Statistical Methods – JoHo Custom Summary

Table of Contents

Chapter 1 – What are statistics?

As a behavioral scientist, it is important to understand statistics. Research is namely conducted using empirical techniques, of which statistics is an essential part. When you understand which technique should be applied in which situation, you can use statistics correctly. Statistics comprises the arithmetic procedures to organize, sum up and interpret information. By means of statistics you can note information in a compact manner. The aim of statistics is twofold: 1) organizing and summing up of information, in order to publish research results and 2) answering research questions, which are formed by the researcher beforehand. Many students struggle with statistics. Hence, this summary explains the most important and frequently occurring topics. Statistics is not something to panic about, but without studying it may be quite hard. Hopefully, this summary will enlighten statistics for you.

Basic terminology

Often, research is conducted to examine the association between variables. A *variable* is a characteristic or condition that is changeable, or has different values for different individuals, for example age. These are *person variables*. But, variables can also apply to characteristics of the surroundings, for example temperature. Here, they are called *environmental variables*. Variables are noted by means of letters, for example variable X and variable Y. There are different kinds of variables. An *independent variable* is a variable that is being manipulated by the researcher. It often comprises two or more conditions, to which participants are being exposed. The *dependent variable* is the variable that is being observed after manipulating the observed variable. It shows what the effect is of the different conditions of the independent variable. Often, a *control group* is used in an experiment. This group receives no treatment, or a placebo to see is there is a difference between the experimental condition and the control group. Variables can also be subdivided into *discrete* and *continuous variables*. A discrete variable comprises different categories. For example, a class can consist of 18 or 19 children, but can not consist of 18.5 children. For a continuous variable, there are infinite numbers or values possible between two observed values. Think for example of length and weight.

Many variables that are being examined, are hypothetical constructs. Think for example of self-confidence. These constructs can not be measured directly. To measure these constructs, definitions of these constructs have to be formed that can be examined. For example, intelligence can be examined by using an IQ test. An *operational definition* describes how this construct should be examined. For example, hunger can be described as ‘the state in which someone is after not eating for at least 12 hours’. This is an example of an operational definition.

Research designs

Researchers can use four different research designs to test hypotheses:

1. *Descriptive research*: with descriptive research, the behavior, thoughts and feelings of a group of individuals are described. Developmental psychologists for example try to describe the behavior of children of different ages.
2. *Correlational research*: with correlational research, the association between variables is studied. With correlational studies, no statements can be made about cause-and-effect relationships.
3. *Experimental research*: in experimental studies, a variable (the independent variable) is manipulated to examine its possible effects on behavior (the dependent variable). If this is true (and all other assumptions are met), we can conclude that the independent variable causes these changes. The main feature of an experiment is the manipulation of the independent variable.
4. *Quasi-experimental research*: this type of design is used when researchers are, for whatever reason, not able to manipulate the variable. Think for example of gender and age. The researcher studies the effects of a variable of an event that happens naturally and can not be manipulated. Quasi-experiments do provide less certainty than real experiments.

The research process

The research process comprises seven steps:

1. Select a topic.
2. Demarcate and specify the topic. Study prior research with regard to your topic and specify the research question(s).
3. Set up a plan to answer the research question, and examine which research design is most appropriate for this.
4. Collect data to find an answer to your question.
5. Analyse the data. Look for patterns in your data.
6. Interpret the data; give meaning to your data.
7. Publish the results of your research, and inform others about the results.

The above steps are rarely clearly separated from each other: conducting research is an interactive process in which many steps are intermingled with each other. In addition, sometimes you have to go back to a prior step of the process.

Basic symbols

Below, you see a table with some useful and frequently used symbols:

Sample	Term	Population
$\acute{X}$	mean	μ
S^2	variance	σ^2
S	standard deviation	σ
p_s	proportion	p_O
r_{xy}	correlation	ρ_{xy}
$\hat{Y}$	dependent variable	Y
$\hat{\beta}_x$	beta	β_x

Types of statistics

There are different types of statistics. *Descriptive statistics* is used to describe the data. We can calculate the mean, display the data in a graph or look for extreme scores. *Inferential statistics* refers to making inferences about the population, based on a certain sample. By means of inferential statistics, we try to answer answer this. When a measurement refers to the whole population, it is called a *parameter*. When a measure refers to the sample, it is called a *statistic*. Statistics are thus estimates of the parameter.

Measurement levels

Variables can be subdivided into four different measurement levels, which are summarized below from lowest to highest level of measurement:

1. *Nominal*: the simplest (lowest) measurement level is the nominal scale. For nominal variables, numbers only refer to categories. Measurement on the nominal scale categorize and label observations. The number 1 for example can be used for ‘men’ and the number 2 can be used for ‘women’. One can not calculate something with these numbers, because they are only labels.
2. *Ordinal*: an ordinal variable comprises a set of categories with an ordering. For example, you can order the participants of a singing competition of worst to best on the basis of the applause they received. However, we can not determine perfectly how much more applause one or the other singer received.
3. *Interval*: here, we do speak of ‘real’ number. Equal differences between number on this scale reflect equal differences in strength. However, with interval variables, there is no defined zero point. For example, you can not say a person has zero height. Because there is no zero level, we can not multiply or divide the numbers of an interval scaled variable.

4. *Ratio*: here, we do speak of a zero-level. Because of this, we are able to add, subtract, multiply and divide observations. Examples of ratio scaled variables are weight and reaction time.

Chapter 2 – How can data be collected?

Collecting data

Data collection can be subdivided into three groups:

1. *Observational measurements*: in this case, behavior is observed directly. This can be done in every study in which the behavior that is to be examined, can be seen directly. Researchers can observe the behavior directly, or they can make audio- or video-recordings, from which information about the participants can be deduced. In observational studies the dependent and independent variable(s) of interest are not manipulated. No claims can be made about cause-and-effect relationships between variables.
2. *Physical measurements*: these are used when the researcher is interested in the relation between behavior and not-directly observable physical processes. It refers here to processes of the human body, that often can not be observed by eye. For example, hart rate, sweating, brain activity and hormonal changes.
3. *Self reportage measurements*: participants answer questions on questionnaires or interviews themselves. There are three kinds of self reportages: 1) *cognitive*: these measure what people think 2) *affective*: these measure what people feel and 3) *behavioral*: these measure what people do.

Quantitative versus qualitative data

In statistics, a subdivision is made into quantitative and qualitative data. Quantitative data results from a certain measurement, for example the grade on a test, weight or the scores on a scale. A measurement instrument is used to determine how much a certain characteristic is present in an object.

Qualitative data is also called frequency or categorical data. Qualitative data refers to categorizing objects. For example, 15 people are categorized as ‘very anxious’, 33 people are categorized as ‘neutral’ and 12 people are categorized as ‘little anxious’. The data consists of frequencies for each category.

Associations

Most research is conducted to discover or examine associations between variables, for example to examine the relation between sleeping habits and school achievement. The first research technique to examine relations is the *correlational method*. With the correlation method the researcher observes two variables to discover if there is a relation between them. The *experimental method* is used when the researcher is interested in the cause-and-effect relation between variables. A change in one variable will cause a change in another variable. This method has two essential characteristics. First, there is a *manipulation*. This implies that the researcher changes the values of a variable (X). Next, values of a second variable (Y) are measured to see if changes of X influence the values of Y. Second, there is *control*. This means that the researcher has to keep the research situation constant. When all other variables/conditions are kept constant, the researcher can claim that changes in Y are caused by X and not by another variable. It is important to be aware of the distinction between correlation and causation. A *correlation* implies that there is a relation between variables, but this does not tell us anything about the direction of the effect. Hence, you can not say that changes in one variable are caused by the other variable. Three conditions have to be met in order to make statements about causality:

1. *Covariance*: variables should covary together. A high score on the x-variable should be in accordance with a high score on the y-variable.
2. *Direction*: the cause should precede the consequence.
3. *Exclusion of the influence of other variables*: it may be the case that a third variable (z) influences both x and y.

Interpreting and displaying raw data

Frequency distributions, proportions and intervals

When participants are being measured, the obtained data are called *raw data*. These data are difficult to interpret. Therefore, steps have to be taken in order to process these data. Raw data is only a collection of numbers. Structure can be added by, for example, displaying the data in a graph. When reaction times are measured, one can for example make a *frequency distribution*. In a frequency distribution, you note how often a certain value (here: reaction time) occurred. This helps you to visualize which value (here: reaction time) occurred most frequently. Describing proportions and percentages is also useful in a frequency distribution. A proportion is calculated by dividing the frequency that belongs to a certain X-value by the total amount of participants. For example, if two people, that belong to a class of 20 persons, scored a six (X=6), the proportion for the score six is 2/20 = 0.10. The formula is: proportion = $p = f/N$ (f refers to frequency and N refers to the total amount of participants or observations). Because proportions are always calculated in relation with the total amount of participants or observations (N), we call them *relative frequencies*. Percentages can be obtained by multiplying proportions by hundred. Thus: percentage = $p(100) = f/N(100)$. Sometimes, many different scores are possible. In that case, it is better to make *grouped frequency distributions*. Here, we make groups of scores instead of only looking at individual values. The groups (or intervals) are called *class-intervals*. Instead of noting for example each possible length, you make groups of different length-intervals. For example, a group with the interval of 100 to 120 cm and a group with the interval of 121 to 140 cm. You can note the group behind each frequency.

Graphs

A frequency distribution can be displayed well in a figure. This is called a *graph*. An example is a *histogram*. The horizontal axis is called the x-axis, and the vertical axis is called the y-axis. The categories are displayed on the horizontal axis, and the frequencies are displayed on the vertical axis. To make a histogram, bars have to be drawn. The height of each bar is in accordance with the frequency of the category. A *bar chart* is in principle similar to a histogram, except that the bars are not put directly next to each other. Also the values that differentiate strongly from the other values are displayed. These values are called *outliers* are often (but not always) not useful. Besides graphs, lines can also be applied to the obtained data. The most frequently used line is the normal curve. This line is highest in the middle of the distribution, and decreases symmetrically at both sides of the middle. The normal distribution is symmetric, but not every distribution looks like this. A bimodal distribution for example, has two peeks. If a distribution has only one peek, it is called a unimodal distribution. A distribution can also be asymmetric, because the distribution is longer on on of the sides. A distribution with a 'tail' to the left has a negative skewness, and a distribution with a tail to the right has a positive skewness.

Besides histograms and bar charts, one can also use *stem-and-leaf-plots*. In such plots, each score is subdivided into two parts. The first number (for example the 1 of 12) is called the stem, and the second number (for example the 2 in 12) is called the stem. When you draw a plot, first note all stems (the first number). Next, note each leaf of each score. A stem-and-leaf-plot offers you the opportunity to quickly find individual scores, which may be useful for calculations. This is not possible with a frequency distribution.

Percentiles

Individual scores are called raw scores. However, these scores do not provide much information. For example, if you tell someone you scored 43 points on your exam, it is not clear to the other person whether this is good or bad. To be able to interpret such a score, it should be clear what the mean is. The *rank* or *percentile rank* is a number that implies what percentage of all individuals in the distributions scored below a certain value. Such a score is also called a *percentile*. The *percentile rank* refers to the percentage, whilst the *percentile* refers to a score. To determine percentiles and percentile ranks, it first has to be examined how many individuals score below any value. The result is called *cumulative percentages*. These percentages show what percentage of individuals score below a certain X-value and add up to 100 for the highest possible value of X. An easy way to use percentile is by means of *quartiles*. The first quartile (Q1) is 25%, the second quartile (Q2) is 50% (thus, the mean) and the third quartile is 75%. The distance between the first and third quartile is called the *interquartile range* (IQR). 1.5 times the IQR above Q3 or below Q1 is a criterion to identify possible outliers. All these data can be displayed in a *boxplot*. The so-called 'box' is from the first to the third quartile. In addition, the median is displayed in the box by a horizontal line. In addition, there is a vertical line from the lowest to the highest observations, that also goes through the box. Outliers are displayed with an asterisk above or below the line.

Central tendency

Measurements of the central tendency are measurements that display where on the scale the distribution is centered. There are three ways to do so: the mode, the median and the mean. These manners differ in the amount of data they use.

1. *Mode*: is used least frequently and is often least useful. The mode is simply the most frequently occurring score. In case of two adjacent scores, the mean of these two numbers is taken.
2. *Median*: the score that corresponds to the point of which 50% of all scores falls below when the data are ordered numerically. Therefore, the median is also called the 50th percentile. Imagine that we have the scores 4, 6, 8, 9, and 16. Here, the median is 8. In case of an even number of scores, for example 4, 6, 8, 12, 15, and 16, the median falls between 8 and 12. In this case, we take the mean of the two middle scores as median. Thus, the median is 10 in this case. A useful formula to find the median, is that of the *median location*: (N+1)/2.
3. *Mean*: this measurement of the central tendency measurements is used most frequently, because all scores of the distributions are included. The mean is the sum of the scores divided by the total amount of scores. That is*: $(\sum X)/N$*. A disadvantage of the mean is that it is influenced by extreme scores. Therefore, the 'trimmed' mean is sometimes used. For example, ten scores at both ends of the distribution are excluded. As a result, the more extreme results are excluded and the estimation of the mean becomes more stable.

Measuring variability

The *variability* of a distribution refers to the extent to which scores are spread or clustered. Variability provides a quantitative value to the extent of difference between scores. A large value refers to high variability. The aim of measuring variability is twofold:

1. Describing the distance than can be expected between scores;
2. Measuring the representativeness of a scores for the whole distribution.

The *range* of a measurement is the distance between the highest and lowest score. The lowest score should be subtracted from the highest score. However, the range can provide a wrong image when there are extreme values present. Thus, the disadvantage of the range is that it does not account for all values, but only for the extreme values.

Variance and standard deviation

The *standard deviation* (SD) is the most frequently used and most important measure for spread. This measurement uses the mean of the distribution as comparison point. Moreover, the standard deviation uses the distance between individual scores and the mean of the data set. By using the standard deviation, you can check whether individual scores in general are far away or close to the mean. The standard deviation can be calculated by means of the next four steps:

1. First, the *deviation* of each individual score to the mean has to be calculated. The deviance is the difference between each individual score and the mean of the variable. The formula is: deviation score = X – μ. The X refers to the individual score, and the μ refers to the mean of the variable.
2. In the next step, the mean of the deviation scores has to be computed. This can be obtained by adding all deviations scores and dividing the sum by the number of deviation scores (*N*). The deviation scores are combined always zero. Before computing the mean, each deviation score should be placed between brackets and squared.
3. Next, the mean of the squared sum is computed. This is called the *variance*. The formula of the variance is: $\sigma^2 = \sum (x-\mu)^2$.
4. Finally, draw the square root of the variance. The result is the standard deviation. The final formula for the standard deviation is thus: $\sigma = \sqrt{\sum (x-\mu)^2 / N}$.

Often, the variance is a large and unclear number, because it comprises a squared number. It is therefore useful and easier to understand to compute and present the standard deviation.

