STENDEN HOGESCHOOL  -  HONOURS PROGRAMME

# *HYPOTHESIS TESTING  : BASIC CONCEPTS  AND MOST COMMONLY USED SIGNIFICANCE TESTS*

**Rob Flohr MSc MA**
**Lecturer Statistics, Mathematics, Philosophy of Science**
**2009/2010**

CONTENTS 2

## Introduction

The current theory of testing hypotheses was developed by statisticians like R.A. Fisher (1890-1962), J. Neyman (1894-1981) and E.G. Pearson (1895-1980, son of the famous statistician Karl Pearson, 1857-1936).
R.A. (Sir Ronald Aylmer) Fisher, the eminent statistician, was inspired by the idea that evidence may have a decisive negative impact on a statistical hypothesis, akin to its falsification. He expressed the view that the null hypothesis is never proved or established, but is possibly disproved in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. Fisher was inspired by both the Popperian falsificationist outlook and the ideal of objectivity when, building on the work of Karl Pearson and W.S. Gossett ('Student'), he developed his system of significance tests for testing statistical theories (Howson and Urbach 2006).
The problem which Popper had to face, was that statistical hypotheses ascribe probabilities to possible events and do not say of any hypothesis that they will or will not actually occur. So the simple mechanism by which a hypothesis may be logically refuted by observational evidence, could never work for this kind of hypotheses. Popper tried to solve this by stating that, though a statistical hypothesis may not be strictly falsifiable, it is 'practically falsified' when an event occurs to which the hypothesis attaches a sufficiently small probability. But this does not hold. Take the simple experiment of tossing a fair coin 10000 times. The probability of any sequence of heads and tails has a minuscule value ( $2^{-10000}$), yet it is the probability of every possible outcome of the experiment, one of which will definitely occur.
Fishers proposal was, roughly speaking, that a statistical hypothesis should be rejected by experimental evidence when it is , on the assumption that the hypothesis is true, contained in a certain set of outcomes that are relatively unlikely, relative, that is, to other possible outcomes of the experiment. Somewhat later, Neyman and Pearson made a step forward by introducing the element of an alternative hypothesis in the process of testing a 'null' hypothesis. This might be interpreted in terms of the philosophy of Imre Lakatos according to which not isolated theories and hypotheses are tested, but competing theories and hypotheses in the context of a specific 'research programme' (Dooremalen et al. 2007).

One might conceive this invention in the context of a history of 'the process of inference', of a particular empirical mind that tries to find solutions while dealing with a great number of concrete observations. For a long time this took the form of rather simple measurement in two basic forms: systematization and quantification of observations.
The Dutch notion '*statistiek'* was German in origin (*Statistik*). It came into broad academic use around 1750 through Gottfried Achenwall (1719 – 1772) who was professor at the university of Göttingen. Achenwall developed '*Statistik'* or *Staatenkunde* into a method of structuring remarkable facts concerning the state. It was a method of inquiry to ascertain the political strength of a country and to find out the real state of a state. (Klep and Stamhuis 2002).
In this context it will not be surprising that the Italian noun '*statista'* (plural: '*statisti'*) means *statesman*.

Statistics is based on the concept of probability or chance. In the sixteenth century, the game of dice, or craps as we might call it today, was subjected for the first time to a formal mathematical study by the Italian mathematician Gerolamo Cardano (1501 – 1576). By the end of the seventeenth century, the Dutch astronomer Christiaan Huygens (1625 – 1695) laid the foundation for current probability theory. His work unified various problems that had been solved earlier by the famous French mathematicians Pierre Fermat and Blaise Pascal. Historical records show that there was no real concept of probability in Europe before the mid-seventeenth century, although the use of dice and other randomizing objects was commonplace.

Although probability theory was initially the product of questions posed by gamblers, nowadays countless problems in our daily lives call for a probabilistic approach.
It is essential to the field of insurance, the stock market, “the largest casino in the world”, cannot do without it, the telephone network with its randomly fluctuating load could not have been economically designed without the aid of probability theory, the quality of judicial and medical decisions will in many cases be improved by applying an elementary knowledge of probability theory, airline companies apply probability theory to determine how many service desks will be needed based on expected demand, engineers use probability theory when constructing dikes to calculate the probability of water levels exceeding their margins, etc.
To cut a long story short, probability has become an integral part of our lives (Tijms 2007).

In this file first a brief sketch of the concept of the procedure of hypothesis testing and of related concepts is offered(Chapter I). In Chapters II and III a summary of the most common statistical tests is to be found, related to the type of research question as well as the level of measurement of the test variable(s). The several statistical tests are illustrated with the help of SPSS. Some statistical tests have been elaborated more extensively, the intention is to offer some exemplary illustrations of the application of the procedure of statistical testing and the use of the notion ‘*significant*’.
Statistical tests can be subdivided into parametric and non-parametric tests. *Parametric tests* are based on the assumption of a normally distributed sampling distribution and the variables in question are measured at the interval/ration (SPSS: scale) level.

In case of *non-parametric (or distribution-free) tests*, no distributional assumptions are made about the population under investigation, and the variables in question are measured at the nominal or ordinal level. These kind of tests often make use of the differences between the rank orders of the collected data instead of the differences between the original numerical data. They are less ‘powerful’ compared to parametric tests, which means that an untrue null hypothesis will be rejected less quickly.

Examples of non-parametric tests are: Chi-square test, Mann-Whitney U test, Kruskal-Wallis test, Wilcoxon signed-ranks test.

# I Basic Concepts

When you intend to make a confident statistical probability statement about a specific population on the basis of a (random) sample, taken from that population, you are dealing with inferential statistics.

The statement in question may refer to:

a) the value of the population mean or the population proportion

b) the significance of your sample outcome

ad a) Interval estimation: your sample result is a point estimator of a population mean or a population proportion, and the use of an interval estimate provides a measure of the precision of an estimate, according to the form: point estimate ± margin of error.
Key formulae:
$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ (population mean) or $p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ (population proportion)

ad b) Determining whether your sample outcome is significant

The first step is to describe your sample outcome by means of the relevant descriptive statistics: counts, frequencies, the mean, the standard deviation, cross tabulation etc., depending on the type of research.

The second step involves determining the significance of your sample outcome. We can distinguish the following stages:

a) We start formulating two hypotheses; the null hypothesis ( $H_0$ ) which contradicts or opposes the researcher's theoretical expectation, and the alternative hypothesis ( $H_1$ or $H_a$ ) or research hypothesis which is the counterpart of the null hypothesis, and which is frequently derived from new scientific insights or from new theories.
b) The question to be answered is: should, on the basis of our sample outcome, the null hypothesis be rejected or not?
In science it is standard to rigidly test the alternative hypothesis, requiring very convincing evidence before the null hypothesis is rejected and the alternative hypothesis accepted. To answer the question mentioned above, we have to calculate the probability that an outcome as extreme as our sample outcome, or more extreme than that, will occur when the null hypothesis is true.
c) The calculation of this probability is based on a specific probability distribution, depending on the type of research and the level of measurement of the test variable (for example: the normal distribution, the t-distribution (Student's t-distribution), the Chi-square distribution,

the F distribution, etc. etc.). These are instances of the so-called sampling distribution: a probability distribution showing the various possible values of the sample statistic as a result of different random samples , in other words: the probability distribution of the sample statistic in question.

d) Then we transform our sample outcome into a value, called the test statistic, on the basis of a formula that is related to the probability distribution in question. For example: in case of a t-distribution, the corresponding test statistic is found by the following formula:

$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$, and in case of the Chi-square distribution, the test statistic is calculated by

$$X^2 = \sum \frac{(O - E)^2}{E}$$

(O being the observed counts and E the expected counts)

e) Before calculating the probability, we will have to decide which level of significance (notation: $\alpha$ , alpha) ) to use. Common levels of significance used are 10%, 5%, and 1%. These levels constitute rejection areas: the null hypothesis is rejected when the sample outcome, as translated into the test statistic, falls into the rejection area. The rationale of this being that it is highly improbable that the population parameter is equal to the value as hypothesized in the null hypothesis.

f) The exact probability that our sample result or an even more extreme sample result, as transformed into the test statistic, will occur under the condition that the null hypothesis is true, is called the ‘p-value’ (‘p’ is derived from ‘probability’). This p-value can be either one-tailed or two-tailed, and can be calculated with any statistical software package (for example SPSS).

The one-tailed p-value is the probability that the sample result, or an even more extreme sample result, is found while the null hypothesis is assumed to be true. The two-tailed p-value is twice the size of the one-tailed level and is used when the alternative hypothesis is non-directional.

- When the one-tailed p-value is less than or equal to the level of significance, the null hypothesis is rejected and the directional alternative hypothesis is accepted.
- When the two-tailed p-value (being twice the size of the one-tailed p-value) is less than or equal to the level of significance, the null hypothesis is rejected and the non-directional alternative hypothesis is accepted.

In both cases the value of the test statistic (based on our our sample result) is said to be significant.

The set of values of the test statistic that lead to the rejection of the null hypothesis is called the critical region or rejection region, and the set of values of the test statistic that do not lead to the rejection of the null hypothesis is called the acceptance region. The size of the critical region is of course determined by the desired significance level (alpha) of the test.

g) There are two cases when the test leads to a correct result. These occur when the null hypothesis is true and the test leads to its acceptance, and when the alternative hypothesis is true and the test leads to the rejection of the null hypothesis.

On the other hand, there are two cases when the test leads to an incorrect result. These occur when the null hypothesis is true but the test leads to its rejection (a Type I error or error of the first kind), and when the alternative hypothesis is true but the test leads to acceptance of the null hypothesis (a Type II error or error of the second kind). The above mentioned significance level ( $\alpha$ , alpha) is the probability of making a type I error (can you see this?).

