We also indicate with r_i the rows' totals and with c_j the columns' totals, and these quantities are called marginal counts. The sample size is n and coincides with the grand total, in our case 1000.

| Size\Manufacturer | VW | BMW | Opel | Mercedes | Totals |
|-------------------|-----------|-----------|-----------|-----------|--------|
| Small | $n_{1,1}$ | $n_{1,2}$ | $n_{1,3}$ | $n_{1,4}$ | r_1 |
| Intermediate | $n_{2,1}$ | $n_{2,2}$ | $n_{2,3}$ | $n_{2,4}$ | r_2 |
| Large | $n_{3,1}$ | $n_{3,2}$ | $n_{3,3}$ | $n_{3,4}$ | r_3 |
| Totals | c_1 | c_2 | c_3 | c_4 | n |

This is a multinomial experiment with a total of $n = 1000$ trials, $k = 3 \cdot 4 = 12$ cells or possible outcomes, and probabilities $p_{ij} = \frac{n_{ij}}{n}$ for the cells. If the 1000 recent buyers are randomly chosen, the trials are considered independent and the probabilities are viewed as remaining constant from trial to trial. We also define the marginal probabilities for rows and columns as $p_{r_i} = \frac{r_i}{n}$ and $p_{c_j} = \frac{c_j}{n}$.

In a two dimensional classification experiment usually we are interested in checking whether one variable can influence the other. It may be helpful calculating the row percentages $\frac{n_{i,j}}{r_i}$ and column percentages $\frac{n_{i,j}}{c_j}$ as follow:

| Size\Manufacturer | VW | BMW | Opel | Mercedes | Totals |
|-------------------|-------|-------|-------|----------|--------|
| Small | 38.0% | 15.7% | 43.8% | 2.4% | 100.0% |
| Intermediate | 31.8% | 20.7% | 35.9% | 11.6% | 100.0% |
| Large | 30.4% | 23.6% | 31.4% | 14.7% | 100.0% |

| Size\Manufacturer | VW | BMW | Opel | Mercedes |
|-------------------|--------|--------|--------|----------|
| Small | 46.0% | 33.9% | 47.3% | 11.9% |
| Intermediate | 37.0% | 42.7% | 37.1% | 54.8% |
| Large | 17.0% | 23.4% | 15.7% | 33.3% |
| Totals | 100.0% | 100.0% | 100.0% | 100.0% |

Using row percentages we can show, for example, that among all small cars only 2.4% are produced by Mercedes compared to Opel which has 43.8% of the market. Using column percentages instead we see that among all Mercedes cars 11.9% are small compared to 33.3% of large ones.

SPSS: Analyze → Descriptive Statistics → Crosstabs → Cells

Therefore, in a two dimensional classification experiment we are not interested in whether the observed counts follow a predetermined distribution, since they are also influenced by the marginal counts (which depends on our sample's choice). We instead test whether the two classifications, manufacturer and size in our example, are independent.

- H_0 : row variable and column variable are independent, i.e. $np_{ij} = \frac{r_i \cdot c_j}{n}$
- H_1 : row variable and column variable are dependent, i.e. there is a couple i, j for which $np_{ij} \neq \frac{r_i \cdot c_j}{n}$.

That is, if we know which size car a buyer will choose, does this information give us a clue about the manufacturer of the car that is going to be bought? In a probabilistic sense we know that independence of events A and B implies $P(A \cap B) = P(A) \cdot P(B)$. Similarly, in the contingency table analysis, if the two classifications are independent, the probability that an item is classified in any particular cell of the table is a product of the corresponding marginal probabilities. Thus, under the hypothesis of independence, we must have: $\tilde{p}_{11} = p_{r_1} \cdot p_{c_1}$, $\tilde{p}_{12} = p_{r_1} \cdot p_{c_2}$ and so forth. To test

the hypothesis of independence, we use the same reasoning as in the one-dimensional tests. First we calculate the predicted count in each cell assuming that the null hypothesis of independence is true, multiplying n by the cell predicted probability $n \cdot \tilde{p}_{ij} = n \cdot p_{r_i} \cdot p_{c_j} = n \cdot r_i/n \cdot c_j/n = r_i c_j/n$. In our example

| Size\Manufacturer | VW | BMW | Opel | Mercedes | Totals |
|-------------------|--------------|--------------|--------------|-------------|--------|
| Small | 413·341/1000 | 413·192/1000 | 413·383/1000 | 413·84/1000 | 413 |
| Intermediate | 396·341/1000 | 396·192/1000 | 396·383/1000 | 396·84/1000 | 396 |
| Large | 191·341/1000 | 191·192/1000 | 191·383/1000 | 191·84/1000 | 191 |
| Totals | 341 | 192 | 383 | 84 | 1000 |

| Size\Manufacturer | VW | BMW | Opel | Mercedes | Totals |
|-------------------|-------|------|-------|----------|--------|
| Small | 140.8 | 79.3 | 158.2 | 34.6 | 413 |
| Intermediate | 135.0 | 76.0 | 151.7 | 33.5 | 396 |
| Large | 65.1 | 36.7 | 73.2 | 16.0 | 191 |
| Totals | 341 | 192 | 383 | 84 | 1000 |

As it can be seen, marginal counts have remained the same.

We use the chi square statistic to compare the observed and predicted counts in each cell of the contingency table $\chi^2 \approx \frac{(157-140.8)^2}{140.8} + \frac{(65-79.3)^2}{79.3} + \frac{(181-158.2)^2}{158.2} + \frac{(10-34.7)^2}{34.7} + \frac{(126-135.0)^2}{135.0} + \frac{(82-76.0)^2}{76.0} + \frac{(142-151.7)^2}{151.7} + \frac{(46-33.3)^2}{33.3} + \frac{(58-65.1)^2}{65.1} + \frac{(45-36.7)^2}{36.7} + \frac{(60-73.2)^2}{73.2} + \frac{(28-16.0)^2}{16.0} \approx 45.81$. Degrees of freedom are $(3 - 1) \cdot (4 - 1) = 6$, significance is²⁰ 0.000003%: therefore we reject the hypothesis of independence and we conclude that the size and manufacturer of a car selected by a purchaser are dependent events. Using instead the critical regions method, for a significance level of 5% we get a critical value of 12.59 and therefore the statistic value falls into the rejection region.