In a sample with *n* scores, the first *n-1* scores can vary, but the last score is definite. The sample consists of *n-1 degrees of freedom* (in short: df).

Systematic variance and error variance

The total variance can be subdivided into 1) systematic variance and 2) error variance.

- *Systematic variance* refers to that part of the total variance that can predictably be related to the variables that the researcher examines.
- *Error variance* emerges when the behavior of participants is influenced by variables that the researcher does not examine (did not include in his or her study) or by means of measurement error (errors made during the measurement). For example, if someone scores high on aggression, this may also be explained by his or her bad mood instead of the temperature. This form of variance can not be predicted in the study. The more error variance is present in a data set, the harder it is to determine if the manipulated variables (independent variables) actually are related to the behavior one wants to examine (the dependent variable). Therefore, researchers try to minimize the error variance in their study.

Chapter 3 – What do reliability and validity mean?

Reliability and validity

Reliability and validity are two central themes within statistics. The *reliability* refers to the phenomenon that the measurement instrument provides consistent results. If you repeat the same measurement twice, a reliable instrument will provide the same result. *Validity* describes whether the construct that is aimed to be measured, is indeed being measured by the instrument. The validity is dependent upon the aim of the study: an instrument may be valid for one concept, but not for another. A valid measurement is always a reliable measurement too, but the reverse does not hold: if an instrument provides consistent result, it is reliable, but does not have to be valid.

Measurement error

The score of a participant on a measurement consists of two parts: 1) the true score of the participant and 2) measurement error. In short: observed score = true score + measurement error. The true score is the score that a participant would have had if the measurement technique was perfect and hence no measurement errors have been made. However, the measurement techniques that researchers use are (almost) never flawless. All measurement techniques consist of measurement error. Because of these measurement errors, scientist can never reveal the exact score of a participant.

Measurement error and reliability

Measurement errors reduce the reliability of a measurement. When a measurement has a low reliability, the measurement errors are large and the researcher knows little about the true scores of the participants. When a measurement has a high reliability, little measurement error occurred. The observed scores of a participant are then a good (but not perfect) reflection of the true score of the participant.

Reliability as systematic variance

Scientist are never completely certain how much measurement error is persistent in a study and what the true scores of participants are. In addition, they do not know completely certain how reliable their measure is, but they can estimate how reliable it is. If they determine that their measure was not reliable enough, they can try to make their measure more reliable. If making their measurement more reliable is not possible, they can decide not to use the measure at all.

- The total variance in a data set of scores consists of two parts: 1) variance by true scores and 2) variance by measurement errors. In formula form, this is: total variance = variance by true scores + variance by measurement errors.
- We can also say that the proportion of total variances that is in accordance with the true scores of the participants is the *systematic variance*, because the true scores are systematically related to the measurement.
- The variance that is caused by measurement errors is called *error variance*, because this variance is not related to what the scientist examines.
- We therefore can say that the reliability can be computed by dividing the systematic variance by the total variance. Thus, reliability = systematic variance / total variance. The reliability of a measurement is somewhere between 0 and 1. A reliability of 0 implies that the scores solely exist of measurement errors and that there is no true score variance present in the data. The scores only refer to measurement errors. The reverse applies to a reliability of 1: now, only true score variance is present, and there is no variance caused by measurement errors. The rule-of-thumb is that a measure is reliable when the reliability is at least .70. This implies that 70% of the variance in the data refers to true score variance (systematic variance).

Types of reliability

Researchers use three types of reliability for analyzing their data: 1) test-retest reliability 2) inter-item reliability and 3) inter-rater reliability. A *correlation coefficient* is a statistic that indicates the strength of the relation between two measurements. This statistic lies between 0 (no relation between the measurements) and 1 (perfect relation between the measurements). Correlation coefficients can be positive or negative. When this statistic is squared, we see what proportion of the total variance of both measures is systematic. The higher the correlation, the more related the two variables are. Below, the three types of reliability are discussed.

Test-retest reliability

A test-retest reliability refers to the consistency in the responses of participants throughout time. Often, participants are measured twice with some time (a few weeks) between the measurement occasions. If we assume that a characteristic is stabile, the person should get the same score twice when the test is similar. If someone scores 110 on an IQ-test the first time, this person should score around 110 on the second measurement occasion. This is because IQ is a relatively stabile concept. However, both measurement occasions will not be completely similar, because measurement errors always occur. If the correlation between both tests is high (at least .70), a test (here: IQ-test) has a high reliability. We expect a high test-retest reliability for intelligence-, attitude- and personality tests. For less stable characteristics, such as hunger of fatigue, measuring test-retest reliability is useless.

Inter-item reliability

The inter-item reliability is important for measurements that consists of more than one item. *Inter-item reliability* refers to the extent of consistency between multiple items on a scale. Personality questionnaires for example often consist of multiple items that tell you something about the extraversion or confidence of participants. These items are summed up to a total score. When researchers sum up the answers of participants to receive a single score, they have to be certain that all items measure the same construct (for example extraversion). To check to what extent items are in accordance with each other, the *item-total correlation* can be computed for each combination of items. This is the correlation between an item and the rest of all items combined. Each item on the scale should correlate with the remaining items. An item-total correlation of .30 or higher per item is considered to be sufficient. Next to calculating whether each item is in accordance with the remaining items, it is also necessary to calculate the reliability of all items combined. In the past, the split-half reliability was calculated.

- For the *split-half reliability* all items are subdivided into two sets. Next, a total score is computed for each set. Then, the correlation between both sets is calculated. If the items in both sets measure the same construct, there should be a high correlation between the tests. The correlation (and hence split-half reliability) is considered high if it is .70 or higher. The disadvantage of the split-half reliability is that the correlation that is found depends on which items are placed in which set. If you subdivide the items a little differently, it may result in a different split-half reliability.

Because of this reason, we recently calculate more often the '*Chronbach's alpha coefficient'*. The Chronbach's alpha is used to calculate (by means of a simple formula) the mean of all possible split-half reliabilities. Researchers assume that the inter-item reliability is sufficient when Chronbach's alpha is .70 or higher. The Chronbach's alpha can be computed as follows:

$$\alpha = \frac{K}{K-1} 1 - \frac{\sum_k V(X_k)}{V(\sum_k X_k)}$$

That is:

$$\alpha = \frac{items}{items-1} 1 - \frac{\sum of\ variances\ of\ all\ items}{total\ variance\ of\ complete\ scale}$$

Inter-rater reliability

Inter-rater reliability is also called ‘*inter-judge*’ or ‘*inter-observer*’ reliability. It refers to the extent to which two or more observers observe and code the behavior of participants equally. When the observers make similar judgements (thus, a high inter-rater reliability), the correlation between their judgements should be .70 or higher.

Validity

Measurement techniques should not only be reliable, but also valid. Validity refers to the extent to which a measurement technique measures what it should measure. The question is thus whether we measure what we want to measure. It is important to note that reliability and validity are two different things. A measurement instrument can be reliable, whilst not being valid. A high reliability tells us that the instrument measures *something*, but does not tell us exactly *what* the instrument measures. To discover that, it is important to check the validity of the instrument. Validity is not a definite characteristic of a measurement technique or instrument. A measure can be valid for one aim, whilst not being valid for another aim.

A subdivision is made into *internal validity* and *external validity*. Internal validity refers to drawing right conclusions about the effects of the independent variable. Internal validity is warranted by experimental control. This causes namely that only the independent variable differs between the conditions. If participants in different conditions differ systematically on more than only the independent variable, we are facing *confounding*. External validity refers to the extent to which the research results can be generalized to other samples. Researchers distinguish three kinds of validity: 1) face validity 2) construct validity and 3) criterion-validity.

Face-validity

Face-validity refers to the extent to which a measure *seems* to measure what it should measure. A measure has face-validity when people think that is the case. This form of validity can thus not be computed statistically, but is more an assessment of people who assess the measure based on their feelings. The face-validity is determined by the researcher, the participants and/or field experts. If a measurement does not have face-validity, the participants think it is not important to really participate. If a personality test has nog face-validity, but participants have to fill in the questionnaire, then they do not see the added value of the test. This decreases their motivation to participate in the study. It is important to remember three things: 1) If a measurement has face-validity, it does not mean per se that the measure is valid too 2) If a measurement does not have face-validity, it does not mean per se that the measurement is not valid 3) Some researchers try to hide their aims. For example, they are afraid that participants will not answer sensitive questions correctly. Therefore, the researcher may decide to design a measurement instrument that has no face-validity, because the measure does not seem to measure what it should measure.

Construct validity

Often, researchers are interested in *hypothetical constructs*. These are constructs that can not be observed directly by empirical evidence. The question arises how to determine whether the measurement of a hypothetical construct (that can not be observed directly) is valid. Chronbach and Meehl say that the validity of the measurement of a hypothetical construct can be determined by comparing the measure with other measures. Scores on an instrument for self-confidence for example should correlate positively with measures for optimism, but negatively with measures for insecurity and fear. A measurement instrument has construct validity when 1) it correlates strongly with instruments with which it should correlate (*convergent validity*) and 2) it does not correlate (or correlates to a small extent) with instruments to which it should not correlate (*discriminant validity*).

Criterion validity

Criterion validity refers to the extent to which a measurement instrument makes sure that we are able to distinguish between participant on a certain *behavioral criterion*. For example, the question is whether different scores on a motivational test in secondary school tell us something about the school achievements at university. The behavioral criterion is here achievement in university. Criterion validity is used often in applied research settings. For example, educational settings or job applications. Researchers distinguish between two primary types of criterion validity: 1) concurrent and 2) predictive validity. The main difference between these two types is the amount of time between the measure and the determination of the behavioral criterion.

- *Criterion validity* occurs when two measures are used at almost the same time. The question is whether the measurement instrument distinguishes well between people who score high and people who score low on the behavioral criterion on that moment. If scores on an instrument are relate to behavior to which it should be related *on that specific moment*, the measurement instrument has concurrent validity.
- *Predictive validity* occurs when a measurement instrument is able to distinguish between people on a behavioral criterion in the future. Thus, it refers to whether the instrument can give a reliable prediction of some future behavior. This is mainly important in the educational setting.

Chapter 4 – Which distributions emerge in statistics?

Normal distribution

The normal distribution is a symmetric, bell-shaped distribution. The normal distribution is the most important distribution because of four reasons:

1. We expect that many of the dependent variables, with which we work, are normally distributed in the population.
2. If a variable is (approximately) normally distributed, we are able to make claims about the values of that variable (it is often a prerequisite to do analyses).
3. When an infinite number of samples is drawn from a population, the distribution of these samples tends towards a normal distribution.
4. Most statistical programs assume that the observations are normally distributed.

The normal distribution uses so-called z-scores. To discuss the normal distribution, we therefore first have to explain what z-scores are and how they can be used.

Standard scores

Often, individual scores are transferred to standard scores, also called z-scores. This is done to determine exactly the position of each score on a distribution. Z-scores are used to standardize the whole distribution. By using z-scores, we are able to compare different distributions.

The z-score describes the exact position of a X-value in two ways: 1) via the sign and 2) via the value. The plus- or minus-sign of the z-score describes whether the X-value is above of below the mean (the mean of a standard deviation is always zero). The value of the z-score describes the distance between the X-value and the mean in terms of number of standard deviations (a z-score of 1.00 means that the X-value is 1 standard deviation away from the mean). In a distribution with $\mu = 100$ and $\sigma = 15$, a score of $X = 130$ receives a z-score of +2. This is obtained by 130-100 = 30, and 30 divided by 15 is 2. The mean μ lies always in the middle of the curve. To the right side of the mean, the z-scores have a positive sign. Z-scores at the left side of the mean obtain a minus sign.

The formula for the calculation of the standard scores is: $z = (X-\mu)/\sigma$. The deviation score is deviated by the standard deviation. This way, the z-score describes how many standard deviations an individual score is away from the mean. An IQ-score of 70 is exactly two standard deviations below the mean: $(70-100)/15 = -2$. In this formula, $(X-\mu)$ refers to the *deviation score*. By subtracting the mean of a score, we see directly whether the score is higher than or lower than the mean. This formula is useful for transferring raw scores to z-scores, but the reverse does not hold. Therefore, one can rewrite the formula.

The z-score and a normal distribution

The standard normal distribution has a mean of 0 and a standard deviation of 1. The distribution is thus N(0,1). The normal distribution is symmetric; the highest frequency is in the middle, and the frequencies decrease to the left and the right of the distribution. Z-scores for normal distributions are given in terms of standard deviations. A z-score of +2 means that the scores is two standard deviations above the mean. For a normal distribution, the following rules apply:

- ± 68% of the observations falls between 1 standard deviation of the mean
- ± 95% of the observations falls between 2 standard deviations of the mean
- ± 99.7% of the observations falls between 3 standard deviations of the mean

Central limit theorem

How to determine whether a sample has a normal distribution? Think for example of the number of hours per day watching television. The main part of the people watches television between one and two hours per day. However, exceptions are present of people who watch eight hours per day. The distribution will then be skewed to the right. Although this chance distribution is not normally distributed, the sampling distribution of the sample mean will be normally distributed. This is called the *central limit theorem*. This happens only if the sample size *n* is large enough, that is at least *n* = 30.

Chances, proportions and scores

Imagine: a distribution of intelligence has a µ of 100 and a σ of 15. What is the chance that, by means of random sampling, an individual is selected with an IQ below 130? To answer this question, the IQ-scores (X-values) first have to be transferred to z-scores. Next, the corresponding proportion has to be determined. This is in accordance with the chance that has to be determined. Here, the z-score is +2. This score is computed as follows: (130-100)/15 = 2. According to table A, the corresponding proportion is 0.9772. Thus: P(X<130) = 0.9772. Thus, there is a 97.72% chance to randomly select someone with an IQ below 130. What to do when you want to examine the proportion between two values? For example, the mean driving speed is 58. The standard deviation is 10. How many of the cars that pass by will drive between 55 and 65 kilometers per hour? You are actually looking for p(55<X<65) here. First, calculate the z-score for both values. For X=55, the z-score is -0.30, because (55-58)/10 = -0.30. For X=65, the z-score is 0.70. Find the corresponding proportion for both values in the table for normal distributions (table A). The proportion of scores below 65 (z = 0.70) = .7580. The proportion of scores below 55 (z = -0.30) = .3821. Because the question refers to the proportion between these two values, the final answer is: 0.7580 – 0.3821 = 0.3759 = 37.59%.

The binomial distribution

When a variable is measured on a scale with exactly two categories, the resulting data is called *binomial*. Binomial data can also result from a variable that only has two categories. For example, people are either male or female. Only heads or tails can result from tossing a coin. In addition, it may happen that a researcher wants to simplify the data by subdividing it into two categories. For example, a psychologist may use scores on a personality test to classify aggression as high or low. Often, the researcher knows the chances on both categories. For tossing a coin for example, the chance on both head and tails is 50%. For a researcher, it is important to know how often an event occurs when there are multiple runs. For example, what is the chance that someone tosses head 15 times, when tossing the coin 20 times in total?