The probability of making a Type II error is often denoted by $\beta$ (beta).
The power (Dutch: 'onderscheidingsvermogen') of the test, which is the probability of rejecting the null hypothesis when the alternative hypothesis is in fact true, is 1 – $\beta$ .

Summarizing:

| | Accept null hypothesis | Reject null hypothesis |
|---|---|---|
| Null hypothesis is true | Correct decision<br>$P(correct / decision) = 1 - \alpha$ | Type I error<br>$P(TypeI / error) = \alpha$ |
| Alternative hypothesis is true | Type II error<br>$P(TypeII / error) = \beta$ | Correct decision<br>$P(correct / decision) = 1 - \beta$<br>('Power') |

To find $\beta$, you need to assume a specific value for the alternative hypothesis, and then calculate the p-value for the critical value of the test statistic under the condition that the alternative hypothesis is true.

Note:
-As to reporting your results in your research report, it is common to mention the statistical test involved as well as some values (the sample size n, the value of the test statistic, the p value, etc.). See for examples the 'Friedman Test' and the 'One-way analysis of variance'.

-In some SPSS outputs you find the expression "Asymp. Sig." (Asymptotic significance). The distribution of a statistic is said to be asymptotically normal if, as the sample size increases, the distribution of the statistic approaches a normal distribution.

Finally, after having discussed the technical part of interval estimation and hypothesis testing, we should reflect on the meaning of statistical statements.
It is important to realize, that the meaning of the concept of probability basically concerns a philosophical issue (see for example Gillies 2000, Cohen 1989, Mellor 2005). Different conceptions of probability are possible, and are actually employed. One of these conceptions is the so-called 'frequentist' approach, which underlies the classical theory of statistical interval estimation and

hypothesis testing  as developed by Fisher, Neyman and Pearson, that has been explained above. According to this frequentist conception, probability has to be conceived of as a limit of a relative frequency, in the context of repeating the chance experiment many times (compare the experiment of rolling a fair die: only after rolling the die many times, the relative frequency of rolling a six, for example, will tend to 1/6). Besides, according the frequentist approach, a population parameter is no random variable, but a fixed, though unknown, quantity. That means that it is not possible to attach any probability to that population parameter. So, it is incorrect to make the following statements:

“This confidence interval will, with a probability of 95%, contain the population parameter”

respectively

“The probability that the null hypothesis is true is 95%”.

The correct statements are

“If we would repeat the procedure of computing a confidence interval over and over again, 95% of the computed confidence intervals would contain the population parameter”

respectively

“If we would repeat the procedure of testing the null hypothesis over and over again, in 5% of all cases (assuming an α of 0.05) the null hypothesis would be rejected unjustly”.

You can see that the probabilities found do not refer to the confidence interval or the null hypothesis itself, but to the procedure which is used to compute a confidence interval and to test the null hypothesis.
This implies that we have to interpret the classical theory of interval estimation and hypothesis testing primarily as a (rational) decision procedure concerning statements about reality, which is a valuable instrument indeed. You might say that the context is a methodological one, and not an ontological one in which truth and falsehood would be the leading concepts.

According to Bayesian statistics, probability is conceived of as a personal degree of belief, and population parameters are treated as random variables. This implies that, in the context of Bayesian statistical analysis, probabilities can be attached to confidence intervals and hypotheses. However, it is beyond the scope of this document to discuss Bayesian statistics at length.

## II Comparative Research Questions (mainly statistical tests for comparing means)

Chi-square test
Mann-Whitney U test
Kruskal-wallis test
Wilcoxon Signed-ranks test
Friedman test
(Student’s) t  test (Independent-samples t  test)
One-way ANOVA
T test for two paired samples (Paired-samples t  test)

| Grouping variable: NOMINAL | Test variable: NOMINAL |
|---|---|

## Chi-Square test

To be used when:

you want to investigate if a significant difference exists between two (or more) groups with respect to one characteristic measured at the nominal level.
The Chi-Square test can be used to investigate if such a significant difference exists.

To use the Chi-Square Test, some conditions have to be satisfied:
- any given expected cell frequency may not be smaller than 1
- a minimum of 80% of the expected cell counts must have values greater than 5
- the variables must not have too many categories (otherwise the table becomes incomprehensible and the first two conditions are also likely not satisfied).

*SPSS:*
*Contingency Table:*
*Each cell of the contingency table compares the number of observed cases (O) with the number that might be expected on the basis of chance (the expected cell counts: E). When the observed cell counts (O) deviate from the expected cell counts (E), the interesting question is: is this deviation great enough to be significant and not merely dependent on chance? By calculating the test statistic Chi-Square (* $X^2$ *), according to the formula:*

$$X^2 = \sum \frac{(O-E)^2}{E}$$

*we find a specific value for* $X^2$ *: the greater the* $X^2$ *, the smaller the probability that the deviation results from chance. SPSS does not only give the* $X^2$ *, but also the corresponding p-value (which is not only dependent on the size of the* $X^2$ *, but also on the number of degrees of freedom:*
$d_f = (c-1)(r-1)$ *with c = number of columns and r = number of rows).*
*The null hypothesis under the Chi-Square Test states that there exists no difference between the groups with respect to the characteristic in question.*

*ANALYZE ; DESCRIPTIVE STATISTICS ; CROSSTABS ; (define the grouping variable in 'Colum(s)' and the test variable in 'Row(s)') ; STATISTICS ; CHI-SQUARE ; CONTINUE ; OK*

*SPSS OUTPUT*
*(example: is there a difference between men and women insofar as marital status is concerned?)*

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| Marital/family status * Sex | 996 | 99,6% | 4 | ,4% | 1000 | 100,0% |

**Marital/family status * Sex Crosstabulation**

Count

| | | Sex | | Total |
|---|---|---|---|---|
| | | man | woman | |
| Marital/family status | alone | 61 | 89 | 150 |
| | with partner | 168 | 208 | 376 |
| | with partner and children | 270 | 200 | 470 |
| Total | | 499 | 497 | 996 |

**Marital/family status * Sex Crosstabulation**

| | | | Sex | | Total |
|---|---|---|---|---|---|
| | | | man | woman | |
| Marital/family status | alone | Count | 61 | 89 | 150 |
| | | Expected Count | 75,2 | 74,8 | 150,0 |
| | with partner | Count | 168 | 208 | 376 |
| | | Expected Count | 188,4 | 187,6 | 376,0 |
| | with partner and children | Count | 270 | 200 | 470 |
| | | Expected Count | 235,5 | 234,5 | 470,0 |
| Total | | Count | 499 | 497 | 996 |
| | | Expected Count | 499,0 | 497,0 | 996,0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 19,904[a] | 2 | ,000 |
| Likelihood Ratio | 19,981 | 2 | ,000 |
| Linear-by-Linear Association | 18,309 | 1 | ,000 |
| N of Valid Cases | 996 | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 74,85.

*Explanation:*

*The difference in the marital status of men and women is clearly significant. Chi-Square ( $X^2$ ) is 19.90, and the probability that the observed marital status difference between men and women at two degrees of freedom results from chance is 0.000. You can conclude with a certainty greater than 95% (in this case even greater than 99.9% ) that a significant difference exists in the marital status of the men and women. (The Likelihood Ratio is comparable to the Chi-SquareTest).*

*To put it otherwise: if the null hypothesis (there is no difference between men and women insofar as marital status is concerned ) is true, then the probability of obtaining a Chi-Square that is as extreme or more extreme than 19.90 is 0.000 (p-value). Assuming a level of significance ( α ) of 0.05, we reject the null hypothesis because the p-value is smaller than α.*

*(Note that the sampling distribution of $X^2$ is approximated by the $\chi^2$ -distribution)*

| | |
|---|---|
| Grouping variable: NOMINAL | Test variable: ORDINAL |

Two independent samples: Mann-Whitney U test

To be used when:

you want to investigate if a difference between two independent samples concerning an ordinal test variable is due to chance. So, the grouping variable comprises two independent samples and the test variable has been measured at at least the ordinal level.
By applying the Mann-Whitney U test, you are testing the null hypothesis that the two samples originate from identical populations (populations with identical distributions).
The several cases from both samples are joined together and then ranked. Subsequently, the rank orders of both samples are separately totaled. If both samples originate from the same population, the totaled rank scores will be about equal.

The test statistic U refers to the extent to which the rank scores are different from each other. Whether or not this difference is significant depends on the sample sizes and the extent of the differences, as is shown by the formulas:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

with $n_1$ and $n_2$ being the sample sizes, and $R_1$ and $R_2$ being the totalled rank scores of both samples.