²⁰ Significance can be found and test can be performed in different ways: (1) using English Microsoft Excel function CHIDIST(45.81;6) which gives us the probability of the right tail of chi square distribution; (2) looking into chi square tables which usually provide critical values for different significance levels; (3) using English Microsoft Excel function CHIINV(5%;6) which gives us the critical value corresponding to 5% significance level; (4) test can be performed also using English Microsoft Excel function CHITEST which, given the observed table and the predicted table (which must be manually built), gives us the value of chi square statistic and then using CHIDIST significance can be found.

5. Which test to use?

While for some data situation it is evident which statistical test to use, such as for example multinomial experiments or when facing with a single distribution, when having to compare two distributions there are several tests that may be used, according to what we want to check.

Mann-Whitney test check whether two distributions have the same position or not. We use it when we have two sets of data and we are simply interested into testing whether they come from distributions in the same position or not. **Student's t test for two populations** is a test for the same situation but which tests whether the expected values of the distributions are the same and does not test their position.

Wilcoxon signed rank test checks whether *two distributions of paired data* have the same position. It can be used only when data are paired and it analyses the difference case with case (intra-case, inside the same case). Same thing for **Student's t test for paired data**, which analyses the difference between expected values of the two samples on a case by case basis.

Spearman rank correlation coefficient checks whether *two distributions of paired data* have the same order or a perfectly reverse order. It does not matter whether the data come from similar distribution or not, the important thing is that they are in order. It can be used only when data are paired and when we are simply interested in the order. **Pearson correlation coefficient** applies to the same situation, but checking the effective data's values and not the order.

There are however many cases where all the tests can be performed. Some theoretical examples:

- if data are in perfect reverse order (1, 2, 3, 4, 5 and 5, 4, 3, 2, 1), Spearman is equal to -1 (H_0 rejected, therefore orders are related) indicating that the order is reversed while Mann-Whitney test and Wilcoxon signed rank test accept H_0 indicating that data may be in the same position;
- if data are perfectly shifted (1, 2, 3, 4, 5 and 3, 4, 5, 6, 7), Wilcoxon signed rank test and Mann-Whitney test reject H_0 indicating that data do not have the same position while Spearman is equal to $+1$ (H_0 rejected, therefore orders are related) indicating that the order is the same.

Some more practical examples for non parametric tests, which are however valid also for the corresponding parametric tests provided their prerequisites are satisfied.

- *Given two students and their exams' grades on the same 6 economics subjects, which one will you hire for a position in a bank?*

Data are paired so we can use all three tests. However, we are not interested whether the two students have the same order or not, but simply whether their two distributions have the same position or not (and, if we want to do also one-tailed tests, whether one of the two students has exams' grades shifted to the right). For example, suppose that grades are 30 29 28 27 26 25 and 25 26 27 28 29 30: in this case for us the two students are equivalent, but Spearman is -1 (H_0 rejected, therefore orders are related) indicating that the order is reversed, while Mann-Whitney and Wilcoxon tests both accept indicating that data may come from the same position. Suppose instead that grades are 30 29 28 27 26 25 and 25 24 23 22 21 20: in this case the first student is evidently the best. Spearman is $+1$ (H_0 rejected, therefore orders are related) indicating that the order is the same, while Mann-Whitney and Wilcoxon tests both reject indicating that data

come from different position. So the best choice here is Mann-Whitney test, followed by Wilcoxon signed rank test since the differences *exam by exam* are not important. Spearman is not good here since we are not interested in the order.

- *Given two subjects and the grades given to the same 6 students, how can you test whether subject B has marks which have been inflated?*

Data are paired so we can use all three tests. However, we are not interested whether the two exams have the same order or not (because it can happen that a student is good in a subject but bad in another, due to personal preferences), but we are interested whether their distribution have the same position. For example, suppose that grades are 30 29 28 27 26 25 and 25 26 27 28 29 30: in this case, even though the same student got different grades, for each student who has got high grade in A and low in B there is another who compensate with a low grade in A and high in B (and this is not an indication that grades have been inflated, but simply that students good in A are not good in B, due to personal preferences). So subject B has not been inflated. Spearman is -1 (H_0 rejected, therefore orders are related) indicating that marks are in the reverse order, an information which is totally useless here, while the two Wilcoxon tests do not reject indicating that marks may come from the same distribution. Suppose instead that grades are 25 24 23 22 21 20 and 30 29 28 27 26 25: in this case exam's grades are evidently inflated. Spearman is $+1$ (H_0 rejected, therefore orders are related) indicating that marks are in the same order, an useless information, while Mann-Whitney and Wilcoxon tests both reject, indicating that marks' distribution do not have the same position. So the best choice here is Mann-Whitney test, followed by Wilcoxon signed rank test since the differences *subject by subject* are not important. Spearman is not good here. If, on the other hand, we want to concentrate the attention on the grades inflation subject by subject, Wilcoxon signed rank test is the best choice, followed by Mann-Whitney test.

- *Given two subjects and the grades given to the same 6 students, how can you test whether grades are consistent?*

Data are paired so we can use all three tests. Consistent here means that good students in one subject are also good in the other. So in this case we are interested in discovering whether grades have the same order or not. For example, suppose that grades are 30 29 28 27 26 25 and 25 26 27 28 29 30: in this case it is clear that good students in first subject are bad in the second. Spearman is -1 (H_0 rejected, therefore orders are related) indicating that the order is different while the two Wilcoxon tests do not reject indicating that marks may come from the same distribution, a useless information in this case. Suppose instead that grades are 25 24 23 22 21 20 and 30 29 28 27 26 25: in this case, even though second exam's grades are evidently inflated, at least good students in first exam are still the best ones in the second. Spearman is $+1$ (H_0 rejected, therefore orders are related) indicating that the order is the same, while Mann-Whitney and Wilcoxon tests both reject indicating that data have different positions. Therefore Spearman is the best choice, and Mann-Whitney and Wilcoxon tests are not appropriate.