To answer questions about chance on the binomial distribution, you first have to exam the binomial distribution. The formula for the binomial distribution is: *p(*X) = CNX *p***q*(N-X) = $\frac{N}{X(N-X)} p*q(N-X)$

p(X) = the chance on X successes
N = the number of trials
P = the chance of success on 1 trial
Q = (1-*p*); the chance of failure
CNX = the number of combinations of *N* things that happen *X* times.

Mean and variance

When *p* = *q* = 0.50, for example when tossing a coin, the binomial distribution will be symmetric. The formulas for mean, variance and standard deviation are:

- Mean = *N**p
- Variance = *N**p*q
- Standard deviation = $\sqrt{N*p*q}$

For the binomial distribution, it applies that the distribution becomes more normal when *p* and *q* are close to 0.50. In addition, the distribution becomes more symmetric and more normal, when the number of trials increases. The rule-of-thumb is that, when *N*p* and *N*q* do not exceed 5, the distribution is close to normal. Then, estimations are reasonably well when we treat the distribution as normal.

Categorical data and Chi-square

When we are facing categorical data, the data exists of frequencies of observations that are subdivided into two or more categories. For these data, we can use the Chi-square test.

The Chi-square distribution

The formula of the Chi-square distribution differs from other functions, because it only has one parameter. The others are constants. The normal distribution has two parameters (μ and σ), the Chi-square only has k as parameter, which refers to the X² degrees of freedom (df). For example, three degrees of freedom are presented as $\chi^2 3$ or $\chi^2(3)$. The larger k becomes, the more symmetric the distribution will be. The mean and the variance increase, when k increases. Moreover:

- Mean = *k*
- Variance = *2k*

The Chi-square distribution uses the observed frequencies and the expected frequencies. The observed frequencies are the actual frequencies in the data. The expected frequencies are the frequencies that you would expect, if the null hypothesis is true. The formula for the Chi-square is: $\chi^2 = \sum \frac{(O-E)^2}{E}$. Thus, you compute for each category the $(O-E)^2/E$ and sum these up. The O refers to observed frequencies. The E refers to expected frequencies.

Table of the Chi-square distribution

Now that we have a value for X², we have to compare this value with the X² distribution to determine the chance that a value of X² is at least as extreme, given the null hypothesis. To do so, you can use the standard table distribution of X² (table F). The table uses the degrees of freedom. For a uni-dimensional table, it applies that: *df = (k-1)*: thus the number of categories minus one. If the obtained X² value is larger than the value in the table, the null hypothesis can be rejected. The problem is however, that the Chi-square distribution is continuous, while the possible values of Chi-square are discrete (especially for small sample sizes). Fitting a discrete distribution into a continuous distribution results in a bad fit.

Two classification examples

In the previous examples, we discussed one dimension (or classification variable). However, often multiple classification variables are present and one wants to examine whether these are independent. When the variables are not independent, they are to a certain extent *contingent* or *dependent* upon each other. In a *contingency table*, we can place the distribution of each variable against each other.

In a contingency table, you note the frequencies you would expected if the variables are independent (between brackets). The expected frequencies are obtained by multiplying the row totals by the column totals (the *marginal totals*) and dividing this by the total sample size. The formula is: $E_{ij} = R_i C_j / N$. E_{ij} is the expected frequency for the cell with row *i* and column *j*. R_i and C_j are the row and column totals.

The chance that an observation belongs to row 1 is the total of that row divided by the number of cells within that row. This also applies to the columns. The expected frequency, if all observations are independent, can be obtained by multiplying these two chances, and multiplying this result by *N*. The value of X² can again be calculated with the same formula. The degrees of freedom can be deduced from the contingency table by: *df* = (*R* – *1)*(C* – *1)* with *R* and *C* number of rows and columns in the table.

Prerequisite of the Pearson Chi-square

One of the main prerequisites to use the Chi-square, is a reasonable size of expected frequencies. Small expected frequencies may cause problems, because they cause a limited number of contingency tables and hence a limited number of values for the Chi-square distribution. The continuous X² can not describe this discrete distribution well.

In general, the rule is that all expected frequencies should be at least five. For smaller frequencies, it is advised to use *Fisher's Exact Test*, because this test is not based on the X² distribution. For 2x2 tables with expected frequencies of 1, the X² can be found with the following formula:

$$\chi^2 adj=(\chi^2 * N)/(N-1)$$

The Fisher's Exact Test is used when expected frequencies are larger than one.

Measuring agreement

With categorical data, it is important to measure to what extent observes agree in their judgements. Imagine that we want to measure the problems of 30 adolescents, with a subdivision into (1) no problems (2) problems at school (3) problems at home. We ask the two observers to examine this, so that we can compare their judgements. By means of a contingency table, we examine how often each observed assessed each score. Imagine that they agree 20 out of the 30 times (the diagonal cells), then there is an agreement of 66%. This is the percentage of agreement.

The problem with calculating a percentage, is that we have to take into account the possibility that the observes agree by chance. To correct for this, Cohen developed the statistic kappa (K). The formula for the kappa is:

$$\frac{\sum f_0 - \sum f_e}{N - \sum f_e}$$

in which f_0 is the observed frequency on the diagonal and f_e is the expected frequency on the diagonal. Assume the kappa is K = 0.33. This implies that –after correction for chance- the agreement between the two observers is 33%. This is much lower that the prior computed value of 66%.

Chapter 5 – How to construct a sample?

A *population* is the unity of events or participants in which a researcher is interested, for example all children of twelve years in a country. Populations can vary to a great extent in size. Because it is (often) not possible to measure the whole population, samples are used in a study. A *sample* is a number of participants or observations from the full population, which are being measured. A *random sample* is preferred. This means that all participants from the population have an equal chance of being selected for the sample. This results in a representative sample. A sample is representative if a certain characteristic occurs as frequently in the sample as in the population. Often however, the sample is not a perfect representation of the population. The difference between a sample and the corresponding population is caused by *sampling error*. A *parameter* refers to a value that describes the population. For example, the mean school achievement in a population of Dutch children. A *statistic* refers to a value that describes the sample. Often, a chance sample is used. Such a sample can be achieved in several ways.

1. Simple random sampling

With *simple random sampling*, a sample is chosen in such a way that each possible sample has an equal chance of being selected from the population. When a researcher for example wants to select a sample of 100 participants from a population of 5000 participants en each combination of 100 participants has an equal chance of being selected as sample, it is a *simple random sample*. To select such a sample, the researcher should use a *sampling frame*. That is a list for the whole population from which the sample will be drawn. Participants are selected randomly from this list. A disadvantage of the simple random sampling is that it requires to know beforehand how many participants there are in the population, and how many are required for the sampling frame. In some situations, forming a sampling frame is impossible. In such situations, a *systematic sampling* is chosen. Every ..th person is chosen to participate in the sample. For example, every 10th person that enters a building is selected to participate.

2. Stratified random sampling

Stratified random sapling is a variant of simple random sampling. Here, participants are not selected directly from the population, but are first subdivided into multiple strata. A *stratum* is a part of the population that is in accordance with a certain characteristic. For example, we can subdivide the population into men and women or into three age categories (20-29, 30-39 and 40-49). Next, participants are chosen randomly from each stratum. By means of this procedure, researchers can control that an equal number of participants is drawn from each stratum. Therefore, researchers often use a *proportional sampling method* in which individuals are selected from each stratum proportionally. That means that the percentage of participants (from a certain stratum) is in accordance with the proportion in which this stratum occurs in the population.

3. Cluster sampling

When it is difficult to receive information beforehand about how many and which participants are present in the population, the *cluster sampling* method is used frequently. Here, the researcher does not draw individuals from the population, but clusters of possible participants. These clusters are often based on naturally occurring clusters, such as regions within a country. Often, *multistage sampling* is used with cluster sampling. With multistage sampling, large clusters are determined first. Next, smaller clusters within these large clusters are determined. This continues until a sample emerges with randomly chosen participants from each cluster.

Sampling errors (bias)

It is difficult to make a fully representative sample. There are different ways in which a sample can not be representative. These are called sampling errors or *bias*, and may result in misleading research outcomes. Two types of bias exist: systematic and non-systematic. Non-systematic bias occurs always. These are the result of sampling variance. For example, psychology students from one year are not the same as psychology students from another year, which may result in a different mean of the measured variable. However, you assume that the higher the number of participants in your sample, the smaller the influence of non-systematic bias will be. The researcher can control the *systematic sampling error* or *systematic bias*. Systematic bias can arise by means of the following different causes:

- *Selection bias*: The way in which the participants are selected, causes a biased view. For example, EUR psychology students may have a higher IQ than the total population of students. Another example can be found in inter-questionnaires. People without internet are automatically excluded from such a study.
- *Non-response bias*: A biased view arises, because the people that are willing to participate in your study, are different from the people that do not participate. For example, an IQ test for psychology students is voluntarily. People who consider themselves to be clever, may me more tempted to participate in the IQ test than people who consider themselves to be not so clever.
- *Response bias:* A biased view arises, because the answers that are given are not in accordance with the truth. For example, students do not feel like participating in an IQ test, but the test is mandatory. As a result, these students randomly fill in some answers. This might also happen when people participate only to receive a reward for their participation.

Other samples

In some situations, it is not useful or not possible to select a chance sample. In those situations, a *nonprobability sample* is drawn. In that case, the researchers do not know to what extent their sample is representative for the population. Many psychological studies are conducted with samples that are not representative for the population. Nevertheless, these samples are very useful for certain studies. Nonprobability samples are appropriate for studies in which testing hypotheses is important, and in which the population is not described. The faith in validity increases when different samples (about the same topic) result in similar results. There exist three types of nonprobability samples:

- *Convenience sampling*: A convenience sample is a sample in which researcher use participants that are directly available. A main advantage of a convenience sample is that by using this method it is much easier to recruit participants than it would be with representative samples.
- *Quota sampling*: With a quota sample, the researcher determines beforehand what percentages should be met. The sample is drawn based on these percentages. For example, a researcher might say that he wants to select exactly 20 men and 20 women for his study instead of randomly drawing 40 participants from the population without paying attention to gender.
- *Purposive sampling*: With a purposive sample, the researchers have strong ideas about which participants are typical for the population. Based on these ideas, they select which participants may participate in their study. The problem with purposive sampling is that it is highly subjective.

Drawing conclusions about the population

After collecting the results about the sample, the aim of the study is not achieved yet. The idea is that you make statements about the population based on these results. A couple of prerequisites have to be met before being able to draw conclusions about the population. These prerequisites are discussed below.

Reducing sampling errors (bias)

Sampling errors (bias) refers to deviations of your result from the true parameter. Imagine that the true parameter is 70, but you found a value of 69 in your study, then the sampling error is 1.

Sample size

A large sample is not a guarantee for a representative sample. The way in which the sample is drawn is at least as important as the sample size. However, there are guidelines that tell you how large your sample at least should be. In general, it is the case that the smaller the population, the larger the part has to be that is included in your sample. For example, if the population consists of 50 people, you need approximately 49 to obtain representative results. A rule-of-thumb is that, for small populations (<500), you select at least 50% for the sample. For large populations (>5000), you select 17-27%. If the population exceeds 250.000, the required sample size hardly increases (between 1060-1840 observations). In sum: the smaller the population, the larger the required sample ratio.

Confidence interval (CI)

As mentioned before, you can never be sure that your results are exactly in accordance with the true population parameter. To indicate this, you can calculate a *confidence interval*. That is a range of numbers below and above the estimate parameter, in which the true parameter will likely be. For example, if a 95% confidence interval runs from 30 to 33, you can say that you know with 95% confidence that the true population parameter is somewhere between 30 and 33. The sample size influences the confidence interval. The larger the sample size, the smaller the confidence interval. That implies that you are able to do a more precise estimation of the parameter based on a larger sample.

Chapter 6 – What is statistical inference?

Inferential statistics

There is a way to discover whether the difference in group means is caused by error variance or by systematic variance. Inferential statistics are used for this purpose. *Inferential statistics* refers to drawing conclusions. This method assumes that the independent variable has had an effect, when the difference between the means of the conditions is larger than is expected based on chance alone. Therefore, we compare the group means that we found with the group means that we expect to find if there is only error variance. Unfortunately, this method does not provide certainty. We are only able to determine the *chance* that the differences in group means are caused by error variance.

Testing hypotheses

Scientists try to test their hypotheses by analyzing different group means. First, they formulate a *null hypothesis*. This hypothesis states that the independent variable did not have an effect on the dependent variable. On the contrary, there is an *experimental hypothesis* which states that the independent variable did have an effect on the dependent variable. The experimental hypothesis may (*directional*) or may not (*non-directional*) indicate a direction of the effect. A directional experimental hypothesis is called *one-sided*. With a one-sided hypothesis, the researcher indicates whether he expects the independent variable to cause a decrease or increase of the dependent variable. When the researcher does not have an indication of the direction of the expected effect, a *two-sided* hypothesis can be used. With a two-sided hypothesis, the direction of the effect is not indicated. Based on statistical analyses, the null hypothesis can be rejected or preserved (*failing to reject the null hypothesis*).

Rejecting the null hypothesis implies that the independent variable did have an effect on the dependent variable. By rejecting the null hypothesis, you indicate that there is a difference between the means. The independent variable thus had an effect, and there is some systematic variance. Rejecting the null hypothesis implies that the difference in group means is larger than expected on error variance only. When the null hypothesis is preserved, it does not mean per se that the independent variable did not have an effect on the dependent variable. Instead, it means that the effect that is found is not large enough to reject the null hypothesis. Whether a null hypothesis is rejected, is dependent upon the confidence interval, which is controlled by the researcher (the researcher determines the significance level). It is important to understand that not rejecting the null hypothesis does not mean that there is no effect.

Type-I and Type-II errors

When drawing conclusions, four scenarios are possible:

- Correct decision: the null hypothesis is incorrect, and the researcher rejects the null hypothesis.
- Correct decision: the null hypothesis is correct, and the researcher does not reject the null hypothesis.
- *Type-I error*: the null hypothesis is correct, but the researcher rejects the null hypothesis. The researcher falsely assumes that the independent variable had an effect. The chance if making a type-I error is called *alfa level*. It is common practice to use an alfa level of 5%. This implies that the null hypothesis is rejected when there is a 5% chance that the found differences between the group means are caused by error variance. Thus, there is a 5% chance that the researcher wrongly rejects the null hypothesis. Sometimes, researchers use a stricter alfa level, for example an alfa of 1%. Then, they only have a 1% chance of wrongly rejecting the null hypothesis.
- *Type-II error:* the null hypothesis is wrong, but the researcher does not reject the null hypothesis. Thus, the researcher assumes that the independent variable did not cause an effect, while in fact it did cause an effect. The chance on a type-II error is called *beta*. Unreliably measuring the dependent variable raises the beta. Effects that do exists, are namely noticed less often with a larger beta. In addition, mistakes in collecting and

coding the responses, highly heterogeneous samples and bad experimental control may cause an increased beta. To reduce the chance on a type-II error, researcher try to design studies with a high power.

The z-test

Generally, we do not know the value of σ , and we have to estimate it with the sample standard deviation (*s*). However, if we do know the standard deviation of the population we can use the z-test.