*SPSS:*
*ANALYZE; NONPARAMETRIC TESTS; 2 INDEPENDENT SAMPLES; indicate the variables you wish to include in the analysis (your test variable in the 'Test Variable List', the grouping variable in 'Grouping Variable'); using "Define groups", you can indicate which categories or groups are to be distinguished. Type 1 in the field next to "Group 1" and 2 in the field next to "Group 2". Sex is consequently numerically coded with a number 1 for men and a number 2 for women. CONTINUE; OK*

*SPSS OUTPUT*
*(example: is there a difference between men and women in the extent that they are happy about the life that they are leading?)*

**Ranks**

| | Sex | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| THAP | man | 484 | 571,97 | 276831,50 |
| | woman | 500 | 415,58 | 207788,50 |
| | Total | 984 | | |

**Test Statistics**[a]

| | THAP |
|---|---|
| Mann-Whitney U | 82538,500 |
| Wilcoxon W | 207788,500 |
| Z | -8,906 |
| Asymp. Sig. (2-tailed) | ,000 |

a. Grouping Variable: Sex

*Explanation:*

*The first table indicates the number of respondents for both groups. The mean happiness rank is 571.97 for men and 415.58 for women. This makes it clear that the women score lower in happiness, insofar as the mean rank is concerned. In addition to the mean rank, the sum of ranks is also indicated. This test is based on the smallest sum (207788.50). The second table (Test Statistics table) lists the test statistic U and the z-value derived from U. The U value is calculated on the basis of the smallest sum of the ranks, along with the sample sizes. This U value can, with the aid of a formula, be converted into a z value. The (two-tailed) probability of error corresponding to the U value of 82538.500 is 0.000. Consequently, the probability that the  difference between men and women is a result of chance is smaller than 1/1000 and the difference is therefore significant.*

*If, on the basis of certain considerations, you might expect that women are unhappier than men (or vice verse), then you will need to conduct a one-tailed test.*

| Grouping variable: NOMINAL | Test variable: ORDINAL |
|---|---|

Three or more independent samples: Kruskal-Wallis test

To be used when:

more than two groups are involved in your study and you want to investigate if a difference between three or more independent samples concerning an ordinal test variable is due to chance. So, the grouping variable comprises three or more independent samples and the test variable has been measured at at least the ordinal level.
By applying the Kruskal-Wallis test, you are testing the null hypothesis that the (three or more) samples originate from identical populations (populations with identical distributions).

The Kruskal-Wallis test is based on the same principle as the Mann-Whitney U test. The various groups of respondents or cases (for example visitors of a shopping centre who travel by bike, by car and by public transport) are mixed together to form a total group. All respondents are then ranked and are therefore given a rank score on the test variable (for example income). The rank scores are then totalled for each group and the mean rank score calculated. The differences between them provide the basis on which the test statistic $H$ is calculated and subsequently converted into a Chi-square value.

(test statistic $H$ : $H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{{R_j}^2}{n_j} - 3(N+1)$ with

$k$ = number of samples (groups)

$n_j$ = number of observations in the $j$ th sample

$N$ = total number of observations

$R_j$ = total of the ranks in sample $j$

Arrange these $N$ observations in order of size and replace their values with the corresponding ranks. Under the null hypothesis, for large $N$, with no $n_j$ small, $H$ has an approximate chi-squared distribution with $k-1$ degrees of freedom).

*SPSS:*
*ANALYZE ; NONPARAMETRIC TESTS ; K INDEPENDENT SAMPLES ; indicate the test variable and the grouping variable ; use "Define Range" to indicate the range of the groups: in our example we have three groups, so the range is from 1 to 3. CONTINUE ; OK*

*SPSS Output*
*(example: we want to compare visitors of a shopping centre travelling by bike, by car and by public transport with respect to their level of income)*

Ranks

| | meest gebruikte vervoermiddel | N | Mean Rank |
|---|---|---|---|
| netto maandinkomen huishouden | fiets | 27 | 21,09 |
| | auto | 28 | 53,71 |
| | openb vervoer | 21 | 40,60 |
| | Total | 76 | |

**Test Statistics**[a,b]

| | netto maandinkomen huishouden |
|---|---|
| Chi-Square | 30,315 |
| df | 2 |
| Asymp. Sig. | ,000 |

a. Kruskal Wallis Test

b. Grouping Variable: meest gebruikte vervoermiddel

*Explanation:*
*The Test Statistics table shows a Chi-square value of 30,315 and two degrees of freedom (the number of groups minus 1, so in our example: 3-1 = 2). We reject the null hypothesis that the three samples originate from identical populations because the p-value (significance) is smaller than 0.05 and conclude that the income distributions of the three groups are not identical.*

| Grouping variable: NOMINAL | Test variable: ORDINAL |
|---|---|

Two paired (related) samples: Wilcoxon Signed-ranks test

To be used when:

you want to find out if a difference on an ordinal variable (for example, the extent to which respondents feel themselves to be happy) between two paired samples (for example, husbands and their wives) can be attributed to chance.
So, two related or paired (not independent) samples are involved and the test variable is measured at at least the ordinal level.

Independent and dependent (paired) groups (samples).

In case of independent groups, such as a random sample from a population of women and a random sample from a population of men, the random selection of women from the population does not determine which men are selected from the population in any way.

Two groups are said to be statistically dependent when each unit of analysis (often respondents) within the first group is somehow related to a unit in the second group. For obvious reasons these groups are often referred to as paired groups. Consider a random sample of adult women (group 1) and a second group consisting of their mothers. The goal of such a design could be to determine differences in occupational careers. Another example is a random sample of respondents interviewed at two moments in time; for example, during elections held in 2003 and in 2006. A third example is the comparison of two variables, such as the results on a language test (group 1) and a math test (group 2), while both groups contain the same respondents. A typical characteristic of these three examples is that there is interdependency between the (paired) observations.

By applying the Wilcoxon Signed-ranks test, you are testing the null hypothesis that there is no difference between two paired distributions.
First, the difference between the individuals in each pair of the test variable (in the above mentioned example of happiness, this is the difference of the happiness variable). The absolute differences are then ordered from lowest to highest. Next, the rank scores of all negative differences are summed and similarly all the positive differences. The mean positive rank score is finally compared with the mean negative rank score.
The greater the number of pairs in which the two members are different in terms of the test variable and, especially, the more frequently that these differences are in the same direction, the smaller is the chance that a difference in the test variable between two groups is a result of chance. This is even more the case when the sample is larger. The extent to which the samples differ is converted into a Z-value.

(test statistic : $Z = \dfrac{\frac{1}{4}n(n+1) - T - \frac{1}{2}}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$ with:

$n$ = number of differences

$T$ = the smaller of $P$ , being the sum of the positive signed ranks, and $\frac{1}{2}n(n+1) - P$ )

*SPSS:*
*ANALYZE ; NONPARAMETRIC TESTS ; 2 RELATED SAMPLES ; using the “Test Pair(s) List”, you can mark the first as well as the second variable (both variables are placed one after the other). OK*

*SPSS Output*
*(example: is there a difference between the happiness scores of husbands and their wives?).*
*(In a study of twenty families, the extent to which the husbands, wives and their (oldest) children felt themselves to be happy was investigated)*

**Ranks**

| | | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| HAPWOMAN - HAPMAN | Negative Ranks | 13[a] | 7,50 | 97,50 |
| | Positive Ranks | 1[b] | 7,50 | 7,50 |
| | Ties | 6[c] | | |
| | Total | 20 | | |

a. HAPWOMAN < HAPMAN

b. HAPWOMAN > HAPMAN

c. HAPWOMAN = HAPMAN

**Test Statistics[b]**

| | HAPWOMAN - HAPMAN |
|---|---|
| Z | -3,207[a] |
| Asymp. Sig. (2-tailed) | ,001 |

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

*Explanation:*

*The output consists of two tables: Ranks and Test Statistics. The Ranks table contains the number of 'negative ranks', the number of 'positive ranks' and the number of 'ties' ('ties': different respondents receive the same score). Of the 20 pairs, 13 involve a negative rank, which is to say a lower score for the women. A higher rank, in which the woman's score is higher than her husband's, only occurs once and in six cases, the happiness scores of the husband and wife are the same.*

*The test results are presented in the Test Statistics table, showing a Z-value of – 3.21 and a corresponding p value of 0.001. The conclusion is that we reject the null hypothesis: there is a significant difference (Z = - 3.21; p < 0.001) in the experience of happiness between husbands and their wives. It is clear that the men have a higher mean rank score on the happiness scale than their wives do.*

| | |
|---|---|
| Grouping variable:  NOMINAL | Test variable: ORDINAL |

Three or more paired (related) samples:  Friedman test

To be used when:
three or more related (paired) samples are involved and
the test variable is measured at least on the ordinal level

The Friedman test is more or less based on the same principles as the Wilcoxon Signed-ranks test (used when there are two paired samples). You use the Friedman test with three or more related groups or samples, for example, when you wish to compare the happiness score of fathers, mothers and their (oldest) children, or when you are dealing with three happiness scores from the same respondents, those collected in year 1, year 2 and year 3. This latter occurs in the context of a so-called panel study. First, for each related group, for example a family, the scores of fathers, mothers and children are ranked.  Subsequently, the mean rank score is calculated for each group separately (for all the fathers,  all the mothers and all the children). Based on the differences between these, a test statistic $T$ is calculated and converted into a Chi-square value.

( test statistic: $T = \left\{ \frac{12}{bt(t+1)} \sum_{i=1}^{t} R_i^{\,2} \right\} - 3b(t+1)$ with:

$t$ = number of groups

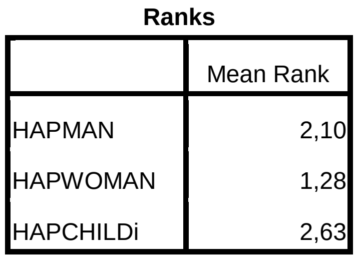
$R_i$ = total of the ranks for group i

$b$ = number of cases

If there are no differences between the groups, then the distribution of $T$ is approximately chi-squared distributed with $(t-1)$ degrees of freedom)

*SPSS:*
*ANALYZE ;  NONPARAMETRIC TESTS ;  K RELATED SAMPLES ;  mark the test variables in "Tests for Several Related Samples" ;  OK*

*SPSS  OUTPUT*
*(example: is there a difference in the happiness of men, their wives and their children?)*



**Ranks**

| | Mean Rank |
|---|---|
| HAPMAN | 2,10 |
| HAPWOMAN | 1,28 |
| HAPCHILDi | 2,63 |

**Test Statistics**[a]

| | |
|---|---:|
| N | 20 |
| Chi-Square | 23,903 |
| df | 2 |
| Asymp. Sig. | ,000 |

a. Friedman Test

*Explanation:*
*The Ranks table contains the mean rank scores for the happiness of the different groups. You can see that the women have the lowest scores (1.28). Their husbands score higher (2.10), and their children even higher (2.63).*
*The test results are presented in the Test Statistics table. The Chi-square value is 23.90, the number of degrees of freedom is 2 (number of groups minus 1, so 3 – 1 = 2). Additionally, twenty families are shown to be involved. The p value is 0.000. So, we reject the null hypothesis stating that there are no differences between fathers, mothers and their children. The differences in the mean rank scores appear to be significant. The mothers have the lowest score, the men are clearly happier and the children even more so.*

*(In your research report, you could report this result as follows:*
*"A comparison of 20 families has shown that wives are clearly less happy (mean rank score 1.28) than their husbands (2.10). The oldest child in each family appeared to be the happiest (mean rank 2.63). The differences are significant (Friedman Chi$^2$ = 23.90; df = 2; $p < 0.001$).")*

| | |
|---|---|
| Grouping variable: NOMINAL | Test variable: INTERVAL/RATIO |

Two independent samples: (Student's) t test (Independent-Samples T Test)

(including ***Levene's test*** to compare population variances)

To be used when:
you want to test, on the basis of two independent samples, whether the means of two populations are equal.