- *Given two subjects and the grades given to 4 students for subject A and to 8 students (including the previous 4) for subject B, how can you test whether subject B has marks which have been inflated?*

Data are paired only for the first 4. So using Wilcoxon signed rank test implies taking very few subjects and looking at the table with 4 cases we see that we must always accept. Therefore Mann-Whitney test with $4/8$ subjects is the only test possible here.

6. Regression model

An important consideration in merchandising a product is the amount of money spent on advertising. Suppose you want to model the monthly sales revenue of a store as a function of the monthly advertising expenditure. First, you have to decide whether an exact relationship exists between these two variables. That is, whether it is possible to state the exact monthly revenue if the amount spent on advertising is known. We are going to study a situation when this is not possible. There are several reasons. First, sales depend on many variables other than advertising expenditure: time of year, the state of the general economy, inventory, and price structure. These variables can be included, along with the monthly advertising expenditure, in a model, but then it is still unlikely that we would be able to predict the monthly sales exactly. This happens due to random phenomena that cannot be predicted with certainty. For example, people may stop buying microwave appliances because of new findings concerning the harmful effects of electromagnetic radiation.

If we were to construct a model that hypothesized an exact relationship between variables, it would be called a deterministic model. For example, if we believe that y , the monthly sales revenue, will be exactly 5 times x , the monthly advertising expenditure, we write $y = 5x$. This deterministic relationship implies that y can always be determined when x is known. There is no allowance for error in this prediction. If, on the other hand, we believe that there will be unexplained variation in monthly sales – perhaps caused by important but not included variables or by random phenomena – we discard the deterministic model and use a model that accounts for this random error. This probabilistic model includes both a deterministic component and a random error component. For example, if we hypothesized that the sales Y is related to advertising expenditure x by $Y = 5x + \text{random error}$, we are hypothesizing a probabilistic relationship between Y and x .

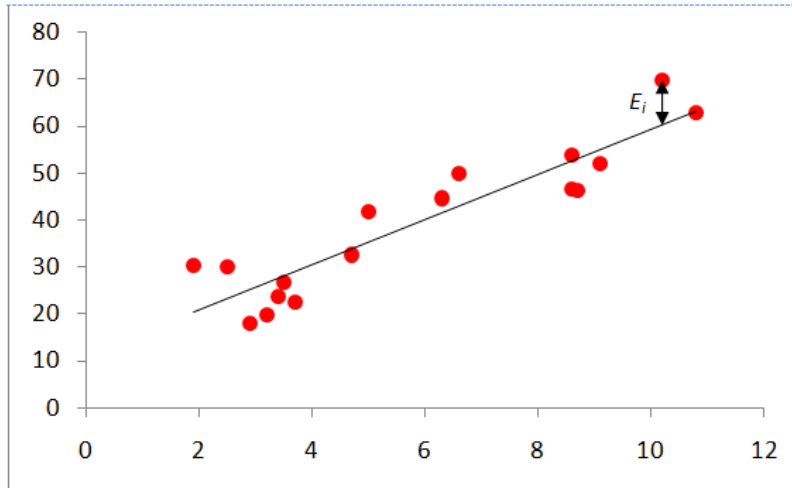
In general the deterministic component may be any function of several variables. The simplest probabilistic model $Y = \beta_0 + \beta_1x + E$ employs a linear function $\beta_0 + \beta_1x$ of one independent variable x as its deterministic component. Here Y is called the dependent variable, x is the independent or predictor variable and E is the random error term. The latter is supposed to have zero mean and finite variance, i.e. to randomly fluctuate around a null value. Then $E(Y) = \beta_0 + \beta_1x + E(E) = \beta_0 + \beta_1x$ implying that expected value of Y follows a straight line $\beta_0 + \beta_1x$. The Greek symbols β_0 and β_1 are the model's parameters. They are not known and we have to estimate them from the available data.

The most common uses of a probabilistic model for making inferences can be divided into two categories. The first is the use of a probabilistic model for estimating the value of Y for a specific value of x which is in the set of our data. The second use of the model usually entails predicting a value for Y corresponding to a new (that is, which is not in the set of data we are dealing with) value of x .

6.1 The least squares approach

Given coupled observations (x_i, y_i) , we now want to find estimates for the parameters which fits to this set of data the best. We start with choosing a mathematical model $y = \alpha_0 + \alpha_1x$. Plotting the above couples in a Cartesian xy plane, we obtain the scatterplot corresponding to this data set. It is very unusual that all points belong to the same straight line. If it does not seem to be possible to have a single straight line passing through all of the couples, we may try to look for a

straight line which deviates the least from them. As a measure for the deviation we may consider the sum of squared distances between the observed couples and the couples predicted by the line $\sum_{i=1}^n (y_i - \alpha_0 - \alpha_1 x_i)^2 = \sum_{i=1}^n E_i^2$. This is the essence of the least squares approach, which tries to find estimates for parameters β for which this quantity is the minimum possible value²¹.



Estimates are $\hat{\beta}_1 = SS_{xy}/SS_{xx}$ and $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$ (remembering from section 4.12 on page 32 the definitions of SS). The straight line given by the linear function $\hat{\beta}_0 + \hat{\beta}_1 x$ is called the least squares line or regression line. The value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ can be considered as a prediction or estimate for y_i .

SPSS: Analyze → Regression → Linear

Since the numerators in the expressions for $\hat{\beta}_1$ and Pearson's correlation coefficient r (see section 4.12 on page 32) are identical, we see that $r = 0$ if and only if $\hat{\beta}_1 = 0$ and that r has the same sign as $\hat{\beta}_1$.

