1. Population, random sampling, scale

1.1. Making a statistical study

The making of a statistical study can be roughly divided into five main points.

1. The planning of the study - Choosing the topic, limiting the topic, choosing the study type and methods, making a research plan
2. Obtaining the material - Interviews, collecting questionforms, etc.
3. Researching the material - Modifying the material to electronic form, describing the material
4. Applying mathematical models - Testing, regression, correlation
5. Making the conclusions - Writing the raport, generalization to population

You have to spare enough time to each step. The step requiring the greatest amount of time is usually obtaining the material. This happens when there aren't any usable materials available already. While preparing the study take notice, that the more care you put to phase one, the easier the rest is.

There are lots of differentv types of researches, because we need specific qualities for different researches. The types are non-exclusive so that one research can be a mixer of various types. Few types are listed below.

THEORETICAL STUDY

Most studies in natural sciences are theoretical studies. In these studies we draw new conclusions and theories based on old previous ones.

EMPIRICAL STUDY

Many empirical studies rely on statistics. We collect and analyze material and based on this analysis we create theories.

BASIC STUDY

Usually before the real statistical study we do a basic study. In this study we'll find out the basic nature of the phenomenon and it answers to the question "what kind of". This study has to be done because otherwise limiting the real topic can be really difficult.

APPLIED STUDY

Often a corporation has a problem, which they want to solve. In an applied study we find the solutions to this specific problem. This study has the features of a case-study, which doesn't aim to form a global theorem but only aims to find the best possible solution to this specific problem.

QUALITATIVE STUDY

A qualitative study is like basic study. We aim to describe the phenomenon and answer questions "why", "how", "what kind of".

QUANTITATIVE STUDY

A quantitative study is quatifiable study. The aim of this study is to get generizable results. It answers to questions "how much", "how often".

1.2. Basic concepts

UNIVERSE = POPULATION

= the group under study aka. all those who we want to research. For example, if we are making a study, where we want to discover finnish 20-year-old feedinghabits, the population is all the 20-year-old finnish people.

STATISTICAL UNIT

= the object of the study. In the example above one statistical unit is a 20-year-old finnish person.

SAMPLE

= a smaller subset from the population. A sample is like the miniature population. The presentability of the sample means how similar the sample and population are. The aim is to get the qualities of the population to the sample in the right ratio.

SAMPLE SIZE

= the amount of statistical units chosen to the sample.

VARIABLE

= the quality that is being measured/researched. Every question in a form measures one variable. Measuring is the event where a statistical unit includes a number or a symbol to

the variable. So in a form the question "age" measures the variable "age" and when a person writes down his/hers age, the measuring happens.

CONSTANT

= the variable that doesn't change values, aka. that has the same value with every statistical unit.

1.3. Scale of the statistical variable

The level of the measuring is described with a scale. We group variables to four groups according to their scale. This is important to know because the scale defines what can be done with the variable.

Nominal level

The variables in the nominal level can be classified to exclusive groups (the values of the variable). These values are mere labels; they express no mathematical properties. The values are merely describing the variable. A good example is hometown.

Ordinal level

To separate the ordinate level from the nominal level, the values indicate the relative position of items, but not the magnitude of difference. These variables are still qualitative variables, so the magnitude of difference is abstract. Many oppinions are measured on a likert-scale (I totally disagree, I partially disagree, I don't know, I partially agree, I totally agree). This is a ordinal level, because we cannot define the difference between these.

Interval level

A variable measured in the interval level differs from the previous variables so that number values indicate also the magnitude of difference. This means that we can do calculus with these variables. The only thing missing is the absolute zero point or it's contractual. A good example is the temperature in celsius degrees. There the absolute zero is $-273,15°C$. This means that the ratio of two values is not unambiguous. When we say "Today's twice as warm than yesterday", it's not mathematically correct way to say it.

Ratio level

The ratio level differs from the interval level so that the absolute zero is zero. Good examples are weight and hight. Most variables in the ratio level are continuous but amounts

are an exception. For example the amount of people living in a city is also measured in the ratio level because the absolute zero is zero.

OBS! In statistical studys variables indicating oppinions are usually formed from several individual questions by adding the values together (sumvariable). This is why oppinions are usually treated as measured in the interval level although they are measured in the ordinal level.