*SPSS:*
*The T test is a so-called parametric test which means that the test variable must be measured on the interval/ratio level, and that the corresponding sampling distribution is supposed to be normally distributed. The grouping variable is mostly measured on the nominal level.*
*The T test is based on the assumption of two random samples and, as stated above, a normally distributed sample distribution.*
*The sample distribution is normally distributed when:*
*- the variable in the population is normally distributed or*
*- when both samples have a sample size of 30 or more*
*The null hypothesis holds that both population means are equal (so, there is no difference between the two groups):*

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

*ANALYZE ; COMPARE MEANS ; INDEPENDENT-SAMPLES T TEST ;*
*indicate the variables you wish to include in the analysis*
*using "Define Groups" , you can indicate the two groups involved: Group1, type 1. Group 2, type 2.*
*CONTINUE ; OK*

*SPSS OUTPUT*
*(example: is the mean age of the men and that of the women comparable with or different from each other?)*

**Group Statistics**

| | Sex | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Age | man | 498 | 42,0723 | 10,06872 | ,45119 |
| | woman | 490 | 37,8102 | 6,96079 | ,31446 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| Age | Equal variances assumed | 127,612 | ,000 | 7,728 | 986 | ,000 | 4,26209 | ,55154 | 3,17977 | 5,34440 |
| | Equal variances not assumed | | | 7,750 | 884,889 | ,000 | 4,26209 | ,54996 | 3,18271 | 5,34146 |

*Explanation:*
*Table 'Group Statistics' contains descriptive statistical measures for the two sample groups.*
*Table 'Independent Samples Test' , left half, refers to 'Levene's Test for Equality of Variances': the outcome of this Test (which gives an answer to the question: do the standard deviations or variances of both populations differ significantly from each other or not?) determines whether you should use the numbers from the row 'Equal variances assumed' or from the row 'Equal variances not assumed'. The null hypothesis of Levene's Test holds: the variances in both populations are equal.*
*In the 'Independent Samples Test' table we see a high F-value (the test statistic F indicates if the variances differ from each other ) that is also significant, so we reject the null hypothesis of Levene's test and use the numbers from the row 'Equal variances not assumed'.*
*We find a t-value of 7,75 and a two-tailed p-value of 0,000 (for 884,889 degrees of freedom), so we reject the null hypothesis that the population means do not differ significantly from each other. From the '95% Confidence Interval of the Difference', we find that in 95% of all cases the difference between the mean age of men and women is between 3,18271 and 5,34146. So, a zero difference does not fall in this interval, which corresponds to the conclusion on the basis of the T Test (reject the null hypothesis).*

| Grouping variable: NOMINAL | Test variable: INTERVAL/RATIO |
|---|---|

Three or more independent samples: one-way analysis of variance (One-way ANOVA)

To be used when:
you want to find out if the mean scores on a test variable measured at the interval/ratio level for three or more independent samples differ from each other. The grouping variable is usually measured at the nominal level.

Other conditions:
-the test variable is normally distributed; this requirement is less important when the sample is larger, but you should nevertheless always check the distribution
-the $k$ independent samples each contain a minimum of 25 respondents
-the grouping variable is of a nominal or a ordinal nature and does not involve too many categories. If the grouping variable is measured at the interval/ratio level, you might consider calculating the relationship between the grouping variable and the test variable.
-the standard deviations in the samples are about the same

(if you are dealing with small samples and/or skewed distributions, it might be more advisable to conduct a Friedman test).

Here we are dealing with the so-called ‘analysis of variance’. It concerns a statistical test about the population means of three or more independent groups. The null hypothesis states that the population means of all ( $k$ ) groups are equal.
The analysis of variance is based on the variation of the sample data. This variation is expressed in terms of the sum of the squared deviations of all scores from the mean. This sum is called Sum of Squares ( $SS$ ). By dividing this Sum of Squares by the number of degrees of freedom, we obtain the Mean Square ( $MS$ ).

Here we limit ourselves to an analysis of variance (ANOVA) involving one group variable (One-way ANOVA), also known as a factor.
Suppose the test variable concerns age, and the grouping variable concerns marital status (single, with partner, and with partner and children). The total variance for age is divided into the between-group variance and the within-group variance.
The probability that any difference between groups results from chance is smaller when:
-the variance between groups is larger (hence the differences in (mean) age between singles, people with partners and people with partners and children are larger);
-the variance within the groups is smaller (hence the age differences within the separate marital status groups are smaller)
-the size of the independent samples being compared is larger

The F value is calculated to determine if there is a significant difference in the means of the groups.

The statistic F represents the ration between the mean within-group variance and the mean between-group variance, expressed as a Sum of Squares.
The 'within-groups sum of squares' is the sum of the squared differences between the individual scores and the group's mean.
The 'between-groups sum of squares' is the sum of the squared distances from the scores to the general mean (the mean when you consider all groups together).
To determine the mean 'within sum of squares' ($MS_{within}$), the 'within sum of squares' is divided by the number of samples minus 1 (= degrees of freedom 'within').
To determine the mean 'between sum of squares' ($MS_{between}$), the 'between sum of squares' is divided by the total number of respondents minus the number of samples (= degrees of freedom 'between').

The formula for the F value is: $F = \frac{MS_{between}}{MS_{within}}$ .
So, when the variance between is larger than the variance within, F will be larger than 1. The differences in age are then mostly the consequence of the differences in the grouping variable (marital status in our example).

The significance of the F value is, of course, dependent on the chosen level of significance (alpha), but also on the values of the degrees of freedom.
The probability of significance is greater when the number of respondents is larger and the number of groups is smaller.

*SPSS:*
*ANALYZE ; COMPARE MEANS ; ONE-WAY ANOVA ; move the test variable to the 'Dependent List' , and the grouping variable to 'Factor frame' ; OK*

*(when you click the "Options" button before clicking the OK button, a menu is displayed containing the "Descriptive" command. In this case you will get not only information if there are differences between the groups and if these differences are significant , but also concerning the question between which groups the differences exist. In most cases you will need this information.*
*By clicking on 'Descriptive' you are given information about the mean and the standard deviation of each subgroup. By clicking on 'Homogeneity of variance test' you get the possibility of confirming that the condition requiring the variances of the groups to be the same has been satisfied. By doing this, the "Levene's Test for Equality of Variances" will be performed.*
*To find out if, and if so, which groups are different from each other, you can click the "Post Hoc" button in the "One-Way ANOVA" menu. In this window, you will see that first you have to check if the variances are the same. If such is the case, you can then make use of the Bonferroni procedure, a simple and transparent technique to investigate post hoc (i.e., after the variance analysis has been executed) which precise groups have significant differences from each other.*
*Summarizing:*
*OPTIONS ; DESCRIPTIVE : HOMOGENEITY OF VARIANCE TEST ; CONTINUE ; POST HOC ; BONFERRONI ; CONTINUE ; OK)*

*SPSS OUTPUT*
*(Example: are there significant differences in the (mean) ages of singles (1), people with partners (2), and people with partners and children (3)?*

**Descriptives**

Age

|  |  |  |  |  | 95% Confidence Interval for Mean |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | N | Mean | Std. Deviation | Std. Error | Lower Bound | Upper Bound | Minimum | Maximum |
| alone | 150 | 28,0200 | 2,94803 | ,24071 | 27,5444 | 28,4956 | 25,00 | 34,00 |
| with partner | 365 | 35,3041 | 4,59492 | ,24051 | 34,8311 | 35,7771 | 28,00 | 42,00 |
| with partner and children | 469 | 47,4286 | 5,44099 | ,25124 | 46,9349 | 47,9223 | 35,00 | 55,00 |
| Total | 984 | 39,9726 | 8,92447 | ,28450 | 39,4143 | 40,5309 | 25,00 | 55,00 |

*The 'Descriptives' table shows that the mean age of singles is the lowest and that of people with partners and children is the highest. Furthermore, you can see that the standard deviation of age within the singles group is much smaller than in the other groups.*

**Test of Homogeneity of Variances**

Age

| Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|
| 37,678 | 2 | 981 | ,000 |

*Levene's test also demonstrates that the difference in variance amongst the various groups amongst the various groups cannot very likely be attributed to chance (p < 0.001).*
*The conditions for the variance analysis are not, in fact, satisfied. In this example this is not a real problem as it involves large samples.*

**ANOVA**

Age

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 55457,218 | 2 | 27728,609 | 1191,229 | ,000 |
| Within Groups | 22835,041 | 981 | 23,277 | | |
| Total | 78292,259 | 983 | | | |

*The 'ANOVA' table indicates that the variance between groups (Mean Square Between : 27728.61) is much larger than the variance within the samples (Mean Square Within : 23.28).*
*The relationship between the 'between variance' and 'within variance' , the F value, is large : 1191.23.*
*The difference proves to be very significant (p < 0.001).*

**Multiple Comparisons**

Age

Bonferroni

| (I) Marital/family status | (J) Marital/family status | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| alone | with partner | -7,28411* | ,46793 | ,000 | -8,4062 | -6,1620 |
| | with partner and children | -19,40857* | ,45256 | ,000 | -20,4939 | -18,3233 |
| with partner | alone | 7,28411* | ,46793 | ,000 | 6,1620 | 8,4062 |
| | with partner and children | -12,12446* | ,33676 | ,000 | -12,9320 | -11,3169 |
| with partner and children | alone | 19,40857* | ,45256 | ,000 | 18,3233 | 20,4939 |
| | with partner | 12,12446* | ,33676 | ,000 | 11,3169 | 12,9320 |

*. The mean difference is significant at the 0.05 level.