Step 1: formulating a hypothesis

First, a hypothesis is formulated. There are two hypotheses: the null hypothesis and the alternative hypothesis. The *null hypothesis* means that the independent variable did not have an effect. The hypothesis implies that there is no change or difference. For the null hypothesis the symbol H_0 is used. The *H* refers to hypothesis, and the 0 refers to zero effect. Second, there is the *alternative hypothesis* (H1) which indicates that there is a change or difference. In the context of an experiment, it indicates that the independent variable (for example a treatment method for depression) did have an effect on the dependent variable (extent of depression). The H1 can be one- or two-sided. When the null hypothesis for example is that the mean depression score is 30 in the population of depressed people, the alternative hypothesis can be that the mean does not equal 30 ($\mu \neq 30$). In some cases, the direction of the difference is also specified. For example, if it is expected that the group that received treatment has a higher mean, it applies: H1: $\mu_1 < \mu_2$. It is for example possible to indicate with H1 that the mean is lower than 30 ($\mu < 30$) or higher than 30 ($\mu > 30$). The latter possibility is in this example superfluous, because it is almost impossible that a treatment will cause an increase in the degree of depression. Hypotheses always refer to the population, although samples are used to test hypotheses.

Step 2: criteria to make a decision

To make a founded decision about the (in)correctness of the null hypothesis, certain criteria have to be used. We use the *level of significance* or the *alfa level* (α) as criterion. The alfa level is a limit in the normal distribution that distinguishes between scores with a high chance and scores with a low chance of occurring in the sample, if the hypothesis is true. An alfa of 5% ($\alpha = 0.05$) implies that there is a 5% chance that the result is found by chance. The alfa level is a chance value that is being used to determine highly unlikely sample results, if the null hypothesis is true. The are that is demarcated by the significance level in the tale(s) of the distribution is the critical area. The *critical area* consists of extreme sample values that are highly unlikely is the null hypothesis is true. When a found value falls within this critical area, it differs significantly from the mean and the null hypothesis is rejected. For an alfa of 5%, it implies that 5% of the scores in the tails fall within this critical area; between $z = -1.96$ and $z = 1.96$. These values are the limits of the critical area for $\alpha = 0.05$.

Step 3: collecting data and calculations

Data are collected after the hypotheses are formulated. This way, the data can be tested by means of the hypothesis; the researcher can evaluate the data in an objective manner. After collecting the raw data, the statistics are calculated. The researcher calculates for example the sample mean. The, the mean can be compared to the null hypothesis. To do so, the researcher has to compute a z-score that describes the position of the sample mean compared to the mean of the null hypothesis. The z-score for the sample mean is: $z = (x - \mu) / \sigma$. The formula implies that the z-score can be calculated by subtracting the population mean from the sample value, and dividing this by the population standard deviation (or the squared root of the sample standard deviation divided by the sample size). The z-score for testing hypotheses is an example of a test statistic.

Step 4: making a decision

A researcher uses the z-score from the previous step to make a decision about the null hypothesis. The first possibility is that the researcher rejects the null hypothesis. This is the case when the statistics falls within the critical region. This means that there is a significant difference between the sample and the null hypothesis. The statistic is found in the tail of the distribution. Referring to the example of treating depression, it means that the researcher has found that the treatment had a significant effect. However, it is also possible that the null hypothesis can not be rejected. This is the case when the statistic does not fall within the critical area.

Effect size

Some researchers critize the process of testing hypothesis. The main critique refers to the interpretation of a significant result. When testing hypotheses, most attention is paid to the data instead of to the hypotheses. When the null hypotheses is rejected, we make statements about the sample data and not about the null hypothesis. Bases on the sample data, the null hypothesis is rejected or not. We do not know whether the null hypothesis is truly false or true. Another point of critique is that a significant effect does not imply anything about the effect size. Something is significant or not, but it does not imply anything about the size of the effect. Thus, a significant effect is not equal to a large effect. To provide more insight into the size of an effect, Cohen (1988) proposed the so-called *effect size*. His measure for effect size is called *Cohen's d*. This measure can be computed by first calculating the difference between the sample mean and the original population mean (M - µ). Next, this outcome is divided by the standard deviation of the population. The outcome of Cohen's d is classified as a small effect for d = 0.2, a medium effect for d = 0.5 and a large effect for d = 0.8.

The t-test

As mentioned before: In general, we do not know the value of σ and we have to estimate it with the sample standard deviation (*s*). However, if we replace σ by s, we can not use the z-formula, but we have to use the t-test. The t-test uses s^2 as approximation of σ ^{2}. The t-distribution uses *n-1* degrees of freedom. The larger the value of df for the sample, the better *s* (standard deviation of the sample) represents σ (standard deviation of the population). The t-statistic can be calculated with the following formula: t = (M - µ) / sM. sM refers to the standard error, which can be calculated as: sM = s/√n. This is used as estimation of the real standard error. Below you find a useful scheme to use when you test a hypothesis with the t-test:

	Right-sided	Left-sided	Two-sided
1. Formulate the null and alternative hypothesis	H0: μ ≤ 123 H1: μ > 123	H0: μ ≥ 126 H1: μ < 126	H0: μ = 122 H1: μ ≠ 122
2. Decision of the test statistic	$T=\frac{\acute{X}-\mu_0}{s/\sqrt{n}}$	$T=\frac{\acute{X}-\mu_0}{s/\sqrt{n}}$	$T=\frac{\acute{X}-\mu_0}{s/\sqrt{n}}$
3. Determine distribution of the test statistic	*T t(n-1)*	*T t(n-1)*	*T t(n-1)*
4. Intuitive area of rejecting H0	$X_{t \gg 0}^{\gg 123}$	$X_{t \ll 0}^{\ll 123}$	$X_{t \ll 0}^{\ll 122}$ $X_{t \gg 0}^{\gg 122}$
5. Determine level of significance	∝ = 0.05	∝ = 0.05	∝ = 0.05
6. Looking up the critical values	t99, 0.05 = 1.660	-t99, 0.05 = -1.660	t99, 0.05 = 1.660 -t99, 0.05 = -1.660
7. Compare the observed value test statistic with the critical value	$t=\frac{125-123}{10/\sqrt{100}}=2$ 2 > 1.660, so reject H0: the mean IQ on the EUR is not smaller than or equal to 123, for ∝ = 0.05	$t=\frac{126-123}{10/\sqrt{100}}=3$ 3 > -1.660, so do not reject H0 for ∝ = 0.05	$t=\frac{122-123}{10/\sqrt{100}}=-1$ so do not reject H0 for ∝ = 0.05.

Assumptions for the one-sample t-test

There are two assumptions that have to be met in order to conduct a t-test.

1. First, the scores in the sample have to be *independent* observations. That means that one score can not influence another score. The chance on a certain outcome for a score can thus not be influenced by another score.
2. Second, the population, from which the sample is drawn, has to be normally distributed. In practice, violation of this assumption has little influence on the t-statistic, especially when the sample size is high. With quite small samples, it is nevertheless important that the population is normally distributed. When you are insecure whether the distribution of the population is normal, it is best to use a high sample size.

Effect size of the t-test

The effect size can be computed with Cohen’s d. In that case, first the difference between the sample and population mean has to be determined. This has to be divided by the standard deviation of the population. Often, the standard deviation of the population is unknown. Hence, an *estimated d* is constructed by dividing the difference between sample and population mean by the standard deviation of the sample.

Proportion of explained variance (r^2)

A different way to determine the effect size, is by looking at how much variance between the scores is explained by the effect. An effect can namely cause an increase or decrease of the scores. The proportion of explained variance can be found by squaring the t-statistic and dividing it by the same number plus the degrees of freedom. In formula, this is: $r^2 = t^2 / (t^2 + df)$. The degrees of freedom are the number of scores minus one. A proportion explained variance of 0.01 refers to a small effect. A value of 0.09 refers to a medium effect. A proportion of 0.25 refers to a large effect. The r^2 is used presented in percentages in the literature.

The t-test for independent samples

The t-test is used often to test differences between two *independent samples*. For example, when we compare the achievements between a control group and an experimental group (which received a treatment). We want to examine whether the difference is large enough to assume that the two samples originate from different populations.

When we compare means of two different populations, we test the null hypothesis H0: µ1 - µ2 = 0. This comprises a sampling distribution of all possible difference scores between the population means. In case of two normally distributed populations, the distribution of the difference scores is also normally distributed. The variance of this distribution can be found with the *variance sum law*: the variance of the sum of the difference of two independent variables equals the sum of there differences:

σ^2X1-x2 = σ^2X1 + σ^2X2 = $\sigma^2$1/n1 + $\sigma^2$2/n2.

The formula for the t-statistic is:

$$T_s=\frac{\acute{Y}_1-\acute{Y}_2-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

in which $(\mu_1-\mu_2)$ equals zero and thus can be omitted.

Assumptions for the t-test with two independent samples

1. The observations in each sample are independent.
2. The populations from which the samples are drawn, are normally distributed. When the researcher assumes that the populations are not normally distributed, it is advised to use large samples.
3. The two populations have equal variances. We call this *homogeneity of variances*. Pooling sampling variance is only useful when both populations have the same variance. This assumption is very important, because a correct interpretation depends upon the research findings. You can check if this assumption is met with Levene's test in SPSS.

Pooled variance

The above formula can only be used when both samples have the same sample size (n1 = n2). In this case, the variance of both samples is precisely the middle of the two variances separately. When the samples do not have an equal sample size, this formula is not appropriate because the two samples receive the same weight in this formula, while a smaller sample should receive a smaller weight than a large sample. Consequently, the outcome is biased towards the small sample. To correct for this, a formula is used that combines the variances: *the pooled variance*. The pooled variance is found by taking the weighted mean of both variances. The sum of squares of both samples is divided by the degrees of freedom. The degrees of freedom are lower for a smaller sample, so that this smaller sample will receive a lower weight. As mentioned before, the variance of a sample ($s^{2)}$ can be obtained by dividing SS by df. To calculate the pooled variance (s^2_p), a different formula is used: (SS1 + SS2) / (df1 + df2). The estimated standard error of M1-M2 is found by taking the square root (√) of the outcome (s^2_p */ (n1 + n2)).* A different formula for the pooled variance is:
s^2_p = ((n1-1)s21 + (n2-1)s22 / (n1+n2-2). The new t-formula is then:

$$(\acute{X}_1-\acute{X}_2)/\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Effect size

As mentioned before, Cohen's d can be computed by dividing the difference between the means by the standard deviation of the population. For two independent samples, the difference between the two samples (M1 – M2) is used to estimate the difference in means. The pooled standard deviation ($\sqrt{s_p^2}$) is used to estimate the standard deviation of the population. The formula to estimate Cohen's d is thus: d = (M1 – M2) / $\sqrt{s_p^2}$.

Paired t-test

A *paired t-test* is used when there is a matched design or when there are repeated measures. The paired t-test takes into account that participants in two conditions are similar to each other. In this case, there are two different samples, but each individual from the one sample is matched to an individual from the other sample. Individuals are matched basis on variables that are considered to be important for the study. This causes an increase of the power: if the independent variable truly has an effect, it is more likely that this will be found in the study. The lower the error variance, the higher the power of the experiment. A high power results in a lower pooled standard deviation (sp). The lower the pooled standard deviation, the higher the t-value.

The t-statistic for related samples is, with regard to its structure, similar to the other t-statistics. The main difference is that the t-statistic for matched samples is based upon difference scores instead of raw scores (X-values). Because participants before and after the treatment are examined, each participant has a difference score. The difference score is:

D = X2 – X1

In this formula, the X2 refers to the second measurement (often: after the treatment). When D is a negative number, it implies that the extent of occurrence of the variable X has decreased after the treatment. A researcher tries to examine whether there is a difference between two conditions in the population by using difference scores. He wants to know what would happen when each individual in the population would be measured twice (before and after the treatment). The researcher strives to know what the mean of the difference score (µD) in the population is.

The null hypothesis is that the mean of the difference scores is zero (µD = 0). According to this hypothesis, it is possible that some individuals in the population have positive difference scores. In addition, it is possible that some individuals have negative difference scores. However, the main question is whether the mean of all difference scores equals zero. The alternative hypothesis H1 states that the mean of the difference scores does not equal zero (µD ≠ 0). The t-statistic for the difference scores is calculated as:

$$T_s = \frac{\acute{X}_1 - \acute{X}_2 - (\mu D)}{S_D / \sqrt{n}}$$

Assumptions for the paired-samples t-test

1. The scores within each condition are independent.
2. The difference scores (D) are normally distributed. Violation of this assumption is not a big matter, as long as the sample sizes are large. For a small sample, this assumption has to be met. A large sample size refers to a sample with at least 30 participants.

If one or more assumptions of the t-test for repeated measures are not met, an alternative test can be used. This is the *Wilcoxon-test*, in which rank scores are used for comparing difference scores.

Effect size

The two most frequently used measures for effect size are Cohen's d and r^2 (proportion of explained variance). Because Cohen's d assumes population parameters (*d = µD – σD)* it is useful to estimate d. The estimated d can be computed by dividing the mean of the difference scores by the standard deviation (d = MD/s). A value higher than 0.8 is considered a large effect. The proportion of explained variance van be computed as: $r^2 = t^2 / t^2 + df.$

Summary of formulas t-test

T-test	Formula	
T-test for equal variances (independent samples)	$T_p = \frac{\acute{Y}_1 - \acute{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$s_p^2 = \frac{(n1-1)s_1^2 + (n2-1)s_2^2}{n1+n2-2}$ $S_p = \sqrt{} S_p^2$
T-test for unequal variances (independent samples)	$T_s = \frac{\acute{Y}_1 - \acute{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	*df* are given in the exercise
Paired sample T-test (dependent samples)	$T_s = \frac{\acute{X}_1 - \acute{X}_2 - (\mu D)}{\frac{S_D}{\sqrt{n}}} t(n-1)$	$S_D^2 = S_1^2 - S_2^2 - 2 r_{1,2} S_1 S_2$ $S_D = \sqrt{} S_D^2$

Confidence intervals

Confidence intervals can assist in describing the results from hypothesis tests. When we obtain a specific estimation of a parameter, we call this a *point estimation*. Next, there are interval estimations, which obtain the limits within the true population parameter(μ) likely is. These are called the *confidence limits*, that make the *confidence interval*. We want to know how high and how low the μ-value can be, for which we do not reject H0. This provides the limits within we keep the null hypothesis.