1.4. Classification of variables

We can classify variable with other criterias:

Describing the variable

- Qualitative variable = answers the question "what kind of". This variable describes, it's values are numbres which don't indicate the magnitude of difference.
- Quantitative variable = answers the question "how much", "how often". It's values are numbers which indicare the magnitude of difference.

The possible values

- Discrete variable = is a variable, which can only have single values (for example whole numbers).
- Continuous variable = is a variable, which can have all real values from an interval.

Causal connection

- Explanatory variable = independent variable. We are trying to change this variable's values and then explain the other variable's behaviour. Usually correlative studies try to find causality, and that's why we need to explain the behaviour of one variable with the help of another.
- Explainable variable = dependent variable. We try explain this variable's behaviour.

1.5. Random sampling

Census study

In a census study we focus the study to the whole population. This is how we get results with no error. These results are easily generalized but getting them is usually really difficult time consuming. In several cases census studies are impossible to pull through, because we cannot reach all population.

This is why we need to take a sample. Sampling studies are easier to control, beacuse the material is more limited and the results are therefore faster to get. The problem with these is the generalization. When we calculate statistics from the sample we make a minor error concerning the population statistics. This is called the sampling error. This is why we inform the propability that the found difference is statistically significant.

Sample study

A sample is a smaller group chosen from the population, a subset. By the definition of a sample, the choosing must be based on propability. Aka. every member of the population must have equal chance to be chosen. If this criteria is not met, the result is a bogey. We end up to a bogey when the researcher has to use his/hers own judgement while choosing statistical units.

While choosing the size of the sample, we aim to make the sample as like the population as possible. This is called the presentability. So the sample is like a miniature from the population. This means that the statistics calculated from the sample diverge as little as possible from the corresponding population statistics. To get this, the sample must be chosen very carefully. There are different ways to form the sample, because we have different kind of studies. Below are listed the most common ways to form a sample.

Simple random sample

This is a very easy way to form a sample. It is the simplest and fastest way to form the sample if so called background variables don't matter (so it doesn't matter wether the chosen is a 20-year-old or 80-year-old etc.) The steps to a simple random sample:

1. You give ID numbers to each member of the population.
2. You decide the sample size.
3. You draw random numbers as many as the sample size tells you to.
4. You collect the members with the right ID numbers from the draw.

Systematic random sample

Systematic sampling is best to use in a line or in an assembly line. This is because the chosen ones are picked with a constant gap.

1. Arrange the population into a line.

2. Decide the size of the sample and calculate the gap = $\frac{\text{size of the population}}{\text{size of the sample}} = \frac{N}{n}$.
3. You draw the first statistical unit to be chosen to the sample.
4. From the first one you move according to the gap and choose the sample.

Stratified random sampling

If it's important to have for example specific amount of certaing ages to the sample, we have to make sure that the age distribution is correct. This is done with the help of stratified random sampling.

1. You divise the population to exclusive groups called strata based on the desired variable.
2. You decide the amount to be chosen to the sample separately from each strata.
3. You pick from each strata this amount of statistical units.

Cluster sampling

The researcher could for example be interested in how a student's writing skills develope during years 1-6 in school. This is very difficult if the researcher is forced to jump from school to school searching for individual students. It is more simple just to choose a school and study all it's students. This is called cluster sampling.

1. The population is divided into exclusive clusters. A cluster can be a postal area or a school etc.
2. Now the sampling is done in the population of clusters randomly.
3. Each statistical unit from a cluster is then chosen to the sample.

2. Statistics

2.1. Arithmetic mean

The arithmetic mean means the calculated centre of the material. This is the average value.

The arithmetic mean is calculated by adding up all the values and this sum is then divided with the amount of the values: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$.

Example 1

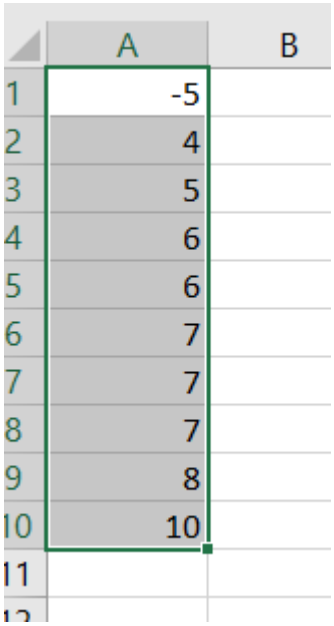
Here are the revenuepercents of 10 corporations in an increasing order: -5, 4, 5, 6, 6, 7, 7, 7, 8, 10. Lets calculate the mean.