*The Bonferroni test indicates that all three groups differ significantly from each other. The probability that the observed differences result from chance is smaller than 0.001.*
*In fact, we should use a test that does not require us to assume the variances to be equal: the Tamhane's T2 test.*

**Multiple Comparisons**

Age

Tamhane

| (I) Marital/family status | (J) Marital/family status | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| alone | with partner | -7,28411* | ,34027 | ,000 | -8,0998 | -6,4684 |
| | with partner and children | -19,40857* | ,34794 | ,000 | -20,2423 | -18,5748 |
| with partner | alone | 7,28411* | ,34027 | ,000 | 6,4684 | 8,0998 |
| | with partner and children | -12,12446* | ,34780 | ,000 | -12,9566 | -11,2923 |
| with partner and children | alone | 19,40857* | ,34794 | ,000 | 18,5748 | 20,2423 |
| | with partner | 12,12446* | ,34780 | ,000 | 11,2923 | 12,9566 |

*. The mean difference is significant at the 0.05 level.

*Applying Tamhane's T2 Test however, gives the same result: the probability that the observed differences are a result of chance is for all the comparisons (group 1 with group 2; group 1 with group 3 ; group 2 with group 3) smaller than 0.001.*

*Once it has been established that a significant difference exists, it may be interesting to know the extent to which the difference in age can be explained by marital status. For this purpose, the squared éta (* $\eta^2 = \frac{SS_{Between}}{SS_{Total}}$ *) is often calculated, a statistic that expresses the relationship between the differences among the groups and the differences among the totaled scores. It is a measure of the extent to which the variance in the test variable (at the interval level) can be explained by a discrete or nominal grouping variable. (SS = Sum of Squares or squared deviation scores). In our example, we get 55457 / 78292 = 0.71, which means that 71% of the differences in age can be explained in terms of the difference in marital status.*

*To sum up:*
*Based on a one-way analysis of variance, we have observed that significant differences in age exist between singles (n = 150), people with partners (n = 365), and people with partners and children (n = 469); (* $F$ = 1191.23 ; $p$ < 0.001).

*On average, singles are the youngest (mean age 29.02 ; sd 2.95). People with partners and children are, on average, the oldest (mean age 47.43; sd 5.44). For people only living with a partner the mean age is 35.30 and the sd is 4.59.*
*A post-hoc comparison with the Tamhane's method shows significant differences in all pair-based comparisons (* $p$ < 0.001).

*Seventy-one percent of the differences in age can be explained by marital status (* $\eta^2 = 0.71$ ).

| | |
|---|---|
| Grouping variable: NOMINAL | Test variable:  INTERVAL/RATIO |

Two paired (related) samples: t test for two paired samples  (Paired-Samples T Test)

To be used when:

-you wish to compare the means of two dependent or paired samples;
-the samples are taken randomly
-the test variable (in our example educational level) is normally distributed; this requirement is less important when the sample is larger
-the sampling distribution (of the mean difference) is normally distributed (parametric test). This is the case when the population is normally distributed or when the samples are large enough ($n \geq 30$). With a smaller sample, it is more reasonable to use the Wilcoxon Signed-ranks test;
-the grouping variable is measured at the nominal level; if the grouping variable is measured at the ordinal or interval level, it might be better to calculate the correlation between the grouping variable (e.g. age) and the test variable (e.g. income) than testing for a difference between two groups on a given variable;
-the test variable (in our example educational level) is measured at the interval/ratio level.

(when you wish to compare three or more related groups, you have to use a multivariate variance analysis, which we will not discuss in this file).

As stated before, in case of two dependent or paired groups, each unit (or case) within the first group is somehow related to a unit in the second group. Some examples follow:
-consider a random sample of adult women (group 1) and a second group of their mothers, with the goal of determining differences in occupational careers;
-consider a random sample of respondents interviewed at two moments in time, for example during elections held in 2003 and in 2006;
-a third example is the comparison of two variables, such as the results on a statistics test (group 1) and a math test (group 2) as part of the Honours programme, while both groups contain the same respondents;

A typical characteristic of these examples is that it is likely that the occupational careers of mothers and their daughters, the answers of the respondents in 2003 and 2006, and the test results on statistics and math of the same students, are more similar than any randomly chosen pair from the sample of mothers and the sample of daughters etc. Thus, the unit of analysis is not a single unit but a pair of units with two scores that are to be compared.

In a $t$-test for two paired samples, first the differences between the two scores are calculated for each pair or case. In this way, a new variable is created: the paired differences. For this new variable, the mean ($\bar{d}$), the standard deviation ($s_d$) and the standard error are calculated. The statistical test concerns the question whether the mean of the paired differences $\bar{d}$ significantly deviates from 0 (no difference between the paired sets of scores), so the null hypothesis states that the mean difference in the population, $\mu_d$, is equal to 0.

The formula for the $t$ value follows: $t = \frac{\bar{d}}{s_d / \sqrt{n}}$, so in our example (see below):

$$t = \frac{-0.44334}{2.33176 / \sqrt{1006}} \approx -6.030$$

The question whether this $t$ value is significant or not, is answered on the basis of the number of degrees of freedom, which equals the number of pairs minus 1 ($d_f = n - 1$).

*SPSS:*
*ANALYZE ; COMPARE MEANS ; PAIRED-SAMPLES T TEST ; move the first variable to 'Variable 1 ' and the second variable to 'Variable 2' in 'Paired variables' ; OK*

*SPSS OUTPUT:*
*(our example deals with inequality between men and women: is the educational level of women lower than the educational level of their spouses? – Educational level has been measured with total years of education to obtain an interval variable).*

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | opleidvrouw | 12,1163 | 1006 | 2,31572 | ,07301 |
| | opleidman | 12,5596 | 1006 | 2,37858 | ,07499 |

*The first table contains, among other things, the mean of the test variable for both groups, as well as the standard deviation. You will note that the mean educational score for the men (12.5596) is higher than that of the women (12.1163).*

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | opleidvrouw & opleidman | 1006 | ,507 | ,000 |

*The correlation coefficient (0.507) refers to the relationship between the level of education of the woman and the man in any pair.*

**Paired Samples Test**

| | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | Lower | Upper | | | |
| Pair 1 opleidvrouw - opleidman | -,44334 | 2,33176 | ,07352 | -,58760 | -,29908 | -6,030 | 1005 | ,000 |

*In this table the results of the* $t$ *test are to be found: it shows that the observed difference of – 0.44334 ( = 12.1163 – 12.5596) is significant because the p value is (much) smaller than 0.05. So, the mean educational level of men and women differs significantly. The 95% Confidence Interval indicates the values between which the mean of the differences will be found. Note that the value 0 is not found in the interval. This corresponds with the result of the t-test:*
$H_0 : \mu_d = 0$ *is rejected.*

## III Correlation Research Questions (measures of association; relationship between variables)

Chi-square test
Cramér’s V
Spearman’s rank correlation coefficient (rho)
Pearson’s product-moment correlation coefficient

## Chi-Square test

To be used when:

you want to investigate if a significant relationship exists between two nominal variables.
By applying the Chi-Square test, you measure if a significant relationship exists between two nominal variables. However, the value of the Chi-Square does not say anything about the strength of the relationship (see : Cramér's V).

To use the Chi-Square Test, some conditions have to be satisfied:
- any given expected cell frequency may not be smaller than 1
- a minimum of 80% of the expected cell counts must have values greater than 5
- the variables must not have too many categories (otherwise the table becomes incomprehensible and the first two conditions are also likely not satisfied).