Note that, dividing by SS_{xx} , we are assuming that $SS_{xx} \neq 0$. In fact, SS_{xx} is equal to zero only when all the x values are identical, a case where it is clearly impossible to estimate Y basing on

²¹ In formal terms, given $n \geq 2$ couples (x_i, y_i) , we consider the following function of two arguments $F(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. This is, in fact, a sum of squared deviations of actually observed values y_i from the quantities assigned to points x_i by the linear function $\alpha_0 + \alpha_1 x$. We want to find a couple of values $\alpha_0 = \hat{\beta}_0$ and $\alpha_1 = \hat{\beta}_1$ such that this quantity F has the minimum possible value (note that F is always positive and it is 0 only when all the (x_i, y_i) lie on a line). In order to find this minimum we derive the function F with respect to α_0 and α_1 : $\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 2n(\bar{y}_n - \beta_0 - \beta_1 \bar{x}_n)$ and $\frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2 \sum_{i=1}^n (y_i x_i - \beta_0 x_i - \beta_1 x_i^2)$. Equating the above partial derivatives to zero, we obtain

$$\begin{cases} \bar{y}_n - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_n = 0 \\ \sum_{i=1}^n y_i x_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

The first equation implies that $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$. Substituting this expression in the second equation, we get $\sum_{i=1}^n y_i x_i - n \bar{y}_n \bar{x}_n - \hat{\beta}_1 (\sum_{i=1}^n x_i^2 - n \bar{x}_n^2) = 0$, and, remembering from section 4.12 on page 29 the definitions of SS , $\hat{\beta}_1 = SS_{xy}/SS_{xx}$ and $\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$.

x .

The simplest way to measure the quality of the linear model is to evaluate the contribution of x in predicting y . We define the sum of squares of errors

$$sse = F(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

which is a measure of how close to 0 our errors are. However, this quantity depends strongly on the scale we are using: if we divide all our numbers by 10, this quantity would be reduced by 100! In order to have a scale invariant measure we introduce

$$d = \frac{SS_{yy} - sse}{SS_{yy}} = 1 - \frac{sse}{SS_{yy}}$$

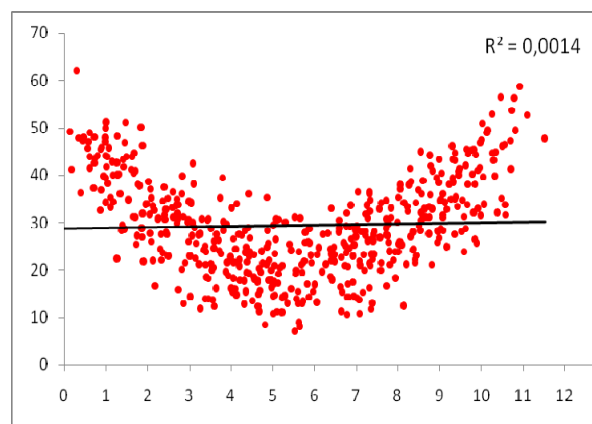
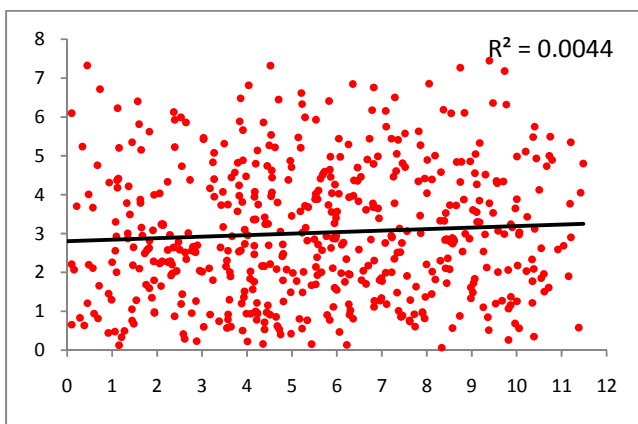
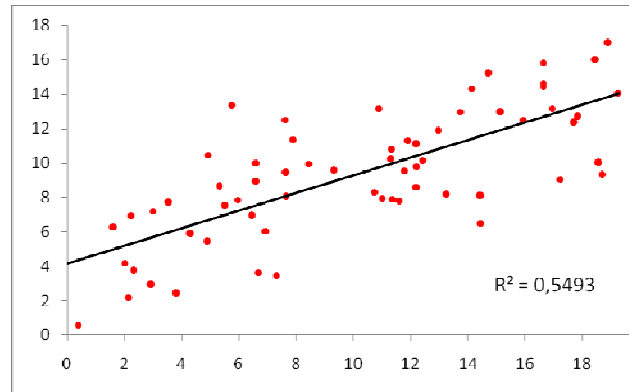
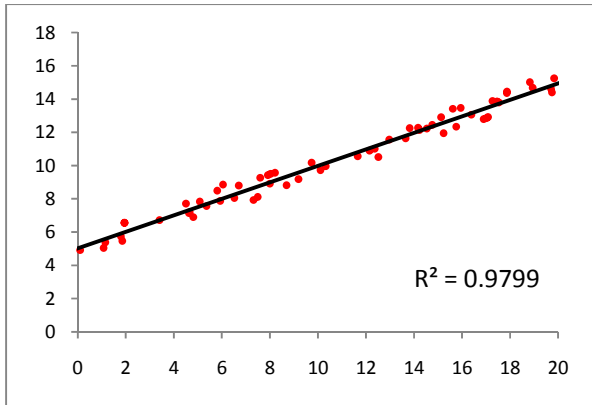
which belongs to $[0; 1]$ and is called coefficient of determination. It is interpreted as the proportion of the total sample variability around \bar{y}_n that has been explained by the linear relationship $Y = \hat{\beta}_0 + \hat{\beta}_1 x$ between x and Y . The difference $SS_{yy} - sse$ shows how much the total variability (when x is not involved at all) has been reduced by using the best possible (in the least square sense) linear approximation. Dividing by SS_{yy} , we get the proportion of this reduction as measured against SS_{yy} . Obviously, the larger is d , the better is the linear approximation. Indeed, a larger d implies a smaller sse . In other words, a smaller deviation of predictions \hat{y}_i from actually observed y_i . For example, $r^2 = 0.6$ means that the sum of squares of deviations of predicted values from actually observed ones has been reduced by 60% by using the least squares linear predictions \hat{y}_i instead of \bar{y}_n .