- z-test confidence interval: $\acute{X} \pm Z\alpha/2 \frac{\sigma}{\sqrt{} n}$
- one sample t-test confidence interval: : $\acute{X} \pm t_{n-1,\alpha/2 \frac{s}{\sqrt{} n}}$
- t-test for independent samples with equal variances: $\acute{Y}_1 - \acute{Y}_2 \pm t_{n-2,a/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t-test for independent samples with unequal variances: $\acute{Y}_1 - \acute{Y}_2 \pm t_{df,a/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- t-test for paired samples: $\mu D = M_D \pm t * S_M D$

Power

Besides measuring the effect size, it is also possible to measure the power of a statistical test. Power refers to the extent to which a study is capable of detecting the effects in the examined variable. A study with a high power is able to detect existing effects, while a study with a low power will likely not detect these effect. The power is influenced by many things. One of these is the number of participants. In general, it applies that the more participants there are, the higher the power is. Strong effects are easier to identify than weak effects. A study with a low power will often not identify weak effects, but may identify the strong effects. To identify a weak effect, a high power is required. For identifying weak effects, it is also useful to have data from many participants. Power can be calculated as 1 – the chance on a type-II error. Researcher often require a power of 0.80. The power of the test is influenced by three factors:

1. First, the sample size is important. The larger the sample size, the higher the chance of rejecting the null hypothesis when the null hypothesis is actually false. This means that the power of the test increases when the sample size increases.
2. Second, the power of the test decreases when the alfa level is lowered. When the alfa for example is decreased from 5 to 1%, the chance that a true effect is found (thus that the null hypothesis is rejected correctly) decreases.
3. Third, the power increases when a two-sided test is transferred to a one-sided test.

Chapter 7 – What are correlation, regression and linear regression?

Correlation versus regression

Correlation and regression are two topics that are strongly related to each other, but nevertheless differ. With simple correlation and regression, we examine *N* participants who are observed or tested on two variables. Imagine that we examine the running speed of mice in a maze (Y). This is a random variable that we can not control. When we examine the relation of Y with the number of times that the mouse presses a button before it finally succeeds (X) we have two random variables that we can not influence. In this case, we use the *correlation*. Both variables will vary per experiment (it are random variables) and we are facing a sampling error. When X is a predefined variable, it is specified by the researcher (for example: the number of food grains the mouse receives per succeeded attempt), we have a *regression*.

Correlations

A correlation measures three characteristics of the association between X and Y:

1. The *direction* of the relation. A *positive correlation* (+) emerges when two variables are moving in the same direction. If the value of X increases (for example the length of a person), the value of Y also increased (for example the weight of a person). A *negative correlation* occurs when two variables are moving in different directions. If X increases, Y decreases (or vice versa).
2. The *form* of the association. It can be for example linear.
3. The *degree* of the association. A perfect correlation has a value of -1 or 1. A correlation of 0 implies that there is no association between the two variables. A correlation of 0.8 is therefore stronger than a correlation of for example 0.5

Pearson correlation

The most well-known measure for correlation is the *Pearson correlation*. This correlation measures the degree and direction of a linear relation between two variables. The Pearson correlation is denoted with *r* and calculated as follows: *r* = covariance of X and Y / the variability of X and Y separately. To be able to calculate Pearson r, a new concept has to be introduced, namely the sum of products of the deviations (SP). In previous parts, we used the sum of deviations (SS) to measure the variability of one variable. Now, we use the SP to measure the degree of covariance between two variables. Two formulas can be used to calculate the SP. For one formula however, the mean values of X and Y have to be calculated beforehand.

1. The formula which requires to calculate the means beforehand is:
   SP = ∑(X - MX) (Y - MY).
2. Another formula (which does not require to calculate the means beforehand) is:
   SP = ∑XY – (∑X∑Y)/n.
3. Also, the formula r = (SP / $\sqrt{SS*SY}$) can be used to calculate the sum of squares.
4. Finally, the Pearson correlation can also be calculated for z-scores. In that case, the formula is: r = ∑(zX*zY)/n.

The proportion explained variance

With the Pearson correlation itself, you can not do so much, because it is not ratio scaled and thus not suitable for calculations. Therefore, you have to multiply it. The value r^2 is called the *coefficient of determination*. This value measures the proportion of variance within one variable, that can be explained by the association of this variable with another variable. A correlation of 0.80 (*r* = 0.80) implies for example that 0.64 (r^2), that is 64% of the variance of scores on Y can be explained by variable X. A r^2 of 0.01 refers to a small correlation and a r^2 of 0.09 refers to a medium correlation. A large correlation is characterized by a r^2 of 0.25 or higher.

Spearman correlation

The Pearson correlation quantifies the linear relation between two variables. This correlation measure is used primarily when data are interval or ratio scaled. Other correlation measures are developed for non-linear relations and other measurement scales. The *Spearman correlation* measures the relation between two variables with an ordinal scale. The Spearman correlation can also be used when data are interval- or ratio scaled and there is no linear relation between X and Y.

The Spearman correlation looks for a *consistent* relation between X and Y, regardless of its form. The original values have to be ordered (from small to large). The Spearman correlation can be calculated as follows: $rs = 1 – 6\sum D^2/n(n^2-1)$. In this formula, n refers to the number of scores and D refers to difference: the difference between each order of a X- and Y-value. For example, one can have the second best score on variable X, but the ninth on variable Y.

The point-biserial correlation

A special variant of the Pearson correlation is called the point-biserial correlation. This correlation is used when one variable consists of number, but the other variable consists only of two categories. A variable with only two categories is called a *dichotomous variable*. An example is gender. To calculate the point-biserial correlation, the dichotomous variable first has to be transferred to a variable with numerical values. One value (for example women) receives a zero and the other value (for example men) receives a one. Next, the formula for Pearson r is used. The point-biserial correlation can also be described as:
$r = SP / \sqrt{(SSx)(SSy)}$. Quaring the point-biserial correlation results in the proportion of explained variance, which is a measure of effect size. The relation between proportion explained variance and a t-test for independent samples is: $r^2 = t^2/(t^2 + df)$. It can also be written as: $t^2 = r^2(1/r^2) / df$

The phi-coefficient (ϕ)

The phi-coefficient (ϕ) measures the relation between two variables that are both dichotomous. To do so, first the values 0 and 1 have to be given to both variables. Next, the Pearson r formula can be applied.

Strong and weak correlations

For large samples, even very small correlation may become statistically significant quickly. A significant correlation tells us nothing more than that the chance is very small that the correlation in the population equals zero. The presence of significance thus does not imply whether the relation between the variables is strong. The strength of a correlation is in accordance with the size of the correlation and not with the statistical significance of the correlation. The rule-of-thumb is that a correlation of .10 is weak, a correlation of .30 is moderate, and a correlation of .50 is strong.

Scatterplot

A useful way to examine the relation between two quantitative variables is a scatter plot. Each participants is displayed by a dot with coordinates, that refer to the values on the variables X and Y. Normally, the predictive variable is presented on the X-axis and the criterion variable is presented on the Y-axis. The criterion variable is predicted from the predictor variable. However, if it concerns a correlation coefficient, it is not always clear which variable is X and which is Y. In that case, it does not matter how the variables are labelled. In a scatterplot, a line is drawn through the cloud of dots as best as possible. That line is called the *regression line* of Y predicted by X (that is: Y on X) which gives the best approximation of Yi for a value Xi. If the regression line is straight, the relation between the variables is linear. If the regression line is curved, it is called a *curvilinear relation*.

The degree to which the dots lie around the regression line is related to the correlation (r) between X and Y. The closer the dots (the observed results) lie around the regression line (the predicted results), the higher the correlation. The correlation coefficient ranges from -1 to +1, in which a perfect correlation (all points are on the regression line) is referred with 1. The plus and minus sign indicate the direction and do not influence the relation between the variables.

Simple regression

The general formula for a simple regression is Y = b0 + b1X + e, in which Y refers to the dependent variable and X to the independent variable. The parameters that have to be estimated are called the intercept (b0) and the regression weight (b1). The error (e) is the difference between the estimated and observed value of Y. For example, you have to pay 5 euros per hour next to the 30 euros entrance fee for a tennis club. In this case, the regression formula is: Y = 5X + 30. The regression coefficient (*slope*) is b1, which shows how Y changes when X increases by one point. In this example, the regression coefficient is 5, because the total cost increases with 5 euros per hour. The value of b0 is called the *intercept*, because it shows the value of Y when X is zero. If the regression coefficient equals zero, the regression line is horizontal.

The relation between X and Y can also be displayed graphically. The most frequently used method to do an optimal prediction is the *least squares* method in which parameters are chosen such that the sum of the squared predicted errors is as small as possible.

Assumptions for regression

A few assumptions have to be met. First, there has to be *homogeneity of variances*. That means that the variance of Y is the same for each value of X in the population. In addition, the values of Y that are in accordance with the X-values have to be normally distributed.

When examining the sample correlation, we replace the regression model assumptions with the assumption that we draw a sample from a bivariate normal distribution. The *conditional distributions* in this distribution are the distributions of Y and X given a specific value of X or Y. When we look at all Y-values, independent of X, we call it the *marginal distribution* of Y. Finally, we assume that the relation between X and Y is linear.

Predicted values

To determine how well a line fits the data, we first have to calculate the distance between the line and each data point. For each X-value, the linear regression line determines the value for the Y variable. This value is called the predicted value ($\hat{Y}$). The distance between this predicted value and the actual Y-value is determined by the following steps:

1. Distance = Y - $\hat{Y}$. This distance measures the error between the line and the actual data.
2. Because some distances are negative, and others are positive, the next step is to square each distance, so that only positive values remain.
3. Finally, the total distance between the line and the data has to be calculated. The squared values from step 2 are summed up: ∑(Y - $\hat{Y}$)^2. This is called the *total squared error*.

An example

Assume we want to examine the association between stress and mental health. The latter is measured with a checklist. The first step to calculate the correlation is to calculate the *covariance* (covXY or SXY), which refers to the degree to which the two variables vary together. It looks much like the variance, because if we replace all Y's with X (or all X with Y) we obtain s^2X (or s^2Y). The formula is given as: (∑XY - ∑X∑Y / N) / (N – 1). We expect a strong positive relation: higher values of X (stress) give higher values of Y (mental health). This will result in a high covariance value. If there would have been a strong negative relation, the sum of products of deviations from the mean would have been large and negative. If there would have been no relation between the variables, the sum would be approximately zero. The covariance of the example is 1.336. To calculate the correlation coefficient, we have to take the standard deviations of X and Y into account.

R = covXY / sXsY

The correlation ranges from -1 to +1. In the example, the correlation is *r* = .529. This does not mean that there is a 53% relation between stress and symptoms. It only indicates the strength of the relation between the two variables; values closer to +1 indicate a stronger relation. The plus and minus indicate the direction of the relation, in which a positive correlation indicates that higher values of X are in accordance with higher values of Y.

Standardized regression coefficients

When the data is standardized, a difference of one unit in X refers to a difference of one standard deviation. If the slope is for example 0.75 (for standardized data), the Y will increase with 0.75 for each increase of one standard deviation of X. The slope of standardized data is called *standardized regression coefficient* or β .

For standardized data, it applies that sX = sY = s2X = 1, in which the slope and correlation coefficient are equal. A correlation of r = .80 implies that an increase of one standard deviatin of X is associated with 8/10 standard deviation increase of Y. However, because it is a correlational association, we can not make claims about cause-and-effect.

Hypothesis tests for regression

The significance of b

If X and Y correlate, and there is a linear relation, the slope of the regression line will not be equal to zero and b will have a value different from zero. This is the case for one predictor variable, but when there are multiple predictor variables, the slope does not have to be significant for each of these variables.
b* is the parametric equivalent of b, namely the slope if we had X and Y measures on the whole population.

The standard error is: $s_b = \frac{Y - X}{X\sqrt{N-1}}$

To test if the population slope is zero, we use the formula:
t = (b-b*)/sb = b / (sY*X / sX * √(N-1)) with N-2 degrees of freedom.

The confidence interval of b* is: CI(b*) = b ± (ta/2)($\frac{Y - X}{X\sqrt{N-1}}$)

The difference between two independent b's

It may happen that two samples have a different slope. You can test if this difference in slope is significant as follows.

H0 = b*1 = b*2 and the sampling distribution of b1 – b2 is normally distributed with a mean of zero and a standard deviation of sb1 – sb2 = $\sqrt{s_{b1}^2 + s_{b2}^2}$ with t = (b1 – b2) / $\sqrt{s_{b1}^2 + s_{b2}^2}$ with N1 + N2 – 4 degrees of freedom.

Filling in the standard error with the formula described above, we obtain:

sb1 – sb2 = $\sqrt{\frac{s_Y^2 - x_1}{s_{x1}^2(N1-1)} + \frac{s_Y^2 - x_2}{s_{x2}^2(N2-1)}}$

If we assume homogeneity of error variances, we can combine the two estimates, weighted by the degrees of freedom:
s^2Y*X = ((N1 – 2)s^2Y * X1 + (N2 – 2)s^2Y * X2(N1 + N2 – 4)

Chapter 8 – What is multiple regression?

Predicting and explaining (causal) relations can also be important when there are more than two variables, because a phenomenon can be predicted by multiple factors. It is wise to take into account as many factor as possible Using a multiple regression has three advantages compared to using Pearson correlations. First, it provides information about optimal predictions of Y by a combination of X variables. Second, it allows you to determine how well the prediction is, by examining what the total contribution is of the set of predictors on the prediction. Finally, it allows you to determine the contribution of each predictor separately. It is important to note that the most optimal prediction is not per se a correct prediction. The last advantage can be used to determine more clearly a causal relation, or to determine the added value of a predictor. The formula for multiple regression is: b0 + b1X1 + b2X2 + … + bpXp.

Multiple correlations

The multiple correlation (R) has a value between 0 and 1, and hence can not be negative which is different from the Pearson correlation. R2 refers to the proportion of explained variance of Y, in which a higher R2 indicates a better prediction. To correct for an overestimation of shared variance, one can use the *adjusted R2* which is calculated as:
1 – ((1-R2)(N-1) / (N-p-1)). The predictors can thus have shared and unique variance. The unique variance can be displayed with the squared semi-partial correlations. Sometimes, there is suppression, that is when the unique contribution of a variables is larger after correction for another variable, than without correction. In other words, the real effect of X1 on Y was suppressed by the relations of X1 and Y with X2.

Partial and semi-partial correlation

The (semi-)partial correlation coefficients control for the effect of one or more other variables.

Partial correlation

The *partial correlation* r01.2 is the correlation between two variables with one or more variables removed from both X and Y. Imagine that we want to examine the relation between income and school achievement. We find a significant correlation between these two variables. However, this does not mean per se that success on school results in a higher income. It might also be explained by for example IQ: this causes both higher school achievements and a higher income. The way to examine this, is by calculating the partial correlation between school achievement and income, after removing IQ from both variables.

For the partial correlation, we conduct a separate regression analysis for both variables with the to be controlled variables (in the example: income with IQ and school achievement with IQ). We take the residual of both analyses. This is the part of variance that is not explained by IQ. The correlation between these is the partial correlation.

The notation of the partial correlation coefficient is r01.23..p in which the correlated variables stand left from the dot, and the variables which are controlled stand right from the dot.

The squared partial correlation is the proportion of explained variance.

Semi-partial correlation

The *semi-partial correlation* is the correlation between criterion Y and a controlled (partialled) predictor variable. While the partial correlation removes a variable from both the criterion and the predictor, here we only remove a variable from the predictor. The semi-partial correlation is the correlation of Y with that part of X1 that is independent of X2 (the residual). The notation of the semi-partial correlation is: r0(1.2) in which we remove variable 2 from predictor 1. For the correlation, it applies that: r20(1.2) = r20.12 – r202.