*SPSS:*
*Contingency Table:*
*Contingency tables are commonly used to describe the association or relationship between variables with low numbers of categories (preferably smaller than 10). Due to limitations imposed by the number of categories, contingency tables are generally used for nominal and ordinal variables only.*

*Each cell of the contingency table compares the number of observed cases (O) with the number that might be expected on the basis of chance (the expected cell counts: E). When the observed cell counts (O) deviate from the expected cell counts (E), the interesting question is: is this deviation great enough to be significant and not merely dependent on chance? By calculating the test statistic Chi-Square ( $\chi^2$ ), according to the formula:*

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

*we find a specific value for $\chi^2$ : the greater the $\chi^2$ , the smaller the probability that the deviation results from chance. SPSS does not only give the $\chi^2$ , but also the corresponding p-value (which is not only dependent on the size of the $\chi^2$ , but also on the number of degrees of freedom:*
$d_f = (c-1)(r-1)$ *with c = number of columns and r = number of rows).*
*The null hypothesis under the Chi-Square Test states that both variables are statistically independent, so there is no relationship.*

*ANALYZE ; DESCRIPTIVE STATISTICS ; CROSSTABS ; (define the grouping variable in 'Colum(s)' and the test variable in 'Row(s)') ; STATISTICS ; CHI-SQUARE ; CONTINUE ; OK*

*SPSS OUTPUT*
*example: is there a relationship between religious affiliation (nominal) and political party preferences (nominal) in the Netherlands?*

*[For finding the observed and the expected counts,*
*ANALYZE ; DESCRIPTIVE STATISTICS ; CROSSTABS ; Cells, choose Observed, Expected, and Column ; CONTINUE ; OK , see the first table below ]*

**politiekepartijvoorkeur * kerkgenootschap Crosstabulation**

| | | | kerkgenootschap | | | Total |
|---|---|---|---|---|---|---|
| | | | Katholiek | Protestant | onkerkelijk | |
| politiekepartijvoorkeur | CDA / CU /SCP | Count | 79 | 126 | 39 | 244 |
| | | Expected Count | 49,3 | 47,9 | 146,8 | 244,0 |
| | | % within kerkgenootschap | 38,7% | 63,6% | 6,4% | 24,2% |
| | PvdA / SP / GroenLinks | Count | 84 | 47 | 409 | 540 |
| | | Expected Count | 109,2 | 106,0 | 324,9 | 540,0 |
| | | % within kerkgenootschap | 41,2% | 23,7% | 67,4% | 53,5% |
| | VVD | Count | 35 | 21 | 113 | 169 |
| | | Expected Count | 34,2 | 33,2 | 101,7 | 169,0 |
| | | % within kerkgenootschap | 17,2% | 10,6% | 18,6% | 16,7% |
| | D66 | Count | 6 | 4 | 46 | 56 |
| | | Expected Count | 11,3 | 11,0 | 33,7 | 56,0 |
| | | % within kerkgenootschap | 2,9% | 2,0% | 7,6% | 5,6% |
| Total | | Count | 204 | 198 | 607 | 1009 |
| | | Expected Count | 204,0 | 198,0 | 607,0 | 1009,0 |
| | | % within kerkgenootschap | 100,0% | 100,0% | 100,0% | 100,0% |

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| politiekepartijvoorkeur * kerkgenootschap | 1009 | 73,4% | 366 | 26,6% | 1375 | 100,0% |

**politiekepartijvoorkeur * kerkgenootschap Crosstabulation**

Count

| | | kerkgenootschap | | | Total |
|---|---|---|---|---|---|
| | | Katholiek | Protestant | onkerkelijk | |
| politiekepartijvoorkeur | CDA / CU /SCP | 79 | 126 | 39 | 244 |
| | PvdA / SP / GroenLinks | 84 | 47 | 409 | 540 |
| | VVD | 35 | 21 | 113 | 169 |
| | D66 | 6 | 4 | 46 | 56 |
| Total | | 204 | 198 | 607 | 1009 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 302,047[a] | 6 | ,000 |
| Likelihood Ratio | 302,116 | 6 | ,000 |
| Linear-by-Linear Association | 48,808 | 1 | ,000 |
| N of Valid Cases | 1009 | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,99.

*Explanation:*
*The null hypothesis which states that there is no relationship between religious affiliation and political party preferences.*
*The Chi-Square ( $\chi^2$ ) value is very high, 302.047, and the probability that the observed association (relationship) between religious affiliation and political party preferences at six degrees of freedom results from chance, under the condition that the null hypothesis is true, is smaller than 0.001. (The Likelihood Ratio is comparable to the Chi-Square value).*
*To put it otherwise: if the null hypothesis (there is no relationship between religious affiliation and political party preferences , so there two variables are statistically independent) is true, then the probability of obtaining a Chi-Square that is as extreme or more extreme than 302.047682 is smaller than 0.001 (p-value). Assuming a level of significance ( α ) of 0.05, we reject the null hypothesis because the p-value is (far more) smaller than α.*

*Now that a significant relationship between religious affiliation and political party preferences has been shown, the next question to be asked would involve the extent to which political party preferences could be explained by religious affiliation. To answer this question, Chi-Square must be transformed into the measure of association known as Cramér's V , which has a value between 0 (no association exists) and 1 (there is a perfect association). (See below for more information ).*

*The point is, that the Chi-square value is influenced by the sample size and the number of degrees of freedom from the contingency table. That is why the Chi-square gives us information about the (non) existence of a relationship between two variables, but not about the strength of the relationship.*

*In order to determine the strength (and the direction) of the relationship between two variables, we need to use a measure of association. Which measure of association is to be used, depends on the level of measurement of the variables .*

*Measures of association for nominal variables only give information about the strength of a relationship. Measures of association for ordinal variables give information about the strength as well as the direction of the relationship. Various measures of association for nominal and ordinal variables are to be found in SPSS (ANALYZE ; DESCRIPTIVE STATISTICS ; CROSSTABS ; STATISTICS).*

*Some of these measures of association are based on the Chi-square, some are based on a so-called proportional error reduction. We will confine ourselves to the first category. A measure of association which is based on the Chi-square is corrected for the sample size and for the number of degrees of freedom from the corresponding contingency table. An example of such a measure of association is Cramér's V.*

*Below, we will apply Cramér's V as measure of association to answer the question how strong the relationship is between religious affiliation and political party preferences.*

## Correlation between two nominal variables: Cramér's V

To be used when:
you wish to know the strength of a relationship between two nominal variables.

First, you should use Chi-square to test if there is a significant deviation from a random distribution (see above). The requirements for calculating Cramér's V are the same as those for the use of Chi-square, namely:
- any given expected cell frequency may not be smaller than 1
- a minimum of 80% of the expected cell counts must have values greater than 5
- the variables must not have too many categories (otherwise the table becomes incomprehensible and the first two conditions are also likely not satisfied).

The findings on the basis of Chi-square are concerned with differences. Now we wish to investigate the strength of any relationship that may exist. In our example, once you know the religious affiliation of a person, to what extent can you predict his or her political party preference?

Cramér's V was developed by the Swedish statistician Harald Cramér (1893-1985). He calculated the maximum possible value for chi-square, given a certain sample size and given a certain number of rows/columns. He then divided the observed value for chi-square by this maximum value and took the square root. Without this square root, a difference between the observed and expected numbers that was twice as large would actually indicate a relationship that was four times stronger. This is due to the squaring of the differences between the observed and expected numbers when chi-square is calculated.
The formula for Cramér's V follows:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

in which:
$N$ = the number of cases
$k$ = the smallest number of colums or rows

(in our example, see below: $V = \sqrt{\frac{302.047}{1009(3-1)}} \approx 0.387$ )

The merit of Cramér's V is that its values are always between 0 and 1. A value of 0 indicates no relationship (the observed numbers are then identical to the expected numbers, so chi-square = 0). The value 1, on the other hand, indicates a perfect relationship. In common research applications, a value of 0.6 is considered exceptionally high. The following indicators for the strength of a relationship are proposed:

>0 – 0.10 = very weak

0.10 – 0.25 = weak
0.25 – 0.35 = moderate
0.35 – 0.45 = strong
> 0.45 = very strong

*SPSS*
*ANALYZE ; DESCRIPTIVE STATISTICS ; CROSSTABS ; (define the grouping variable in 'Colum(s)' and the test variable in 'Row(s)') ; STATISTICS ; CHI-SQUARE ; Phi and Cramér's V; CONTINUE ; OK*

*SPSS OUTPUT*
*(example: what is the strength of the relationship between religious affiliation and political party preferences?)*

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| politiekepartijvoorkeur * kerkgenootschap | 1009 | 73,4% | 366 | 26,6% | 1375 | 100,0% |

**politiekepartijvoorkeur * kerkgenootschap Crosstabulation**

Count

| | | kerkgenootschap | | | Total |
|---|---|---|---|---|---|
| | | Katholiek | Protestant | onkerkelijk | |
| politiekepartijvoorkeur | CDA / CU /SCP | 79 | 126 | 39 | 244 |
| | PvdA / SP / GroenLinks | 84 | 47 | 409 | 540 |
| | VVD | 35 | 21 | 113 | 169 |
| | D66 | 6 | 4 | 46 | 56 |
| Total | | 204 | 198 | 607 | 1009 |

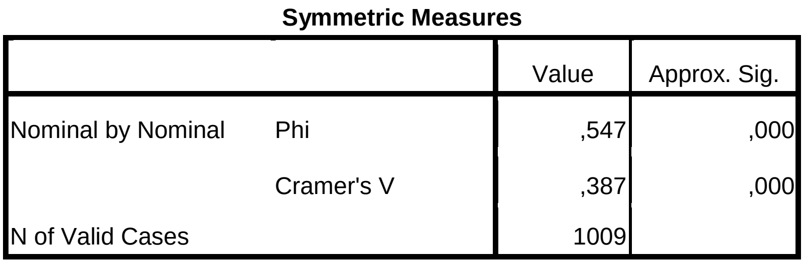
**politiekepartijvoorkeur * kerkgenootschap Crosstabulation**

| | | | kerkgenootschap | | | Total |
|---|---|---|---|---|---|---|
| | | | Katholiek | Protestant | onkerkelijk | |
| politiekepartijvoorkeur | CDA / CU /SCP | Count | 79 | 126 | 39 | 244 |
| | | % within kerkgenootschap | 38,7% | 63,6% | 6,4% | 24,2% |
| | PvdA / SP / GroenLinks | Count | 84 | 47 | 409 | 540 |
| | | % within kerkgenootschap | 41,2% | 23,7% | 67,4% | 53,5% |
| | VVD | Count | 35 | 21 | 113 | 169 |
| | | % within kerkgenootschap | 17,2% | 10,6% | 18,6% | 16,7% |
| | D66 | Count | 6 | 4 | 46 | 56 |
| | | % within kerkgenootschap | 2,9% | 2,0% | 7,6% | 5,6% |
| Total | | Count | 204 | 198 | 607 | 1009 |
| | | % within kerkgenootschap | 100,0% | 100,0% | 100,0% | 100,0% |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 302,047[a] | 6 | ,000 |
| Likelihood Ratio | 302,116 | 6 | ,000 |
| Linear-by-Linear Association | 48,808 | 1 | ,000 |
| N of Valid Cases | 1009 | | |

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,99.