It can be easily demonstrated²² that coefficient of determination is the square of Pearson's correlation coefficient.

$$d = 1 - \frac{sse}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = r^2$$

This implies also that r , r^2 and $\hat{\beta}_1$ are either all 0 or all different from 0.

²² The expression for $\hat{\beta}_0$ implies that $sse = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n (y_i - \bar{y}_n + \hat{\beta}_1 \bar{x}_n - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n [(y_i - \bar{y}_n) - \hat{\beta}_1(x_i - \bar{x}_n)]^2 = \sum_{i=1}^n (y_i - \bar{y}_n)^2 - 2\hat{\beta}_1 \sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2 = SS_{yy} - 2\hat{\beta}_1 SS_{xy} + \hat{\beta}_1^2 SS_{xx}$. Indeed, we have demonstrated above that $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = SS_{xx}$, but the remaining terms may be treated analogously. Inserting here $\hat{\beta}_1 = SS_{yx}/SS_{xx}$, we get $sse = SS_{yy} - 2 \frac{SS_{yx}}{SS_{xx}} SS_{yx} + \left(\frac{SS_{yx}}{SS_{xx}}\right)^2 SS_{xx} = SS_{yy} - \frac{SS_{yx}^2}{SS_{xx}}$. Hence, $d = 1 - \frac{sse}{SS_{yy}} = \frac{SS_{yx}^2}{SS_{xx}SS_{yy}} = r^2$.



6.2 Statistical inference

So far all calculations have been done without any hypothesis concerning the nature of the dependence between the variables in question. Now we want to turn to the distributions used to evaluate the quality of the least squares estimates obtained above. This calls for more assumptions about the structure of the data:

- whenever we consider a value x_i we have the following relation $y_i = \beta_0 + \beta_1 x_i + E_i$. The values β_0 and β_1 are deterministic and unknown to us. E_i are independent (meaning that E_i is independent from E_j for $i \neq j$) observations of a random variable E normally distributed as $N(0; \sigma^2)$, with zero expected value and a fixed variance (the same for all E_i).

The values $\hat{E}_i = y_i - \hat{y}_i$, are called the residuals. They are the estimates for the outcomes of random variable E . It is interesting to note that residuals have the following feature

$$\sum_{i=1}^n \hat{E}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n [y_i - \bar{y}_n + \hat{\beta}_1(\bar{x}_n - x_i)] = n\bar{y}_n - n\bar{y}_n + \hat{\beta}_1(n\bar{x}_n - n\bar{x}_n) = 0$$

which automatically implies that any \hat{E}_i can be written as a function of the others. Therefore \hat{E}_i are not independent.

Prerequisites: E_i are independent observations of a random variable $E \sim N(0; \sigma^2)$

H₀: β₁ = 0

$$\text{Statistic: } \frac{\sqrt{n-2} \cdot SS_{xy}}{\sqrt{SS_{yy}SS_{xx} - SS_{xy}^2}}$$

Statistic distribution: Student's t with (n - 2) degrees of freedom. For n ≥ 31 standard normal distribution.

SPSS: Analyze → Regression → Linear

Arguing about the usefulness of the simple linear regression model, we want to test whether β₁, of whom we only have the estimation $\hat{\beta}_1$, is equal to zero or not. Indeed, when β₁ = 0, the deterministic part of the model does not change as x varies and therefore the model is totally useless since y does not depend on x. Therefore, we test:

- H₀: β₁ = 0
- H₁: β₁ ≠ 0.

The statistic we use is

$$\frac{\sqrt{n-2} \cdot SS_{xy}}{\sqrt{SS_{yy}SS_{xx} - SS_{xy}^2}}$$

which has the Student's t distribution with n - 2 degrees of freedom. When we reject the null hypothesis, we are sure that the slope of the regression line is not zero and therefore there is a deterministic influence of independent variable over dependent variable. On the other hand, when we accept the null hypothesis we are not able to prove that β₁ is different from zero and we can not argue whether there is an influence or not.

6.3 Multivariate and non linear regression model

We may also assume that the dependent variable Y is a function of k independent arguments. For example, we may try to model the dependence of the annual revenue Y of a firm running supermarkets not only from the advertising expenditure x⁽¹⁾, but also from the money x⁽²⁾ invested in the infrastructure of its shops. Trying to recover a relationship between Y and x^(j), j = 1, 2, ..., k, we can assume that it takes a linear form Y = β₀ + β₁x⁽¹⁾ + β₂x⁽²⁾ + ... + β_kx^(k) + E. Searching for the best values for such β coefficients, we can attempt to use again the least squares approach²³ to estimate the $\hat{\beta}_j$ coefficients.

SPSS: Analyze → Regression → Linear

²³ Introducing $F(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n [y_i - \beta_0 - \beta_1 x_i^{(1)} - \beta_2 x_i^{(2)} - \dots - \beta_k x_i^{(k)}]^2$ and looking for a minimum point, ($\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$) of this function of k + 1 parameters β₀, β₁, ..., β_k. As in the case of the simple linear model, the k + 1 estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are obtained as a unique solution to a system of k + 1 linear equations

$$\begin{cases} \frac{\partial}{\partial \beta_0} F(\beta_0, \beta_1, \dots, \beta_k) = 0 \\ \vdots \\ \frac{\partial}{\partial \beta_k} F(\beta_0, \beta_1, \dots, \beta_k) = 0 \end{cases}$$