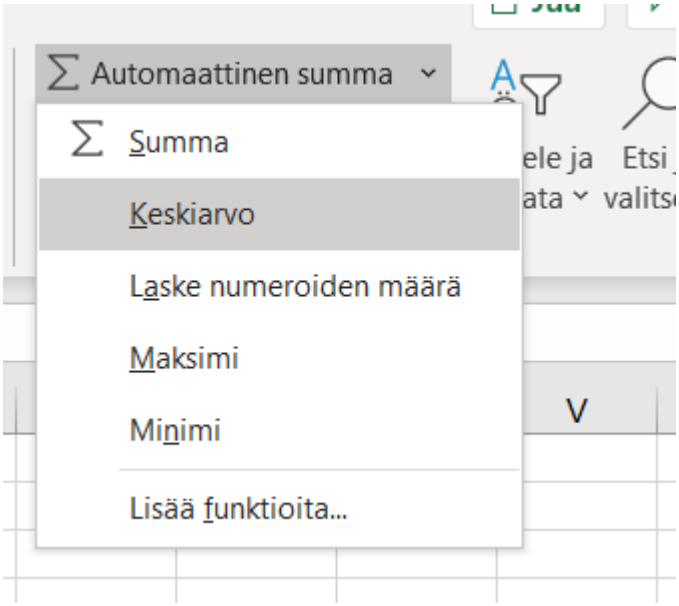
$$\overline{x} = \frac{-5+4+5+6+6+7+7+7+8+10}{10} = \frac{-5+4+5+2\cdot 6+3\cdot 7+8+10}{10} = \frac{55}{10} = 5{,}5$$

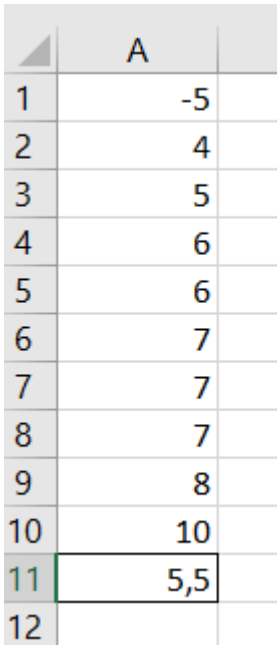
If the material is grouped, the calculation happens with the help of the group centres.

We can do the calculation also with the help of statistical programmes: OBS! In the pictures the excel is in finnish. For english pictures try google search.

Put the values to Excel one below the other. "Paint" the cells and choose the pull-down menu with the picture $\sum$. Choose here "mean".







OBS! you can also write it as a command:

After you have placed the values to cells, go to an empty cell and write =MEAN() and inside the brackets write first and last cell: =MEAN(A1:A10).

2.2. Weighted arithmetic mean

Sometimes all values in the material aren't as meaningful or significant as others. Now the arithmetic mean can give false image from the phenomenon because in the arithmetic mean every value has equal significance. In a weighted arithmetic mean each value has its own coefficient describing the significance of the value: the greater the coefficient, the more significant the value is. The weighted arithmetic mean is calculated by multiplying every value by its coefficient and adding these products together. This sum is then divided with the sum of the coefficients.

Example 2

The consumer price index is a way to study, how prices in general developes in Finland. It is calculated by grouping all buyable services and products and assigning each group it's own index (indicating this groups price development) and coefficient (how much peolple have bought this groups products). In the table below there is presented the groups and indexes and coefficients in february 2020.