Constants and regression weights

In general, the constant does not have an intrinsically value for psychologists and is therefore difficult to interpret. In addition, the interpretation of the regression weights can be difficult, because the measurement units are often arbitrary. This also makes it difficult to determine which predictor is most important. The latter problem can be resolved by using standardized regression weights. Standardized regression weights are noted with the sign β . This way, you are independent of measurement units and you can compare different predictors well. However, this has the negative consequence that you are dependent on the standard deviation within samples, which is especially problematic if you want to compare different studies. Regression weights are always partial, which implies that they are only valid when all variables are included in the equation. Thus, when a correction is applied for the effects of all other variables. You can not examine the regression weights as something separately, but only within the context.

Testing: from samples to population

So far, we only looked at descriptive statistics. However, we can also use inferential statistics to say something about the population, from which the sample is drawn. To determine if the total contribution of all variables differs from zero, the F-test can be used. To determine the unique contribution of each predictor, a t-test can be conducted for each predictor. However, the more predictors (the more t-tests), the larger the chance on a type-I error. Therefore, the F-test is used as a kind of 'gatekeeper' to determine how many t-tests should be considered. If the F-test is significant, t-tests are conducted. The F-test is calculated as:
F = ((-p-1)R2)/p(1-R2) with N participants/observations and *p* predictors. There are *p* and *N-p-1* degrees of freedom involved. For the t-test, the standard error of the statistic is required. That is the variance of the statistic over repeated samples. The test is: t = (bj – bj*)/sbj with N-p-1 degrees of freedom. To test the null hypothesis: bj* = 0, we use t = bj/sbj.

Assumptions

Different assumptions have to be met:

1. The dependent variable should be interval scaled; predictors can be binary or interval scaled. Fortunately, multiple regression is fairly robust for small deviations of the interval level.
2. There is a linear relation between the predictors and the dependent variable. With a standard multiple regression, only linear relation can be identified (and for example no curvilinear relations). Deviations can also be determined with a residual plot.
3. The residuals have (a) a normal distribution (b) the same variance for all values of the linear combinations of predictors and (c) are independent of each other.

The assumption of normally distributed residuals is not very important to consider, because regression tests are robust against violations when the sample is large enough (N > 100). Often, the assumption is checked with a histogram. The assumption of heteroscedasticity (3b) should be checked properly, because regression is not robust against violations of this. A residual plot is used for this. The latter assumption (independence of mistakes, 3c) is very important, but difficult to check. Fortunately, most research designs meet this assumption. Checking assumptions is thus always dependent on the assessment of researchers and can thus be interpreted differently by people.

Multicollinearity and outliers

Outliers are scores of three of more standard errors above or below the mean. It is important to consider why the score of an individual is an outlier in the analysis. In addition, outliers can have a disproportional influence on the regression weights. If you decide to exclude outliers from the analysis, it is good practice to be very explicit about this in you report, and note why you chose to do so.

Different problems may arise when correlations between dependent variables are strong. Sometimes, the regression does not provide any results. In other cases, the estimates are unreliable or it is difficult to interpret the results. To check for multicollinearity, you can check the *tolerance* of each predictor (it should exceed 0.10). Tolerance is calculated as

1 – R2j,in which Rj is the multiple correlation between variable j and all other predictor variables. Furthermore, you can check the VIF which can be calculated as 1/tolerance. This should be as low as possible, at least below 0.10.

Mediating and moderating relations

In psychology, mediators and moderators are important: variables that play a role in the relation between two other variables.

Mediation

A mediator mediates the relation between two other variables. For example: the degree of self-confidence is mediated by the amount of care received from parents and the way someone thinks about raising children (Caring parents result in a high confidence, which results in confidence to raise children).

Baron and Kenny wrote a lot about mediation. They mention three steps that have to be taken, in order to have a mediating effect. First, you have to show that the independent variable has a significant relation with the mediator. Second, you have to show that there is a significant relation between the mediator and the dependent variable and between the independent and dependent variable. Finally, you have to demonstrate that, when the mediator and independent variable are used together to predict the dependent variable, the path between the independent and dependent variable (c) becomes less strong (preferably non-significant). But, when path ‘c’ does not disappear fully and remains significant, what then? One way is the Sobel test, with which we question whether the full mediating path of the independent variable to the mediator to the dependent variable is significant. For this, we need the regression-coefficients and standard errors of the two paths. The standard error of Beta (s β) is not given and should be calculated as: t = β/sβ , so s β = β/t.

Moderation

With moderating relations, the relation between independent and dependent variables changes by a third (moderator) variable. For example: we examine the influence of faily stress-events on the number of symptoms of stress as indicated by the student. In addition, we find that when the student receives much social support, he shows less symptoms than someone who receives little social support.

Chapter 9 – What is logistic regression?

Logistic regression

This chapter is about logistic regression with a categorical dependent variable and quantitative or dichotomous independent variables. In a normal logistic regression, there is always a dependent variable (Y) and a set of independent variables (X’s) that can be dichotomous, quantitative or a combination of both. The dependent variable can be dichotomous (in a binary logistic regression) or categorical with multiple categories, which refers to polytomous or multinomial logistic regression. Binary logistic regression is a technique with which a regression analysis is conducted for a dichotomous dependent variable. It provides a model for the chance that an event occurs dependent on the values of the independent variable. For example, for predicting the response to a treatment for cancer, and the participants can ‘survive’ or ‘not survive’. The categorical independent variables can be both categorical and continuous.

Assumptions logistic regression

- No multicollinearity (when more than two predictive variables correlate strongly).
- No errors in the specification. All irrelevant predictor variables are excluded.
- The independent variables have to measured on a numeric scale, either ratio or interval.
- The errors are independent of each other, so each observation is independent of other observations.
- The dependent variable should be binary.
- Large sample size, preferably 30 times the number of estimated parameters.

Coding binary variables

It is good practice to code the presence of a characteristic with 1 and the absence of that characteristic with 0. Variables that are examined, are labelled with 1 (response group, comparison group, purposive group), others with 0 (reference, basic or control group). The aim of a logistic regression is to predict to which group each individual belongs. This is obtained by calculating the chance that the individual belongs to category 1. An advantage of this coding is that the mean of the dependent variable equals the proportion ones in the distribution. The mean is also equal to the chance to label a random person as 1 in a random sample.

- P = proportion ones.
- Q = proportion zero’s (1 – P)
- PQ = variance
- √PQ = standard deviation

There are more than two categories for the dependent variable with multinomial logistic regression. These are often coded as 1, 2, 3 and so on. The reference group should be identified and the other groups are used as target group in separate analyses.

Graphical displaying logistic regression

The graphical display of a linear regression is a line, with which it is assumed that the proportions are constant. If x changes with a certain amount, y changes with a certain amount for all numbers. With logistic regression, the line is S-shaped. As a result, we can predict the chance on outcome 1, based on the value of the predictor. The first and last values of X hardly result in differences. Difference is present in the middle: the steeper the slope, the more difference is present. A logistic regression is used when the relation is not constant. In those cases, a logistic regression has a high predictive value.

Logistic regression and odds

To be able to conduct a logistic regression, you first have to transform the data with the *natural log transformation*. Below, you’ll find some core definitions:

- *Odds*: for a dichotomous variable, the odds of group membership equal the probability of membership of that group divided by the probability of membership of another group. Odds imply how likely it is that an observation belongs to a certain group, compared to another group.
- *Chances*: the chance to belong to one group divided by the chance not the belong to that group = P / (1 – P). It ranges from 0 till infinity.
- *Odds ratio*: another important concept is the odds ratio, which estimates the change in odds of group membership of a target group per one-unit increase of the predictor. The raw coefficient of the predictor variable indicates the change in the natural logarithm of the odds ratio, which is more difficult to interpret than the odds ratio. This raw coefficient does have a useful function: a positive raw coefficient implies that the predicted odds ratio increases when the predictor value increases and vice versa. For a raw coefficient of 0, the odds ratio is 1 (de odds are the same for each value of the predictor).

We want to calculate what the chance is that an individual belongs to a certain group. To do so, the probability of the event is transferred to chances. This is done by taking the natural log (ln). As a result of this transformation, the data fit the S-curve to predict the group membership as good as possible. The logistic regression equation with *v* independent variables: ln[chances] = grouppred = a + b1X1 + b2X2 + … + bvXv in which grouppred refers to the predicted group membership. The *b* coefficients give the change in log chances for membership for a change of one unit for the independent variables, controlled by the other predictors. The values of *b* (slope) and *a* (constant) are calculated by using the Maximum Likelihood Estimation (MLE), that you can obtain after transforming the dependent variable in the logit. This is a method to change the data to obtain a linear function. The scores are transferred to chances, and then to *log odds*[llog(p/1-p)] with *p* the chance on improvement and 1 – p the chance on no improvement. The log odds are positive for odds larger than one and negative for odds smaller than one.

X is the score of the predictor. This can be either 0 or 1 for dichotomous variables or a range of numeric values for quantitative variables. It implies how likely it is that the observed value of the dependent variable can be predicted from the observed values of the independent variables.

The logistic function can be described as P = en / 1 + en. The logistic function has a range of 0 to 1. If *n* is large and negative, the chance P is small. If *n* is large and positive, the chance P is large. When it applies that: n = 0, then e0 = 1. The corresponding chance is 1/1+2 = 0.5.

In the logistic function, *n* is replaced by a linear regression part: P1 = ea + b1x1 + … / 1 + ea + b1x1 + b2x2 + … P1 is the chance of succeeding (success = 1), a is the constant under B (from the SPSS table), b1 and b2 are the corresponding regression coefficients, x1 andx2 are the corresponding predictors. The outcome is interpreted with the following rule: if P1 is equal than or larger than 0.5, the code is 1, if P1 is smaller than 0.5, the code is 0. The chance ratio can be calculated from the e and the b-coefficient: eb = chance ratio.

Evaluation of the logistic model

2 Log Likelihood Test examines if the set of the independent variables can predict the dependent variables better than on chance alone. The likelihood values are often very small and therefore the natural log is presented in the output. This is calculated by multiplying the log likelihood value with -2, so that the significance can be tested with the X^2 test. This the the -2LL (log likehood). It is tested whether at least one predictor has a significant contribution, different from zero. The higher the -2LL, the less well the model fits the data. The 0-model always fits the data least.

To compare models with each other, the model without predictors is compared to the model with one parameter. The difference between the -2LL values indicates the change in X2 that is caused by adding a predictor. The difference can be examined with 1 df. In the Model Summary* (SPSS), the -2LL shows you the strength of the relation. The -2LL is included in the formula of Hosmer and Lemeshow: RL2 = -2LLmodel 0 - -2LLmodel x / -2LL model0. You always compare the current model, for example model 1 or model 2, with the zero model. RL2 gives the proportion reduction in -2LL. For the null model, see the ‘Iteration history’ in SPSS.

Classification analysis

The percentage accurate classified cases (PAC) is the number of correct classified cases divided by the total number of classified cases. However, a different measure of accuracy can be used. Sensitivity is the percentage of the target group that is classified correctly. Specificity refers to the percentage of the other group that is classified correctly. The negative predictive value is the percentage that is correctly allocated to the other group by the model. If you want to conduct a good prediction for both groups, the mean predictive value over classes is very useful. Finally, it is important to take the generalizability of the results into account, for example by using a cross-validation sample.

Chapter 10 – How to conduct an analysis of variance (ANOVA)?

Analysis of variance (ANOVA) is a way to test hypotheses. By means of an ANOVA, the difference in means between two or more groups are examined. ANOVA has a main advantage compared to the traditional t-test. T-tests can only be conducted for comparing two groups. With an ANOVA, more than two groups can be compared. With an ANOVA, an independent variable or a quasi-independent variable (for example gender) are called a *factor*. The individual groups or treatment conditions that are part of the factor are called *levels* of the factor. Another advantage of ANOVA compared to t-tests is that the chance on type-I remains equal when comparing multiple hypothesis. Normally, there is risk on a type-I error of the selected alfa level (often 5%) for each individual comparison. With multiple conditions, there are multiple hypothesis tests required to compare all of them, and for each hypothesis test there is a chance on a type-I error. These result in an increased risk for the total experiment compared to the alpha level of an individual hypothesis test. The advantage of ANOVA is that all comparisons that are required to test all different hypotheses of one experiment, can be conducted at once. Consequently, the alpha value remains at the selected value (often 0.05).

One-way ANOVA

The one-way ANOVA is an analysis of variance in which only one independent variable is examined. This is for example the case when a therapy for depression is offered to three different conditions. The structural model of the analysis of variance is: $X = \mu + \tau_j + \epsilon_{ij}$. τ_j refers to the difference of the group means and the large mean. ϵ_{ij} refers to the difference between the individual score and the group mean.

Assumptions ANOVA

Three assumptions have to be met to conduct an ANOVA:

1. Homogeneity of variances (homoscedasticity): each group that is used in the study has the same variance. You can check this with Levene’s test for equal variances. The F-test is robust for this assumption when the largest group and the smallest group do differ no more than a factor 1.5.
2. Normal distribution of the error: the second assumption is that the scores are normally distributed for each condition or sample. Because deviation from the mean is also called error, this assumption is also called ‘the normal distribution of the error’. The F-test is robust for non-normality if *n* is larger than or equal to 15 in each group.
3. Independent scores: the third assumption is that the observations or scores are independent. That implies that, if we know one observation, it does not provide any information about another observation. This may go wrong when the participants are not allocated randomly to a group.

The ANOVA is, in general, a quite robust test. That means that the assumptions may be violated to a certain extent, without having much consequences for the test. When the populations are fairly symmetrically distributed, and the largest variance is no more than four times as large as the smallest variance, the ANOVA is still valid. If the sample sizes differ substantially, the test is less robust against heterogeneity of variances.

Hypotheses for ANOVA

Imagine, you examine three conditions. Then the null hypothesis is $\mu1 = \mu2 = \mu3$. This means that the mean of all three conditions is the same. The alternative hypothesis is that at least two population means differ from each other. The alternative hypothesis can also be more specific: $\mu1 \neq \mu2 \neq \mu3$. That means that all three population means differ from each other. With ANOVA, the t-statistic is also called the *F-ratio*. F = variance between sample means / variances expected on coincidence (thus when the treatment does not have an effect). The F-ratio is thus calculated over the variance, and not over the difference in sample means. A found F-value is the same as the squared t ($F = t^2$). If the researcher for example conducts a t-test with two independent samples, a difference of means may result. For example, the researcher finds a t-value of 2.00. If the researcher would have used an ANOVA, the F-value would be 4.00.