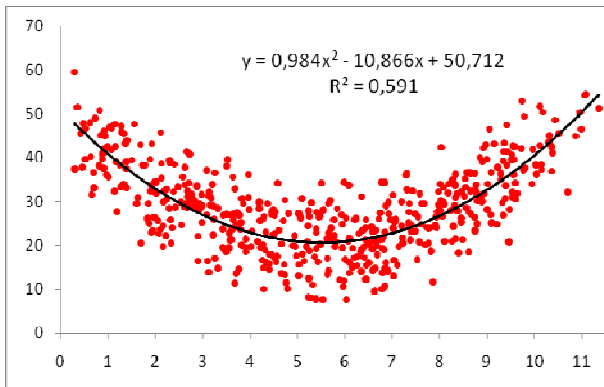
The quality of this approximation may be assessed by looking at the multiple coefficient of determination defined always as

$$d = 1 - \frac{sse}{SS_{yy}}$$

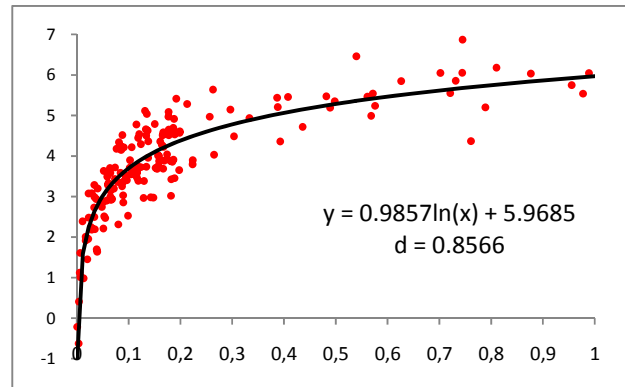
As before, it is interpreted as the ratio of the explained variability to the total variability. Hence the larger is this value, the better fits this linear function to the set of data in question. For multivariate regression models there is no alternative formula and no easy relation with Pearson's correlation coefficients. However, it is still often called r^2 .

The β_0 coefficient has the same meaning as before: it is the predicted value of Y when all the x are equal to 0. Instead, each β_j coefficient is the increment of \hat{y} when the corresponding $x^{(j)}$ increments by 1 unit and at the same time all the other $x^{(1)}, x^{(2)}, \dots, x^{(j-1)}, x^{(j+1)}, \dots, x^{(k)}$ maintain the same value. Thus each β_j for $j \geq 1$ can be seen as the effect of its independent variable on the dependent variable.

In the previous models we have always looked for a linear relation between dependent and independent variables. There are cases, however, where theoretical reasons or the scatterplot itself suggest a non linear relation. For example, the relation may be polynomial $y = \beta_0 + \beta_1x + \beta_2x^2 + \dots + \beta_kx^k + E$ or logarithmic part $y = \beta_0 + \beta_1 \ln x + E$ or exponential $y = \beta_0e^{\beta_1x} + E$ or any complex combination of functions. Clearly, the more complex the function the more coefficients are necessary to be estimated. Estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ may be obtained by the least squares approach as well.



Quadratic model: $Y = \beta_0 + \beta_1x + \beta_2x^2 + E$



Logarithmic model $Y = \beta_0 + \beta_1 \ln x + E$

SPSS: Analyze → Regression → Curve Estimation / Nonlinear

6.4 Multivariate statistical inference

Prerequisites: E_i are independent observations of a random variable $E \sim N(0; \sigma^2)$

$H_0: \beta_j = 0$ for all $j > 0$

Statistic: $\frac{d}{k} \frac{1-d}{n-(k+1)}$

Statistic distribution: Fisher's F with k and $n - (k + 1)$ degrees of freedom.

SPSS: Analyze → Regression → Curve Estimation / Linear / Nonlinear

Since the dependent variable Y now is written as a linear function of k independent variables $x^{(1)}, x^{(2)}, \dots, x^{(k)}$, the model is also termed as a general linear model. Now it is explicitly postulated that the unknown deterministic part is a linear function with unknown coefficients. The value of β_j determines the contribution of the independent variable $x^{(j)}$ and β_0 is the intercept. We define exactly as in the previous case the residuals $\hat{E}_i = y_i - \hat{y}_i$ and we make the same assumptions as before. Now we can test the usefulness of the model

- $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- $H_1: \text{there is at least a } j > 0 \text{ for which } \beta_j \neq 0.$

The statistic we use is

$$\frac{(SS_{yy} - sse)}{k} \bigg/ \frac{sse}{n - (k + 1)} = \frac{\frac{d}{k}}{\frac{1 - d}{n - (k + 1)}}$$

which is distributed as a Fisher's F with k and $n - (k + 1)$ degrees of freedom with only a rejection region on the right. Note that when we reject we know that at least one coefficient is not zero but we do not know which one.

6.5 Qualitative independent variables

People distinguish between two types of data: qualitative and quantitative ones. Quantitative data are recorded in a meaningful numerical scale, whereas qualitative data are measured on a non-numerical or categorical scale. Thus, the Gross Domestic Product, number of sold items, kilowatt per hours of electricity used per day are all examples of quantitative variables. On the other hand, the gender, race, job title, and style of packing are all examples of qualitative variables. The possible values of a qualitative independent variable are referred to as category. For example, the style of packing might have three possible levels: A, B, and C. Even if we designate these values arbitrarily as 1, 2, and 3, then the numbers still represent categories and the variable is still qualitative.

Let us look at regression models with qualitative independent variables. Suppose we want to estimate the mean operating cost per kilometer of cars as a function of the car's manufacturer. Let there be three manufacturers of interest, which we identify as A, B and C. Then the automobile manufacturer is a single qualitative independent variable with three categories, A, B and C. Note that, as always with a quantitative independent variable, we cannot attach a quantitative measure to a given category. Even if we were to call the manufacturers 1, 2, and 3, the numbers would simply be identifiers of the manufacturers and would have no meaningful quantitative interpretation. Our objective is to write a single equation to predict the cost per kilometer based on car's brand. This can be done as follows: $Y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + E$, where

$$x^{(1)} = \begin{cases} 1 & \text{if the car is manufactured by B,} \\ 0 & \text{if the car is not manufactured by B;} \end{cases}$$

$$x^{(2)} = \begin{cases} 1 & \text{if the car is manufactured by C,} \\ 0 & \text{if the car is not manufactured by C;} \end{cases}$$

The variables $x^{(1)}$ and $x^{(2)}$ are not meaningful independent variables as they are in the case of the model with quantitative independent variables. Instead, they are dummy variables that make the model work.