February 2020	**Coefficient**	**Index**	**Coefficient·index**
Groceries and non alcohol drinks	138,71	103,3	14328,74
alcohol drinks and cigarettes	55,30	117,5	6497,75
Clothing and footwear	54,10	96,7	5231,47
Living, water, electricity, gas and other fuels	220,56	105,5	23269,08
Furniture, household machines and general taking care of household	57,98	99,4	5763,212
Health	50,16	108,5	5442,36
Traffic	134,95	102,9	13886,36
Communications	21,90	95,9	2100,21

Culture and free-time	122,78	98,5	12093,83
Education	4,62	107,1	494,802
Restaurants and hotels	72,28	109,3	7900,204
Other products and services	66,66	100,6	6705,996
Together	1000		103714

$$\overline{w} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i} = \frac{103714}{1000} = 103{,}714 \approx 103{,}7$$

2.3. Median

Median is the centre calculated by amount of the values. Median in considered to be the centre according to amount where the arithmetic mean is the centre according to magnitude. If the distribution is symmetrical according to the centre the median and mean are equal.From a discrete material we calculate the median by finding the value in the middle of the material when the material has been organized by the order of magnitude.

- If there are odd amount of the values, we take the one in the middle.
- If there are even amount of the values, we take the mean of the two middle ones.

Example 3

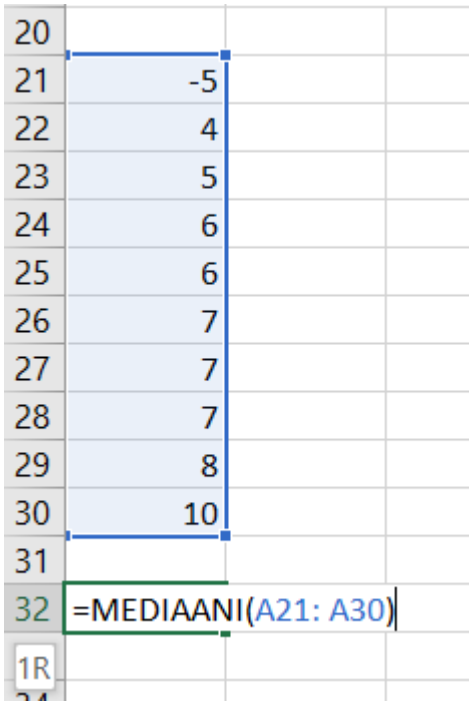
Here are the revenuepercents of 10 corporations in an increasing order: -5, 4, 5, 6, 6, 7, 7, 7, 8, 10.

Lets find the median.

Because there are 10 values, the median is the mean of the two values in the middle.
$Md = \frac{6+7}{2} = 6{,}5$.

If there are lots of values, a frequency table is preferable.

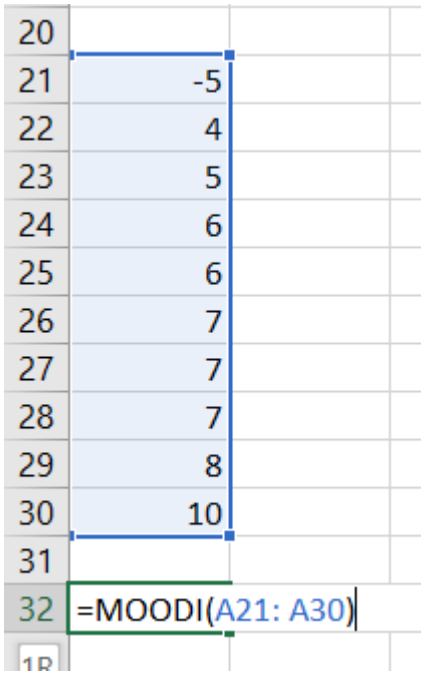
With Excel, the median can be found with the MEDIAN command. Write the material to cells and write to an empty cell =MEDIAN() and inside the brackets the cells as previously first cell : last cell.



2.4. Mode

The mode is the most typical value of the material. It is the class or the value with the greatest frequency. The mode isn't neccesarely unambiguous statistic because multiple values can have the same frequency.

With Excel write =MODE() and inside the brackets the material as previously.



2.5. Standard deviation

In statistics the standard deviation is a measure of variation. It describes the average variation from the arithmetic mean for the values. There are two formulas for the standard deviation and this is because wether you are calculating the whole population's deviation or

sampling deviation. If the material is claadified you'll have to use the class centre for calculations.

standard deviation of the population: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}} = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})}{n}}$

sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n-1}} = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})}{n-1}}$

In the statistical research, we usually calculate the sample standard deviation, because seldom we have census study. Still we know that these two statistics don't have significant difference, when the sample size is above 30, $n > 30$.