ANOVA

Imagine: you have three samples. The first step is to determine the total variance in the full data set. This can be obtained by combining all scores of the samples. Next, the total variance has to be subdivided into parts. The total variability can be subdivided into (1) *between-groups variance* and (2) *within-group variance*. Between-groups variance occurs when one group systematically scores higher or lower than the other group(s). Within-group variance occurs when there is variance within each group. The aim of ANOVA is to discover whether differences between conditions are coincidence or not. If there is a coincidence phenomenon, there there is no effect. In that case, differences between the scores are only present due to the fact that each sample consists of different individuals. If there is an effect, the difference between the groups should be larger than expected based on coincidence alone.

The F-ratio

After subdividing the total variance into two parts (between- and within-group variance) these parts have to be compared. This can be done with the F-ratio. For an ANOVA with independent samples, the F-ratio is de variance between the conditions divided by the variance within the conditions. If there is no effect, the difference between the treatments is the result of coincidence only. In that case, the F-ratio is 1. A large F-ratio implies that the difference between conditions is larger than expected based on coincidence alone. Wit ANOVA, the denominator of the F-ratio is called the *error term*. The error term provides an indication of the variance resulting from coincidence.

Important symbols

1. The letter *k* refers to the number of conditions (the levels of the factor).
2. The letter *n* refers to the number of scores in each condition. The total number of scores in the study is noted as *N*.
3. The total (∑X) for each condition is noted with the letter *T.*

Calculating with ANOVA

It is important to be able to calculate with the ANOVA and to understand the logic behind it. First, the formulas are introduced.

SSB = Sum of Squares Between

$$SSB=\sum_{i=1}^{a} n_i(\acute{y}_i-\acute{y})^2$$

The number of people within one group times the squared difference between the group mean and the total mean. And summing that up over all groups.

SSW = Sum of Squares Within

$$SSW=\sum_{i=1}^{a}(n_i-1)s_i^2$$

The number of people within a group -1 times the variance of that group, summing up over all groups.

$$SST=(n-1)s^2$$

SST = Sum of Squares Total

The total number of people within the study -1 times the total variance.

SSB + SSW = SST

Degrees of freedom for ANOVA

Each degree of freedom is related to a specific SS-value.

1. The number of degrees of freedom of the total (df total) is found by adding up the number of scores (of all conditions combined) minus 1 (df total = N – 1).
2. Next, the degrees of freedom for the within group variance have to be found (df within). This can be obtained by df within = ∑(n-1) = ∑df in each treatment. Another formula is N – k.
3. Finally, the degrees of freedom for the between-group variance have to be determined (df between). Df between = k – 1 (the number of conditions minus one). If the degrees of freedom of the within-group variance and the between-groups variance are added, you obtain exactly the degrees of freedom of the total.

Mean Squares

Next, the variance between and within conditions has to be calculated to find the F-ratio. With ANOVA, the term *mean square* (MS) is used instead of variance. The corresponding formula is the same as for the variance: MS = s^2 = SS/df.

1. To find the MS between groups, use: MSbetween = s^2between = SSbetween / dfbetween.
2. MSwithin = SSwithin / dfwithin
3. Next, the F-ratio can be found by dividing these values by each other:
   F = MSbetween / MSwithin

The F-distribution

The null hypothesis is correct, when the F-ratio is 1. Because the F-ratio is calculated by means of two variances, the F-value is always positive. Found F-values can be looked up in the F-table. These are made such that first the degrees of freedom of the denominator and then the degrees of freedom of the numerator have to be looked up in the table. Then, the F-value can be found in that part of the table. The F-value lies between the two values that are found in the table. The chance on these values are also mentioned in the table. For example, when there is a 1% chance for the obtained F-value, the null hypothesis can be rejected for an alfa of 5%. If there is chance of more than 5%, then the null hypothesis can not be rejected.

Example of a hypothesis test with ANOVA

To conduct an analysis of variance, four steps have to be taken:

1. Formulate the null and alternative hypothesis. For example:
   a. H0: µ1 = µ2 = µ3
   b. H1: at least one of the population means differs from the others
2. Determine the degrees of freedom for the between- and within-group variance to determine the critical region for the F-ratio.
3. Conduct the following computations to find the F-ratio: calculate the MSbetween and MSwthin. Next, the F-ratio is: F = MSbetween / MSwithin.
4. Finally, check if the found F-value lies within the critical area. As with the t-test, the null hypothesis is rejected when the F-value falls within this critical area.

It is important to remember that the size of the sample may influence the results of the ANOVA. The larger the sample size, the higher the chance to find evidence to reject the null hypothesis. Such a problem can be avoided by an alternative statistical analysis: the *Kruskal-Wallis test*. With this test, the data are transferred to ordinal scale, and rank scores are used. The Kruskal-Wallis test can also be used when the assumption of normality is violated. This test uses medians.

Effect size

As with other tests, a significant result only is not sufficient. We also need to know if the results are practically significant. For the F-ratio, we can use the *r*-family of effect sizes. For the ANOVA, the effect size refers how much variance of the dependent variable can be attributed to a treatment effect. Two of the most frequently used statistics are $\eta^2$and ω^2

Eta-square η²

SS_treatment is a measure for the degree of observation differences between the different conditions. SS_total is the measure for the differences in the complete data set. These two SS's divided by each other provide a percentage of variance for the treatment:

$$\eta^2 = \frac{SS_{treatment}}{SS_{total}}$$

When the sum of squares is unknown, the eta-square can also be computed differently:

η² = 1 /(1+(dferror / (F x dfbehandeling)))

η² assumes that the regression line crosses the mean of each group. However, this does not apply when the measures are biased. η² is the effect size with most bias.

Omega-square ω²

The omega-square is a good measure for the effect of *balanced design* (with equal n's). This statistic provides less bias than η².

$$\omega^2 = \frac{SStreatment - (k-1)MSerror}{SStotal + MSerror}$$

Post-hoc tests

The main advantage of ANOVA (compared to t-tests) is that more than two conditions can be examined. If the null hypothesis is rejected by means of the F-ratio, it thus implies that there is a significant difference. But where is that significance difference? This can be examined with post-hoc tests. Post-hoc tests are always conducted after an ANOVA. The null hypothesis should first be rejected and there have to be more than two groups in order to do meaningful post-hoc tests.

With each post-hoc test, two conditions are compared, so pairs of comparisons are made. For example, with three conditions we can compare µ1 with µ2, µ2 with µ3 and µ1 with µ3. Each pair of comparisons has its own hypothesis test to examine which conditions differ. The disadvantage of post-hoc tests is the increase in type-I error.

(Un)planned comparisons

Statisticians often distinguish between planned and unplanned comparisons.

- A *planned comparison* emerges when the researcher forms expectations that are important for the hypotheses of the study. A means of protection against increasing type-I error s to divide it by the number of planned comparisons. For example, if the researcher uses two planned comparisons, he should divide the alfa of 5% by two. Thus, he uses an alfa of 2.5% for these comparisons.
- *Unplanned comparisons* refer to the situation when the researcher has no idea about the effect and conducts all kinds of post-hoc tests in order to find a significant effect. Here, too, the type-I error has to be minimized. This can be obtained by Tukey's HSD test.

Tukey's HSD test

Tukey's HSD test is used frequently in psychological research. With this test, the minimal difference between conditions required to find a significant effect can be determined. This value is called the *honestly significant difference* (HSD). This value is used to compare two conditions. If the mean difference between these two conditions is larger than the predefined HSD, it can be concluded that there is a significant difference between the conditions. This value can be computed as: HSD = q * √MSwithin / n. The value of q can be found in the corresponding table. To find q, you need to know the number of conditions (k) and the degrees of freedom of MSwithin. The n represents the number of scores in each condition, which has to be equal per condition.

Using a-priori contrasts

With the multiple comparison procedure (MPC), group means are compared. A MPC is used when there are at least three groups. A contrast is a weighted combination of the means. Use for example the hypothesis: Does drinking alcohol cause a disturbed subjective perception of physical reality? The three groups are: not drinking alcohol, moderately drinking alcohol, and drinking a lot of alcohol. You determine a-priori contrasts before you start the study, based on your expectations. You can formulate different hypotheses. A contrast is a combination of population means. The degrees of freedom are: DFE = N – I. The alternative hypothesis can be both one- and two-sided. The confidence interval of ψ is c ± t*Sec.

For example, one can ask: Does the no-drinking population scores higher than the drinking population (both moderately and a lot)? The following hypotheses can be formulated:

H0: μ1 = 0.5(μ2 + μ3).
Ha: μ1 > 0.5(μ2 + μ3)

The following contrasts can be made from these hypotheses:
H0: μ1 = 0.5(μ2 + μ3)
H0: μ1 – 0.5(μ2 + μ3) = 0
H0: μ1 – 0.5μ2 – 0.5μ3 = 0
Ψ = μ1 – 0.5μ2 – 0.5μ3

This is thus your final contrast. The contrast coefficients (ai’s) are: 1, -0.5, -0.5. These indeed add up to zero, as it is meant to. When there are multiple contrasts, it is required that these are orthogonal. That means that the products of these contrast coefficients are zero if you add them. Imagine that contrast 1 has the following contrast coefficients: 1, 1, -2. Contrast 2 has the contrast coefficients: 1, -1, 0. These contrasts are orthogonal, because
(1*1) + (1*-1) + (-2*0) = 0

Chapter 11 – Wat is the two-way ANOVA model?

ANOVA with multiple factors

In practice, behavior is often influenced by different interacting factors. To examine these complex effects, researchers often design studies with more than one independent variable. Thus, researcher manipulate two or more variables to observe the effect on behavior. A design with more than one factor is called a *factorial design*. An ANOVA with two factors combines multiple hypotheses. Therefore, multiple hypothesis tests have to be conducted. Again, the F-ratio is used: differences between the sample means / differences expected by sampling errors or coincidence.

Example

You are interested in the degree to which light and temperature influence the speed of learning. You can create two conditions for light: no light and normal light. For temperature, you use three conditions: 10, 20 and 30 degrees. These conditions of the two factors have to be combined. In total, there are six conditions. These are called *cells*, because the combined factors are presented in a matrix. Each cell refers to a combination of the two factors. For example, a condition with 20 degrees and no light, a condition with 30 degrees and light and so on. The researcher is interested in three things:

1. The differences in means between the light levels.
2. The differences in means between the temperature levels.
3. The differences in means that emerge from a unique combination of specific temperature and specific light level. An example is that learning is stimulated by light and 20 degrees.

Main effects

Factors are assigned a letter – the factor light for example receives the letter A and the factor temperature receives the letter B. The aim of an experiment is to check if these factors cause differences independent or combined. The mean of the condition ‘light’ is found by adding all mean scores of the three temperature levels combined with the condition light. In total, there are three row means that have to be used to calculate the mean. The mean of ‘no light’ is obtained by adding the means of the three temperature levels that are combined with the condition no light. Again, there are three row means to add. The difference between these two means is called the *main effect of factor A*. In addition, there are three columns means (for the three temperature levels). The mean of the condition ’10 degrees’ can be found by taking the mean of the condition ’10 degrees with light’ and ’10 degrees without light’. This can also be done for the other two temperature conditions. The difference between the means of these three conditions is called the *main effect of factor B.*

Hypotheses

For an ANOVA with two factors, it is examined whether the main effects of A and B are significant. Thus, two hypotheses arise:

1. For factor A, the null hypothesis is: µA1 = µA2. This hypothesis states that there is no significant difference between the condition ‘no light’ and the condition ‘light’. The alternative hypothesis states that there is a significant difference: µA1 ≠ µA2.
2. For factor B, there is a comparison between the three temperature levels. The null hypothesis is: µB1 = µB2 = µB3. The alternative hypothesis is that at least one mean differs from the others.

Interaction

With a two-way ANOVA it is also possible to examine the unique effect of combinations of factor levels. An *interaction* between two factors emerges when the differences in means between individual levels (cells) differs from expected based on the main effects only. For example, people who can study exceptionally well with light and a temperature of 10 degrees, while this effect does not exists when someone studies with light *or* a temperature of 10 degrees. For the interaction effect too, a hypothesis is formulated:

1. The null hypothesis states that there is no interaction between factor A and B. All differences in means between conditions can be explained by the main effects of the two factors.
2. The alternative hypothesis states that there is an interaction between the two factors. The differences in means between the conditions are not (only) caused by main effects of the two factors.
3. The corresponding F-ratio is: differences in means that can not be explained by the main effects / differences in means that are expected based on coincidence or error.

The interaction-effect can be observed in a graph. You can plot for example the three temperature levels on the X-axis (10, 20, 30 degrees) and the mean scores on the Y-axis. There emerge two lines in the graph: one for mean temperature in combination with no light and one for mean temperature in combination with light. If there is no interaction, the lines are parallel. If the lines are not parallel, there is an interaction-effect.

Testing

An ANOVA with two factors thus has three different hypotheses: the main effect of A, the main effect of B and the interaction effect of A and B. First, divide the total variance in between-groups and within-group variance. Second, divide the between-groups variance into variance of factor A, variance of factor B and variance of the interaction. Within each condition, participants are treated equally. Differences within a condition can thus not be caused by the effects of the condition. The within-group variance can therefore only be caused by coincidence or error. Therefore, we need three types of within-group variance (for factor A, B and the interaction) and we need one within-group variance. Each of these variances is determined by a SS-value and a df-value. MS (mean square) = SS / df.

Formulas

Calculating the F-value for the main effects of A and B is conducted similarly to the calculations for a one-way ANOVA. The difference is that you calculate it for both effects. Next, you can compute it for the interaction between factor A and B: $SS_{A*B} = SS_A – SS_B$. The corresponding degrees of freedom are: $df_{A*B} = df_{\text{between treatments}} – df_A – df_B$. The MS for the interaction is: $MS_{A*B} = SS_{A*B} / df_{A*B}$. The three F-ratio's are: $F_A = MS_A / MS_{\text{within treatments}}$.
$F_B = MS_B / MS_{\text{within treatments}}$. $F_{A*B} = MS_{A*B} / MS_{\text{within treatments}}$.

$FA = MS_A / MS_{\text{within treatments}}$.

Effect size for ANOVA with two factors

For the ANOVA, we use the η^2 (eta-squared) to calculate the proportion of explained variance.

1. For factor A: $\eta^2 = SS_A / (SS_{\text{total}} – SS_B – SS_{A*B})$. This is the same as $SS_A / (SS_A + SS_{\text{within treatments}})$.
2. For factor B: $\eta^2 = SS_B / (SS_{\text{total}} – SS_A – SS_{A*B})$. This is the same as $SS_B / (SS_B + SS_{\text{within treatments}})$
3. For the interaction effect: $\eta^2 = SS_{A*B} / (SS_{\text{total}} – SS_A – SS_B)$. This is the same as $SS_{A*B} / (SS_{A*B} + SS_{\text{within treatments}})$.

Table for a two-way ANOVA

Source	Degrees of freedom	SS	MS	F
A	I – 1	SSA	SSA / DFA	MSA / MSE
B	J – 1	SSB	SSB / DFB	MSB / MSE
AB	(I – 1) (J – 1)	SSAB	SSAB / DFAB	MSAB / MSE
Error	N - IJ	SSE	SSE / DFE	
Total	N – 1	SST		

Using more than two factors

It is possible to use more than two factors for a study. However, if more than three factors are used, the results become incomprehensible. It is therefore best to use no more than three factors.