To understand the meanings of the β coefficients, let $x^{(1)} = x^{(2)} = 0$. This condition means that the car is manufactured by A (neither B nor C is manufacturing; hence it must be A). Then the model becomes

$$Y = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + E = \beta_0 + E.$$

Thus, $Y = \beta_0 + E$. Taking the expected value, $E(Y) = \beta_0$ for cars manufactured by A. Therefore β_0 is the expected value for the cost of cars manufactured by A. Now suppose we want to represent the mean cost per kilometer for manufacturer B. Then we should let $x^{(1)} = 1$ and $x^{(2)} = 0$:

$$E(Y) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 = \beta_0 + \beta_1.$$

We have that $\beta_1 = E(Y \text{ produced by B}) - E(Y \text{ produced by A})$. Therefore this coefficient is the expected difference of cost when switching from A to B. Similarly, $\beta_2 = E(Y \text{ produced by C}) - E(Y \text{ produced by A})$.

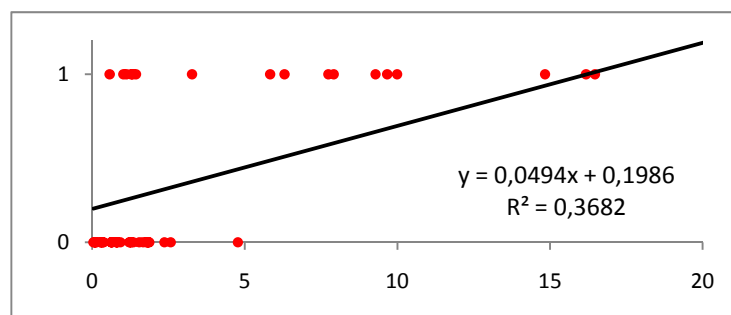
Note that we are able to describe three categories of the qualitative variable with only two dummy variables. This is because the base level (manufacturer A, in this case) is accounted for the intercept β_0 . In general therefore for each qualitative variable we require an amount of dummy variables equal to the categories minus 1.

Since a model with dummy variables is a multivariate regression model, everything we said concerning usefulness tests also applies here. Moreover, it can be freely mixed with quantitative linear and non linear model's components.

SPSS: dummy variables are handled automatically if variables are nominal or ordinal

6.6 Qualitative dependent variable

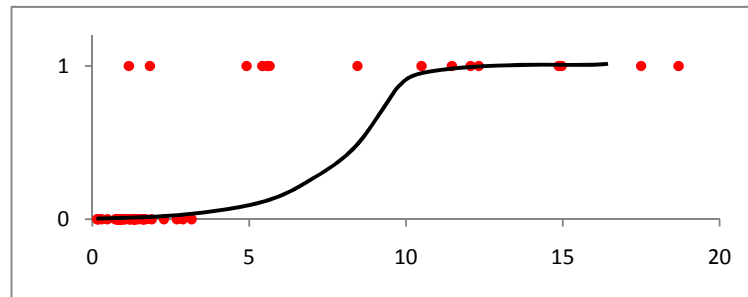
It is also possible to build a regression model with a qualitative dependent variable, provided that it has only two categories arbitrarily indicated with 0 and 1. The difficulty of this model is the fact that the right side of the model's term $f(x^{(1)}, x^{(2)}, \dots, x^{(k)})$ provides continuous values while the left side has only two possible values. For example a linear regression model would yield silly results, larger than 1 or smaller than 0, such as



Therefore instead of using $f(x^{(1)}, x^{(2)}, \dots, x^{(k)})$ as estimation function, we use its logit

$$Y = \frac{1}{1 + e^{-f(x^{(1)}, x^{(2)}, \dots, x^{(k)})}} + E.$$

The function on the right side now goes from 0 to 1. Its shape is much better suited for interpolating values which are always 0 or 1:



Values between 0 and 1 can be interpreted as the probability for the dependent variable to take value 1.

SPSS: Analyze → Regression → Binary Logistic

6.7 Problems of regression models

6.7.1 Number of observations

In order to work correctly, the least squares approach needs a number of observations that is at least equal to the number of parameters it needs to estimate. If this condition is not satisfied, the linear equations system does not have a unique solution and thus parameters' estimates cannot be determined. In practice however, *the number of observations must be much larger than the number of parameters*: as an empirical rule, observations should be at least 10 times the number of parameters used by the model. This is because whenever we add a parameter to the model we get always a larger d ; this seems to indicate that the new model is better, but it is only more complex, i.e. it is simulating reality not simplifying it but simply auto adapting itself. If we build a regression model to understand reality and not simulating it, keeping the model simple must be our priority.

6.7.2 Multicollinearity

Multicollinearity exists in a multivariate regression model when one or more of the independent variables used in regression depend in a deterministic way on each other. In this case the corresponding independent variables contribute redundant information. For example, suppose we want to construct a model to predict the fuel cost of a truck as a function of its load $x^{(1)}$ and the power $x^{(2)}$ of its engine. Clearly, these two variables are dependent since usually a powerful truck carries huge loads. Although both $x^{(1)}$ and $x^{(2)}$ contribute information for the prediction of fuel cost, the contributions are tautological and in the model β coefficients will not reflect the effect of each independent variable.

A simple way to detect multicollinearity is to calculate the Pearson's correlation coefficient between each pair of independent variables in the model. When a calculated value differs significantly from zero, the variables in question are related and a multicollinearity problem exists.

6.7.3 Dependent errors and Durbin-Watson test

As we have seen in section 6.2, residuals are dependent. Residuals are the estimations of the outcomes of random variable E , which in order to perform usefulness tests, must be all independent. Since their estimates are dependent, we may legitimately argue that the independence hypothesis may not hold. In fact, there are many practical situations where it does not hold, in particular when the observed cases are taken at different times, since cyclical component of a time series may result

in deviations from the secular trend that tend to cluster alternately on the positive and negative sides of the trend. For example, if our dependent variable is the monthly Gross Domestic Product of a country, its cases are taken at different times and their values may be influenced by cyclical fluctuations which, not predicted by the deterministic side of the model, end up influencing the errors which are therefore no more independent.