**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | ,547 | ,000 |
| | Cramer's V | ,387 | ,000 |
| N of Valid Cases | | 1009 | |

*Explanation :*
*The third table shows that non-members have a strong preference for left-wing parties. Dutch Catholics traditionally have difficulties in choosing between left- wing (an economical interest) and Christian parties (a cultural interest), while Protestants predominantly prefer Christian parties. The relatively large differences regarding political party preferences are reflected in a strong relationship between religious affiliation and political party preferences (Cramér's V = 0.39, p-value < 0.001), and differs significantly from 0 with all common values of* $\alpha$ *.*

*To put this in other words:*
*the Chi-square test demonstrates that the observed distribution is significantly different from a random one. Specifically, Chi-square is 302.047 and, when df = 6 ((columns – 1)(rows – 1) = (3 – 1)(4 – 1)) the p-value is < 0.001.*
*The strength of the association between religious affiliation and political party preferences is large: Cramér's V = 0.39.*

*You might report this in the following way:*
*A strong association (Cramér's V = 0.39) can be seen to exist between religious affiliation and political party preferences. Non-members have a strong preference for left-wing parties, compared to members of a religious community, though there is a difference between Catholics and Protestants in this respect. This correlation is significant (Chi-square = 302.047 ; df = 6 ; p< 0.001).*

Correlation between ordinal variables: Spearman's rank correlation coefficient (rho)( $r_s$ )

To be used when:

you wish to investigate if a relationship exists between the rank scores of two variables
-that are measured at the ordinal level
-or at the interval/ratio level, when the sample is small ( $n < 30$) and/or contains many extreme scores
-or at the interval/ratio level, when the relationship is nonlinear

In the previous sections we used Chi-square to determine the *presence* of a relationship between variables of which at least one was nominal, and Cramér's V to determine the *strength* of a relationship between the same kind of variables.
If both variables are ordinal, not only can the strength of the relationship between the two be determined, but so can its *direction*. For example, it is obvious to expect a positive relationship between educational level and income: the higher the level of education attained, the more income earned. Likewise, studies show a negative relationship between health care and child mortality: the more a government invests in health care, the lower child mortality will be. In both cases, Cramér's V is not appropriate, as it fails to detect the direction or sign of the relationship.

Because ordinal variables are rank ordered, the relationship between them can also be expressed as the difference in rank order, as argued by psychologist Charles Spearman (1863 – 1945). Suppose that the variables educational level and income are perfectly related. In that case the ranking of education perfectly matches the ranking of income, and Spearman's rank correlation (rho or $r_s$ ) equals 1. When there is no relationship between educational level and income, neither is there a relationship between the rank order of education and income ( $r_s$ = 0). Finally, when a perfectly negative relationship exists between educational level and income, the rank order of both variables perfectly oppose each other ( $r_s$ = -1).

Spearman's rank correlation (rho or $r_s$ ) is calculated using the rank scores of two ordinal variables. To prevent the correlation from falling outside of the – 1 and + 1 range, rank scores are first standardized into $z$ - scores, eliminating the influence of variables measured in different units. For example, the variables educational level and income are difficult to compare because the ranking is measured in different units (levels vs. income classes). An easy solution for this incomparability is to transform them into $z$ - scores.
Next, for each unit of analysis (often respondents) – in SPSS called 'case' – the two $z$ - scores are multiplied and summed across all units to a total. This total sum of multiplied $z$ - scores reaches a positive maximum if both rank orders match perfectly. Conversely, the total sum has a maximum negative value when both rank orders perfectly oppose each other.
However, more units results in a higher total sum. Therefore, the total sum is divided by the total number of units ( $n$ ), resulting in a value that always falls between -1 (maximum negative association) and +1 (maximum positive association), while 0 means no association at all.

Because the rank scores are first transformed into $z$ - scores ('standardization'), the standard deviation becomes the unit of measurement. For example, when $r_s$ = - 0.5, a rise of 1 standard deviation in the ranking of $x$ is associated with a decline of 0.5 standard deviations in the ranking of $y$.

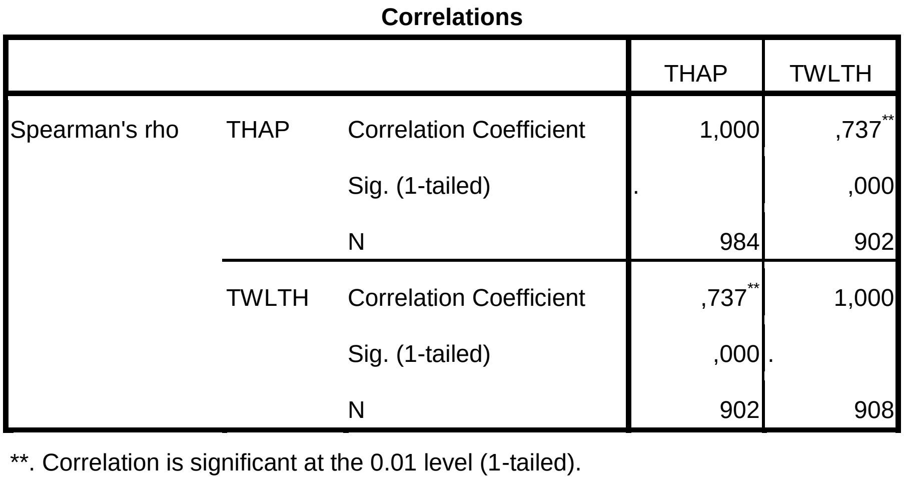
To test Spearman's rank correlations for samples with at least 30 observations, the $t$ - distribution can be used. Again, the $t$ - value indicates the relative distance between the observed rank correlation and the correlation stated in the null hypothesis. Finally, the associated $p$ - value is compared with the selected level of significance ($\alpha$ ).
The significance of the observed rank correlation is determined by the strength of the correlation and the number of sample elements (observations) used to calculate the correlation. For example, an $r_s$ of 0.35 for a one-tailed test at the 5% - level is significant when $n$ = 30, but not significant when $n$ = 10.
For samples with less than 30 observations, it is preferable to use an exact-test, where the $p$ - value is calculated based on the distributions of $x$ and $y$ variables, assuming that there is a zero rank correlation in the population (null hypothesis).

The benefit of Spearman's rank correlation ( compared to, for example, Kendall's tau, another measure of association for two ordinal variables which we will not discuss in this file), is that it is similar to Pearson's correlation coefficient, which is an important measure of association for interval-/ratio variables (see the next section). However, use of Pearson's correlation coefficient requires an approximately linear relationship, which is not required for Spearman's correlation. Therefore, $r_s$ provides a better alternative to describe a bivariate nonlinear relationship between interval variables. In case of a nonlinear relationship, Pearson's correlation coefficient mostly underestimates the real nonlinear association.

*SPSS:*
*ANALYZE; CORRELATE; BIVARIATE; mark the relevant variables and place them in the 'Variables Field'; SPEARMAN; choose one-tailed or two-tailed (box 'Test of Significance' )(if you already have a certain idea about the direction of the relationship, you should perform a one-tailed test); OK*

*(by clicking the 'Options' button, you can indicate what is to be done with the missing values: 'exclude cases pair-wise' or 'exclude cases list-wise')*
*(by default, the computer places two ** behind correlation coefficients that are significant with an alpha of 0.01 (99% reliability) and one * with an alpha of 0.05 (95% reliability).*
*You can also have the ** printed by placing a check in the box next to 'Flag significant correlations'. The exact $p$ - value is then also printed.*

*SPSS-OUTPUT*
*(example: is there a positive correlation between health and happiness?)*



**Correlations**

| | | | THAP | TWLTH |
|---|---|---|---|---|
| Spearman's rho | THAP | Correlation Coefficient | 1,000 | ,737** |
| | | Sig. (1-tailed) | . | ,000 |
| | | N | 984 | 902 |
| | TWLTH | Correlation Coefficient | ,737** | 1,000 |
| | | Sig. (1-tailed) | ,000 | . |
| | | N | 902 | 908 |

**. Correlation is significant at the 0.01 level (1-tailed).

*Explanation:*
*The Spearman's rank correlation between health (TWLTH) and happiness (THAP) proves to be 0.74 and positive. The* $p$ *value for this correlation coefficient is 0.000. The observed relationship between wealth and happiness is significant at an alpha of 0.01 (and even at an alpha of 0.001, as the p value is, after all, smaller than 0.001).*

*You might state this as follows:*
*A strong, positive relationship has been found to exist between the financial resources that respondents have at their disposal and the extent to which they are content about the lives they are leading (* $r_s$ *= 0.74 p < 0.001, one-tailed). The hypothesis that wealthier people are also happier has therefore been confirmed.*

Correlation between interval and ratio variables:

Pearson's product-moment correlation coefficient

To be used when:

you wish to investigate if there exists a relationship between two interval or ratio variables.

Pearson's (Karl Pearson 1857 – 1936) correlation coefficient, or more fully, Pearson's product-moment correlation coefficient, is a statistic (notation: $r$) which indicates the extent to which a linear relationship exists between two variables $x$ and $y$ which are measured at the interval or ratio (SPSS: 'scale') variables.
Pearson's correlation coefficient equals the maximum value of 1 when an increase of 1 unit of variable $x$ is associated with an increase of 1 unit of variable $y$. This relationship is called a perfect positive *linear association*. (In a scatter plot, all points then lie along an upward-sloping straight line). The linear association is perfectly negative ($r = -1$) if every 1-unit increase of $x$ results in a 1-unit *decrease* of $y$.