Chapter 12 – What is ANCOVA?

ANCOVA is a combination of regression analysis and ANOVA and can be used to predict a dependent variable of interval level as accurate as possible based on a number of independent variables. These independent variables are called factors (nominal) and covariates (interval). The combination of different types of predictors allows you to do a better prediction in different kinds of situations. Adding covariates makes it possible to test the effects of factors better by (1) reduced error variance and (2) elimination of systematic bias.

Example

Assume that we want to examine whether large of small calls are easier to control. We have three different car sizes, three different types of participants, with large differences in driving experience. We are not able to match participants on experience, so we assume that the mean driving experience is equal across groups. The dependent variable is the number of driving mistakes and the covariate is driving experience. We want to examine the achievement of participants, independent of their driving experience (thus based on the car size). By adding driving experience as covariate, we reduce the error.

The ANCOVA model

By subdividing the variance into within-group and between-groups, the F-test and other statistical data can be computed. For both ANOVA and ANCOVA, we try to optimally calculate the Y-score for all individuals in the groups. With ANOVA, we only know to which group an individual belongs, so the group mean is the estimated score. With ANCOVA we also know the individual score on the covariate, which allows us to make the prediction of Y more accurate. The ANOVA model can be decomposed into a (1) ANOVA component and a (2) regression component.

An ANOVA model has three components:

- The overall mean: $\acute{Y}$
- The deviation of the group from the overall mean: $a_j = \acute{Y}_j - \acute{Y}$
- The error or deviation of each individual from the group mean: $e_{ij} = \acute{Y}_{ij} - \acute{Y}_j$

This provides the following model: $\acute{Y}_{ij} = \acute{Y} + a_j + e_{ij}$. The variance of Y is subdivided into a between-groups component (a_j) and a within-group component (e_{ij}). In the ANCOVA model, a covariate is added to the formula:

$$\acute{Y}_{ij} = \acute{Y} + a'_j + b_w (C_{ij} – \acute{C}) + e'_{ij}$$

With this formula, we try to predict the Y_{ij} score of each individual *i* from group *j*. The difference between ANOVA and ANCOVA is that with ANOVA we only, we only know to which group the individual belongs, while with ANCOVA we also know the individual score on the covariate. As a result, the prediction of ANCOVA is more accurate. The ANCOVA formula consists of a variance analysis component ($\acute{Y} + a'_j$) and a regression analysis component ($b_w (C_{ij} – \acute{C})$).

Pooled-within vs. total regression and correlation

The regression weight of the covariate is called *bw*, because it refers to the prediction of Y by C within each of the groups (the pooled within-groups regression weight). The assumption that this weight is equal for all groups applies only to the population, not to the sample.

The F-test is ANCOVA

The F-test in ANCOVA is similar to the one in ANOVA, only with an adapted sum of squares and degrees of freedom, in which the overlap with the covariate is filtered. The total adapted sum of squares consists of a between-groups component and a within-group component: SST* = SSb* + SSW*. With the following formula, the total variance of the dependent variable can be calculated: SST* = SST – rYC2SST = (1 – rYC2)SST. rYC2SST refers to the variance explained by the covariate.

The within-group variance is calculated as: SSW* = (1 – rYC(W)2r^{2}YC(W))SSW. The between-group variance can easily be calculated as: SSB* = SST* - SSW*. Before calculating the F-value, the MS (mean squares) has to be calculated by dividing the sum of squares by the degrees of freedom: MSb* = SSb* / k-1 and MSW* = SSW* / N-k-c. Here, the *k* is the number of groups, *N* the sample size and *c* the number of covariates. The F-value is then: MSb* / MSW* with degrees of freedom dfb = k -1 and dfw = N – k – c.

Adapted group means

If the groups differ on the covariate, we use adapted group means. These refer to the best estimate of the means if the groups do not differ on the covariate. The covariance-analysis examines whether the group means differ significantly. The adapted group mean can be obtained from: $\acute{Y}_j = \acute{Y} + a'_j + b_w (C_i - \acute{C})$. Because : $\acute{Y}_j^* = \acute{Y} + a_j$ is the adapted group mean: $\acute{Y}_j^* = \acute{Y}_j$ – bw(C_i - $\acute{C}$). If this is displayed in a diagram, the adapted group mean is found on the intersection of the regression line of the group with the line C = $\acute{C}$.

In general, it applies that the groups with a high mean on the covariate will have a lower mean on the dependent variable after adapting. Groups with a lower mean of the covariate will have a higher mean on the dependent variable after adapting. If the group with the highest score has the lowest score on the covariate, the differences on the dependent variable are thus larger, while if this group also has the highest score on the covariate, the differences will become smaller, disappear or switch. The above applies only for a positive bw. If the bw is negative, the reverse holds.

Reducing error variance

Even if the groups do not differ on the covariate, the second aim of the ANCOVA (reducing error variance) still applies. However, ANCOVA is not the perfect solution, because adding one or more covariates may result in a decreased power and a more difficult interpretation. There are three types of situations for which interpretations may become complex or irrelevant. First, the adapted means may not be equal to the research aim. Second, extrapolation to reality may happen that actually does not exist, or does not mean anything. Third, a covariate may eliminate a part of the effect.

Assumptions

Besides the general assumptions (normal distribution, homogeneity of variances and independent observations), the ANOVA has some additional assumptions about the covariate:

1. *No error in the covariate*. The effect of random error in the covariate is an underestimation of the relation with other variables. The most important consequence of error is a too small adaption in the computation of the adapted group means, which results in false conclusions. However, it is not always better not to include unreliable covariates: some an incomplete adaption is better than no adaption at all.
2. *Linear relation with the dependent variable*. This assumption applies to pooled within-group correlation between the covariate and the dependent variable. The relation between these variables can be displayed best with a straight line, not with a curve. Violation of this assumption results in an underestimation of the relation of the above variables, which results in incorrect adaption of the group means. This assumption refers to the population and can be checked with diagrams, not with a test for non-linearity.
3. *Parallelism*. This assumption states that the regression weight *bw* is equal for all groups. This assumption may have large implications for the results. The assumption of parallelism means implies there is no interaction between the covariate and the treatment. The use of a complex ANCOVA model with separate regression weight and adaptions is not an appropriate solution, because it makes the computation and interpretation more difficult.

Limitations of ANCOVA

The described experimental perspective is not the only possible perspective for the ANCOVA. In addition, post-hoc procedures for the ANCOVA are not described here. As mentioned before, covariates should be used carefully, because they 'cost' a degree of freedom and may be difficult to interpret. In addition, there are limitations in studying existing groups.

Chapter 13 - What are MANOVA (multivariate analysis of variance) and DA (discriminant analysis)?

Sometimes, researcher want to examine the difference of conditions of different dependent variables at the same time. T-tests and ANOVA-analyses can be conducted if there is only one dependent variable. MANOVA (multivariate analysis of variance) is conducted to examine the effect of two or more conditions on two or more dependent variable. MANOVA is quite clear: we have a number of dependent interval variables (p) that we want to predict from one or more nominal variables, divided over k groups. We compare means, but we examine multiple variables at the same time in mutual cohesion (multivariate). With MANOVA, you distinguish between 1 independent variable (one-way), 2 independent variables (two-way) and so on. The question is why not to conduct an ANOVA for both dependent variables. This has two reasons:

- Sometimes, the dependent variables are associated with each other. They may be part of a mutual construct. In that case, the researcher may think it is better to analyse the variables combined, instead of separately.
- The more tests are conducted, the larger the type-I error. The type-I error also increases if we use t-tests or ANOVA's on multiple dependent variables. Because a MANOVA tests the differences of group means on multiple dependent variables at the same time, the alpha remains 5%.

Conducting a MANOVA

The null hypothesis of a MANOVA is: the mean of group 1 on variable 1 equals the mean of group 2 on variable 1 equals the mean of group k on variable 1. If the null hypothesis is true, all groups have the same mean per variable. 'There is no relation between the set of nominal variables and the set of interval variables'. For each dependent variable it applies that the population means are equal. The letter k refers to groups and p refers to the number of dependent variables.

The null hypothesis can be tested with multivariate tests Willks, Pillai's, Hotellings and Roys. They do not always provide the same answers. Which one is best, is difficult to determine. During a MANOVA analysis, you conduct a multivariate version of the F-test. If you reject the null hypothesis, you know that there is a difference for at least one dependent variable between the groups. However, that is not a lot of information. Follow-up analyses have to be conducted, for example the Protected F approach or the Descriptive Discriminant Analysis.

A MANOVA followed up by a series of ANOVA's

If you use a MANOVA to protect yourself against type-I error, you may use multiple univariate tests for each dependent variable after a significant result. This is also called protected F-procedure (with an adapted alfa level by means of *the Bonferroni correction*: a/p, with p dependent variables, so that the alpha level is stricter). However, there is also critique about this method. The main arguments are (1) insufficient protection against type-I errors (2) not taking into account the underlying correlations between dependent variables. To deal with these limitations, a discriminant analysis is a better method.

Assumptions protected F-approach

- Multivariate normality of the errors: each of the dependent variables has a normal distribution of the errors and is also normally distributed for all possible combinations of values for the other dependent variables. When n > 20 per cell, you do not have to worry about this assumption, because multivariate tests are fairly robust.
- Homogeneity of the variance-covariance matrices: equal variances and equal covariances in all groups. If the groups have approximately the same sample size (nmin / nmax < 1.5), you do not have to worry much about this assumption. You can also check this assumption with the Box M test. However, this test is extremely sensitive and quickly significant. Only p-values below .001 can be taken seriously. The only thing

you can do, is report this clearly. You are less certain about the results of your study, what is especially important for determining whether results are significant or not.
- Independent errors: the error of one person does not tell anything about the error of the other person (independent of to which group they belong). This can not be checked with a test, but has to be determined based on the research design.

Discriminant Analysis (DA)

With the discriminant analysis, we try to predict a set of interval variables as good as possible based on differences between groups. With DA, the focus can be on describing differences between groups (DDA; from the group perspective) or predicting to which group an individual belongs (PDA; from the individual perspective). PDA is the opposite of MANOVA, whereas DDA is more an extension of MANOVA. The difference between MANOVA and DDS is that the X and Y are switched.

Discriminant function variates

By using discriminant function variates, we can transfer a set of correlated dependent variables to a new set in which the differences between the groups are presented correctly, but the variables do not correlate anymore. Linear combinations are formed between the dependent variables with the use of two subscripts per weight. A discriminant analysis is presented as:

$$Dj = b1jY1 + b2jY2 + \ldots + bpjYp$$

In which the p refers to the number of dependent variables and j to the variate. The weights for the first discriminant function (D1) are chosen such that there is a maximal difference between the k groups. The same applies to the second and next discriminant function, but the functions can not correlate: they have to be *orthogonal*. The maximum number of discriminant functions is k-1 or p. You have to chose the minimum of these. After formulation the discriminant functions, you have to determine the position of the groups on the functions. For each group on each variate, you replace the Y’s by the group means.

Each variate has an eigenvalue (λi): SSb(i) / SSW(i). The proportion explained variance of the variate is displayed as: Ri1 = SSb(i)/SST(i) = SSB(i)/SSr(i) λi / (1 + λi). The maximum explained variance equals the number of dependent variables p. Thus, after summing up the squared correlation of all variates, you may obtain a value of ∑iRi2 higher than 1. By dividing this by the number of dependent variables, you obtain the proportion of explained variance of the dependent variables.

DA and PCA

DA and PCA differ in the optimization criterion that is used to select the weight. PCA tries to explan as much variance of Y as possible, without taking into account differences between group means. With DA, each discriminant function tries to explain as much differences between the groups as possible. Using DA has two main advantages compared to using the original set of variables: (1) data reduction and (2) easier interpretation. If we use both advantages, we can describe differences between groups (DDA). We can also try to discover underlying dimension, from which the group differences can be reduced.

Disadvantages of DDA

The main disadvantages of DDA are (1) the results are sometimes difficult to interpret (2) the results are more descriptive than confirmative. In addition, DDA is relatively less popular than MANOVA for researchers and journals (although this may be for wrong reasons).

Chapter 14 – What are Random- and Mixed Effects Analysis of Variance Models?

To minimize the type-I error in an ANOVA, we have to control the total alpha-level. But we also have to increase the power. The omnibus F-test can test that total. This test is also used in the ANOVA model. The one-way ANOVA has one independent variable with factors of two or more levels.

In the *random-effects model*, the levels of the independent variable in the sample are randomly taken from all levels in the population. As a result, generalizations can be made to all levels of the populations.

In the *fixed-effects model*, first, the levels of the independent variable are selected. Next, the subject are assigned randomly to levels of the independent variable. In some situations, the researcher can control these assignments, but in other situations he can not. The analysis does not differ between these situations, but the interpretation of the results does differ.

With a fixed-model ANOVA, the treatment groups are selected carefully and remain equal when the experiment is repeated. With a random model, the treatment groups are obtained randomly, and they vary per repeat.

Crossed design with random variable

Sometimes, the design has a fixed factor and a random factor. Assume that we want to test if people recognize capital letters quicker than small letters. The letter size is a fixed factor. By selecting a sample from the alphabet, we use five letters (A, G, D, K, W) in the experiment. The variable ‘letters’ is a random factor. Because one of these factors is random, we will select different letters when repeating this experiment, and we will obtain different F-values. A random effect changes the effect of the test for a fixed effect (here: letter size). To show which effect random factors have, we have to examine the *expected mean squares*:

	Fixed	Random	Mixed
Source	A fixed / B fixed	A random / B random	A fixed / B random
A	σ2e + nb2a	σ2e + nσ2αβ + nbσ2a	σ2e + nσ2αβ + nbθ2a
B	σ2e + naθ2β	σ2e + nσ2αβ + naσ2β	σ2e + nσ2αβ
AB	σ2e + nθ2αβ	σ2e + nσ2αβ	σ2e + nσ2αβ
Error	σ2e	σ2e	σ2e

The F-test for the fixed and random variable model:

- For a random variable: E(F) = E (MSb/MSerror) = (σ2e + nbσ2β) / σ2e
- For the interaction effect: E(F) = E(MSAB/MSerror) = (σ2e + nσ2αβ) / σ2e
- For a fixed variable: E(F) = MSA/MSB = E(σ2e + nσ2αβ + nb σ2a) / (σ2e + nσ2αβ)

The model

In fixed effects models, there is one true effect size. The differences in effect size from different studies result from random sampling error. In behavioral research, studies often switch multiple variables. For example, one study will consists of more women in the study than the other, or different ages and so on. That implies that we are facing with random models instead of fixed models. When we compare effect sizes, the measurement per study consists of random error (as always), but there is also a difference *n* effect sizes caused by the presence of certain variables. We assume that the true effect are random and normally distributed around a certain value. The model for the meta-analysis is then:

Yi = μ + τj + εij

With μ s the overall mean effect, τ as the difference between the true effect in the study and the overall mean effect and ε as the sampling error. We are thus facing a variance by τ and a variance by ε.