Prerequisites: $E_t \sim N(0; \sigma^2)$

H₀: E_t are independent

Statistic:
$$\frac{\sum_{t=1}^{n-1} (\hat{E}_t - \hat{E}_{t+1})^2}{\sum_{t=1}^n \hat{E}_t^2}$$

Statistic distribution: Durbin-Watson's table

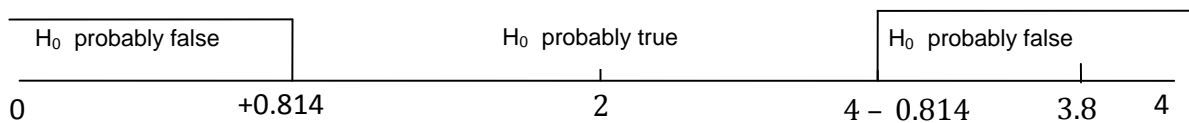
SPSS: Analyze → Regression → Linear → Statistics

Supposing now that the observations are taken at different times, and thus using t as case index, we want to test

- H₀: E_t and E_{t+1} are not autocorrelated for every t
- H₁: E_t and E_{t+1} are autocorrelated for at least a t

The statistic is $\frac{\sum_{t=1}^{n-1} (\hat{E}_t - \hat{E}_{t+1})^2}{\sum_{t=1}^n \hat{E}_t^2}$ and takes values from 0 to 4. Value 0 corresponds to a situation where all the \hat{E}_t are constant, and thus a perfect correlation. On the other hand, when $\hat{E}_t = -\hat{E}_{t+1}$, we have $\frac{\sum_{t=1}^{n-1} (-2\hat{E}_t)^2}{\sum_{t=1}^n \hat{E}_t^2} = \frac{4 \sum_{t=1}^{n-1} (\hat{E}_t)^2}{\sum_{t=1}^n \hat{E}_t^2} \approx 4$, which therefore corresponds to a perfect negative correlation. A value of 2 is the uncorrelation's value, where H₀ is accepted. Critical values are found in Durbin-Watson's table.

For example, if we get a Durbin-Watson's statistic value of 3.8 for $n = 15$ and $k = 3$ (multivariate model with 3 independent variables) we get a critical value of 0.814. This means that the right critical value is $4 - 0.814 = 3.186$ and therefore we reject. This means that errors are autocorrelated.



If we manage to prove that the autocorrelation is not zero, then automatically E_t and E_{t+1} are correlated and thus (since independence implies zero correlation) they cannot be independent and usefulness' test cannot be performed since hypotheses do not hold true.

On the other hand, if the autocorrelation is zero, we cannot directly deduce that E_t and E_{t+1} are independent (since zero correlation does not imply independence). However, when assumption $E_t \sim N(0; \sigma^2)$ holds, zero correlation does imply independence. Therefore, if null hypothesis is accepted, we can hope that errors be independent.

In any case, *even with dependent errors, the regression model continues to work and the determination coefficient continues to have its meaning.* The only thing than cannot be performed is the usefulness test.

6.7.4 Heteroskedasticity

Let the errors E_i in a simple linear regression model have a varying variance, that is even though all of them are independent and normally distributed with zero expected value, they come from random variables with a different variance σ_i^2 . Usually this happens whenever the data come from observations which vary a lot in size, as it is the case when our cases are firms of different sizes, or when data come from aggregated values, such as averages or sums.

In the case when we are able to know a priori the values σ_i^2 , we may build a new model for which errors' variances are the same. If we divide the model, written for each i , by σ_i , we get

$$\frac{y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{E_i}{\sigma_i}$$

and calling $y_i^* = \frac{y_i}{\sigma_i}$, $x_i^* = \frac{x_i}{\sigma_i}$, $E_i^* = \frac{E_i}{\sigma_i}$ it becomes

$$y_i^* = \beta_0 \frac{1}{\sigma_i} + \beta_1 x_i^* + E_i^*$$

which is a particular multivariate regression model with two independent variables, $\frac{1}{\sigma_i}$ and x_i^* , and no intercept. If we calculate the variance of E_i^* we get that they are all the same:

$$\text{Var}(E_i^*) = \text{Var}\left(\frac{E_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} \text{Var}(E_i) = 1.$$

Therefore we can use the new model, which does not present the heteroskedasticity problem, to estimate y_i^* and then multiply by σ_i to get the y_i estimations.

Since in practice the value of the standard deviation of errors is unknown, it is common to divide the regression model by a scale quantity which represent an estimate of the standard deviation, for example the size of the firm (number of employees or total budget).

A typical case where the scaling factor is known are aggregated data with heteroskedasticity. Let y_{ij} be the observation on the j th case in the i th group, and consider the following regression

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + E_{ij}$$

If we do not have the cases' single values but only aggregate observations on each group are available, then these expressions are summed over cases, that is

$$y_i = \beta_0 n_i + \beta_1 x_i + E_i$$

Here $y_i = \sum_{j=1}^{n_i} y_{ij}$, $x_i = \sum_{j=1}^{n_i} x_{ij}$, $E_i = \sum_{j=1}^{n_i} E_{ij}$. If the original errors E_{ij} , those referred to cases, satisfy our assumptions, i.e. are independent identically distributed with expected value 0 and variance σ^2 , then the errors we use in the model E_i are still independent identically distributed with expected value 0 but with variance $n_i \sigma^2$ since

$$\text{Var}(E_i) = \text{Var}\left(\sum_{j=1}^{n_i} E_{ij}\right) \stackrel{E_{ij} \text{ independent}}{=} \sum_{j=1}^{n_i} \text{Var}(E_{ij}) = \sum_{j=1}^{n_i} \sigma^2 = n_i \sigma^2.$$

This means that the errors in our aggregated model are now heteroskedastic. However, we know their variances $\sigma_i^2 = n_i \sigma^2$. Hence, we have to divide the model by $\sigma \sqrt{n_i}$ and perform the ordinary

least squares estimates on the transformed equation. In practice, in this case we do not know the value of σ but we may simply divide by $\sqrt{n_i}$, a value that we know since it is the number of firms in every sector. We will get a model with constant errors' variance σ .

In any case, *even with dependent errors, the regression model continues to work and the determination coefficient continues to have its meaning*. The only thing than cannot be done is the usefulness problem.