Because variables often have different units of measurements (for example, the variables *body weight* and *age* are measured in kilograms and years, respectively), the original scores of both variables are first transformed into $z$ - scores.
Per unit of analysis (SPSS: 'case') these $z$ - scores on $x$ and $y$ are multiplied and finally summed across all units. This total sum is positive when the linear association is also positive, and vice versa. Because the total sum tends to be higher when the total number of observed units (often respondents) is higher, the total sum is divided by the total number of observations, resulting in a correlation that falls within the range -1 and 1. The corresponding formula shows:

$$r = \frac{\sum_{i}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)}{n-1}$$

The correlation coefficient always lies between these two extremes and equals 0 when there is no linear association. This, however, may not mean that there is no association between the variables, as nonlinear association may exist!

As the original scores are first transformed into $z$ - scores with the standard deviation as their unit of measurement, Pearson's correlation coefficient indicates that when the score on one variable ($x$) increases by 1 standard deviation, the score on the associated variable ($y$) will increase by a number of standard deviations equal to $r$. In our example (see below), a positive relationship between height and weight amounts to 0.52. Therefore, for every standard deviation increase of height, weight increases on average by 0.52 standard deviations. An interpretation of the value of $r$, so of the strength of the relationship, is to be based on the following table:

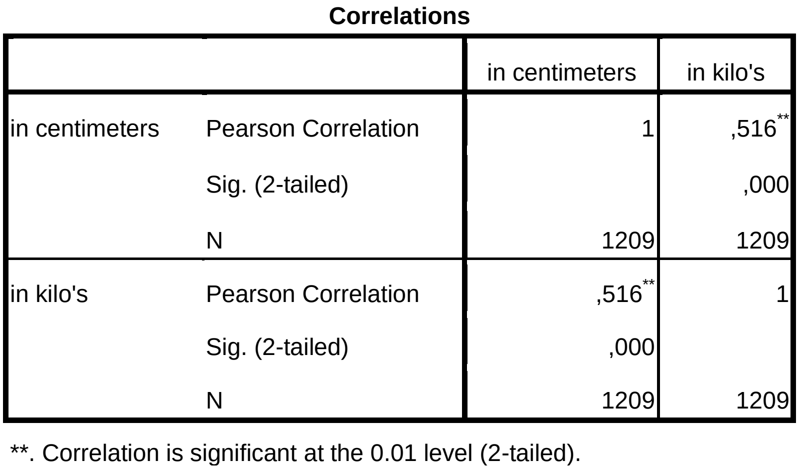
>0 – 0.10 = very weak

0.10 – 0.25 = weak
0.25 – 0.35 = moderate
0.35 – 0.45 = strong
> 0.45 = very strong

As with Spearman’s rank correlation, Pearson’s correlation is statistically tested using a $t$-distribution when the sample has at least 30 observations. In smaller samples, the exact-test provides a more appropriate alternative. The significance of the observed correlation coefficient is dependent on the strength of the correlation and the sample size $n$. A correlation of 0.30 is, for example, not significant at all ( $p > 0.10$) for a sample of 10, significant at the 5% level for a sample of 50 ( $p < 0.05$), and significant at the 1% level for a sample of 100 ( $p < 0.01$).

*SPSS:*
*ANALYZE ; CORRELATE ; BIVARIATE ; place the variables in the ‘Variables field’ ; click on Pearson ; OK*

*SPSS OUTPUT*
*(example: what is the strength of the relationship between height and weight?)*



**Correlations**

| | | in centimeters | in kilo's |
|---|---|---|---|
| in centimeters | Pearson Correlation | 1 | ,516** |
| | Sig. (2-tailed) | | ,000 |
| | N | 1209 | 1209 |
| in kilo's | Pearson Correlation | ,516** | 1 |
| | Sig. (2-tailed) | ,000 | |
| | N | 1209 | 1209 |

**. Correlation is significant at the 0.01 level (2-tailed).

*The correlation, calculated for 1209 respondents, is 0.52 and the associated $p$ value is smaller than 0.001, so it has been demonstrated that a relatively strong positive relationship exists between height and weight.*

*Note:*
*You might be interested in answering the following question: on average, how many kilograms are added to body weight when body height increases by a particular value? The answer can be found by using linear regression>*
*SPSS:*
*ANALYZE ; REGRESSION ; LINEAR ; fill in the dependent and the independent variable ; OK*

*SPSS OUTPUT*
*(example: what is the linear regression line for the relationship between height and weight?)*

*As a detailed discussion of the SPSS output is beyond the scope of this summary, we will limit ourselves to only one SPSS table:*

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -50,540 | 6,069 | | -8,327 | ,000 |
| | in centimeters | ,729 | ,035 | ,516 | 20,920 | ,000 |

a. Dependent Variable: in kilo's

*The equation of the linear regression line appears to be:*

$$y = 0.729x - 50.54$$

*By filling in a particular value for the height ( $x$ ),*
*you can predict the associated value for the weight ( $y$ ).*

# IV One-Sample T test

The One-Sample T test is used to test, on the basis of a sample mean, whether the corresponding population mean equals a given or supposed value.
Or, to put it otherwise:
To be used when
you want to test the null hypothesis that the population mean equals a specific value, on the basis of a sample mean.

*SPSS:*
*The test variable must be measured on the interval/ratio level (parametric test).*

*The T test is based on the assumption of a random sample and a normally distributed sample distribution.*
*The sample distribution is normally distributed when:*
*- the variable in the population is normally distributed or*
*- when both samples have a sample size of 30 or more*
*The null hypothesis holds that the population mean equals a given value:*

$$H_0 : \mu = \mu_0$$

*ANALYZE ; COMPARE MEANS ; One-SAMPLE T TEST ;*
*choose the variable you wish to test (Test Variable(s))*
*indicate the Test Value (in our example: 4, so* $H_0 : \mu = 4$ *)*
*OK*
*(see 'Options' for other confidence intervasl than 95%)*

*SPSS OUTPUT*
*(example: the customers of a specific shopping centre will, on average, visit this centre four times a month)*

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| gem aantal winkelbezoeken/maand | 80 | 4,56 | 2,530 | ,283 |

**One-Sample Test**

| | Test Value = 4 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| gem aantal winkelbezoeken/maand | 1,989 | 79 | ,050 | ,563 | ,00 | 1,13 |

*Explanation:*
*Table 'One-Sample Statistics' contains descriptive statistical measures for the test variable: number of cases (sample size), sample mean, sample standard deviation, standard error.*

*Table 'One-Sample Test' gives the t-value, the number of degrees of freedom (N-1), the significance of the t-value (two-tailed), the difference between the sample mean and the supposed population mean (4 in our example) and the Confidence Interval of the mean difference: in 95% of all cases will the difference between the sample mean and the supposed population mean lie between 0,00 (actually 0,0005330) and 1,13.*

*The t-value is to be calculated with the help of the following formula:*

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{4{,}56-4}{2{,}530/\sqrt{80}}\approx 1{,}989$$ (rounding offs)

*In case of a two-tailed test and on the basis of a 95% confidence interval (* $\alpha = 0.05$ *), the null hypothesis is just rejected (the significance equals* $\alpha = 0.05$*), so the hypothesis that the mean number of shopping visits equals 4 is (just) rejected.*
*In case of a one-tailed test (in that case the null hypothesis refers to a population mean of at least 4, so 4 or less), the significance should be divided by two (0,050 : 2 = 0,025) and the null hypothesis will be (firmly) rejected: the average number of shopping visits is significantly greater than 4 (0.025 is much smaller than 0.05) .*

## Literature

Anderson, D.R., D.J. Sweeney, Th. A. Williams, J.Freeman, E. Shoesmith (2007). *Statistics for Business and Economics*. London: Thomson

Cohen, L.J. (1989). *An Introduction to the Philosophy of Induction and Probability*. Oxford: Clarendon Press.

Baarda, D.B., M.P.M. de Goede, C.J. van Dijkum (2004). *Introduction to Statistics with SPSS*. Groningen/Houten: Wolters-Noordhoff

Dooremalen, H., H. de Regt, M. Schouten (2007). *Exploring Humans. An Introduction to the Philosophy of the Social Sciences*. Amsterdam: Boom

Gillies, D. (2000). *Philosophical Theories of Probability*. London (etc.): Routledge

Grotenhuis te, M. , Th. Van der Weegen (2009). *Statistical Tools. An Overview of Common Applications in Social Sciences*. Assen: van Gorcum

Howson, C. and P. Urbach (2006). *Scientific Reasoning. The Bayesian Approach*. Chicago and La Salle, Illinois: Open Court

Klep, P.M.M. and I.H. Stamhuis (eds) (2002). *The statistical mind in a pre-statistical era: The Netherlands 1750 – 1850*. Amsterdam: Aksant

O'Hagan, A. and B.R. Luce (2003). *A primer on Bayesian Statistics*. Centre for Bayesian Statistics in Health Economics, Sheffield U.K., from www.shef.ac.uk/chebs

Mellor, D.H. (2005). *Probability: A Philosophical Introduction*. London (etc.): Routledge

SPSS On-line Training Workshop: http://www.cst.cmich.edu/users/lee1c/spss/index.htm

Tijms, H. (2007). *Understanding Probability. Chance Rules in Everyday Life*. Cambridge: Cambridge University Press

Upton, G. and I. Cook (2008). *Oxford Dictionary of Statistics*. Oxford: Oxford University Press.

Vocht, A. de (2009). *Basishandboek SPSS 17. SPSS Statistics*. Utrecht: Bijleveld Press