Example 4

Lets calculate the standard deviation of the material, when we know that the mean is 6,70.

x	f	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i (x_i - \overline{x})^2$
1	1	-5,7	32,49	32,49
2	2	-4,7	22,09	44,18
3	3	-3,7	13,69	41,07
4	1	-2,7	7,29	7,29
5	3	-1,7	2,89	8,67
6	6	-0,7	0,49	2,94
7	1	0,3	0,09	0,09
10	1	3,3	10,89	10,89
11	1	4,3	18,49	18,49
14	3	7,3	53,29	159,87
15	1	8,3	68,89	68,89
Yht:				394,87

$$s = \sqrt{\frac{394{,}87}{23-1}} \approx 4{,}24$$

With EXCEL we have two different commands based on, which one you are calculating: the sample deviation or population deviation.

=STDEV.P is the command to calculate the deviation for population and =STDEV.S is the command for the sample deviation.

2.6. Variance

Variance is standard deviation squared. It tries to tell the same story as the standard deviation but its unit is different than of the unit of mean or stardard deviation. Variance is calculated as the deviation but without the square root.

3. Probability

3.1. Defitions of probability

Propability is a measurement, which tells you the frequency of an event, aka. "how likely the event is". It took a long time to reach consensus over the definition of propability because it's human's own concept that has no scientific or philosofic foundations. Other parts of mathematics build on existing facts, meaning new laws and theorems are build on old ones. This wasn't the case with propability.

During time, people have always gambled. Gamblign relies on classical propability, which tells us that all possible outcomes are as equally likely to happen. This means that all outcomes are symmetric. So the propability to take an ace of hearts out of a card deck is as high as to take the two of spades. But ofcourse this definition was too simple for real world. All outcomes aren't always symmetrical. For example in Finland we knoe that the propability for having a girl child is around 48,7 %, which means that for the event "having a boy child" the propability is a little higher.

This is when we speak of statistical propability. Based on a statistical study we can determine the outcome's statistical propability by counting how many times the event occurs.

None of the proposed propability theories alone covered it all. Finally mathematician Andrei Kolmogorov put together the axioms of propability, that connects group theory,

propabilities and measure theory. Kolmogorov didn't explain propability but he solved the problems of different theories. Propability as we know it, is based on Kolmogorovs axioms.

The propability of the event A is always marked P(A), where P comes from the word propability. $0 <= P(A) \leq 1$. If P(A)=1, then A is certain to happen. If on the other hand P(A)=0, then A is impossible.

3.2. Mutually exclusive events

Mutually exclusive events:

- P(A or B) = P(A) + P(B)
- P(A and B) = 0

Example 1

Let us throw a regular dice once. What is the probability of having 2 or 3 as a result?

With a single throw $P(2) = \frac{1}{6}$ but also $P(3) = \frac{1}{6}$. Thus $P(2\ or\ 3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$.

Example 2

Let us throw a regular dice once. What is the probability of having 2 and 4 as a result?

Because it is not possible to have both 2 and 4 with the same throw, this is an impossible event. The probability is 0.

3.3. Not-mutually exclusive events

Not-mutually exclusive events:

- P(A or B) = P(A) + P(B) - P(A and B)
- P(A and B) =
    - $P(A) \cdot P(B)$, when events A and B are independet.
    - $(P(A) \cdot P(B|\ A)$, when evets A and B are not independent. (More from this on the next page.)

Example 3

Let us throw a regular dice twice. What is the probability of having first a 2 or secondly a 4?

P(1. dice is 2) = $\frac{1}{6}$ and P(2. dice is 4) = $\frac{1}{6}$. Now P(1. dice is 2 or second is 4) = $\frac{1}{6}+\frac{1}{6}-\frac{1}{36}=\frac{11}{36}$. This happens because the event "1. dice is 2" doesn't consider what the second dice is. It can be 4 or it isn't a 4. Also the event "2. dice is 4" doesn't consider what the first dice is. Now both events include the event "1. dice is 2 and the second is 4". This isn't any more valuable case than others so you cannot calculate it twice. This means we have to reduce once this events propability from the sum.

P(1. dice is 2 and a second 4) = $\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$.

Example 4

Lets lift a card out of a regular deck of cards (52 cards). What is the probability that both cards are aces when

a) the card that is lifted is returned to the deck,

b) the card that is lifted is nor returned to the deck?

a) $P\left(1.\ ace\ and\ 2.\ ace\right)=\frac{4}{52}\cdot\frac{4}{52}=\frac{1}{169}$

b) $P\left(1.\ ace\ and\ 2.\ ace\right)=\frac{4}{52}\cdot\frac{3}{51}=\frac{1}{221}$

3.4. Complement

The events A complement or opposite means "not A". This is usually marked $\overline{A}$. A and its complement are exclusive cases so $P\left(A\right)+P\left(\overline{A}\right)=1$, where $P\left(A\right)=1-P\left(\overline{A}\right)$.

Example 5

A regular dice is thrown three times. Find the probability of having at most two times number 2.

A="having at most two times number two". Now with three throws the number of number two's is 0, 1, 2 or 3. At most two number 2 means 0, 1 or 2. It is now easier to calculate the complement: $\overline{A}$ = "each throw is a number two".

$P(A) = 1 - P\left(\overline{A}\right) = 1 - \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{215}{216}$.

4. Independence and conditional probability

4.1. Independence

In the last chapter the rule for P(A and B) had two separate forms depending on the fact wether the events A and B are independent or not. The definition of independence is:

The events A and B are independent when $P(A \; and \; B) = P(A) \cdot P(B)$.

This means that exclusive cases are almost always dependent. When A and B are exclusive and $P(A) > 0$ and $P(B) > 0$, then $P(A) \cdot P(B) > 0$ but $P(A \; ja \; B) = 0$. Thus exclusive cases are almost always dependent.

For non-exclusive cases the independence must be based on the definition and looked individually if it's not mentioned in the task.

Example 1

You sign up to the yearly "HeinäHaukku"-event for dogs for saturday's show, sunday's show or both. One year 1372 dogs were signed up, from which 31 were signed only for saturday, 43 only for sunday. Let L be the event "the dog was signed for saturday" and S the event "dog was signed for sunday".

a) Find P(L and S) that year.

b) Are L and S independent that year?

a) If the amount of dogs signed for both days is x, then

$31 + 43 + x = 1372 \Rightarrow x = 1372 - 31 - 43 = 1298$

So P(L and S) $= \frac{1298}{1372} \approx 0{,}94606$.

b) Now $P(L) = \frac{31 + 1298}{1372} = \frac{1329}{1372}$

$P(S) = \frac{43 + 1298}{1372} = \frac{1341}{1372}$

$P(L) \cdot P(S) = \frac{1329}{1372} \cdot \frac{1341}{1372} = \frac{1782189}{1372^2} \approx 0{,}94677$.

Because $P(L) \cdot P(S) \neq P(L \; and \; S)$ so events aren't independent.

4.2. Conditional probability

For independent events always states $P(A \; ja \; B) = P(A) \cdot P(B)$. But when the events aren't independent, we get $P(A \; and \; B) = P(A) \cdot P(B \mid A)$.

From this we can solve $P(B \mid A)$, which is called conditional propability:

$P(B \mid A) = \frac{P(A \; and \; B)}{P(A)}$.

So the event B's propability is looked from the point of view of the event A. If A happens, what is the propability for B to also happen?

Example 2

There are white and red cubes in two boxes. In the box number 1 there are 3 red and 2 white ones and in the box number 2 there are 4 red ones and 1 white in. We choose first randomly one of the boxes and then we take one cube from that box. The chosen cube is white. What is the propability that we chose the box number 2?

Even though as we choose the box, the propability for box number 2 is $\frac{1}{2}$, we must look this regarding the given fact "we chose a white cube". The propability in question is P(we choose box number 2| we choose a white cube). So we must look at all the cases where we choose a white cube.

P(white cube)$= \frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{1}{5} = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}$.

Now the favorable events are those where we choose first the box number 2.

P(We choose box number 2 and we choose a white cube)$= \frac{1}{2} \cdot \frac{1}{5} = \frac{1}{10}$.

Now the propability in question is P(we choose box number 2 | we choose a white cube) $= \frac{\frac{3}{10}}{\frac{1}{10}} = \frac{1}{3}$.

5. Discrete random distributions

5.1. Random distribution

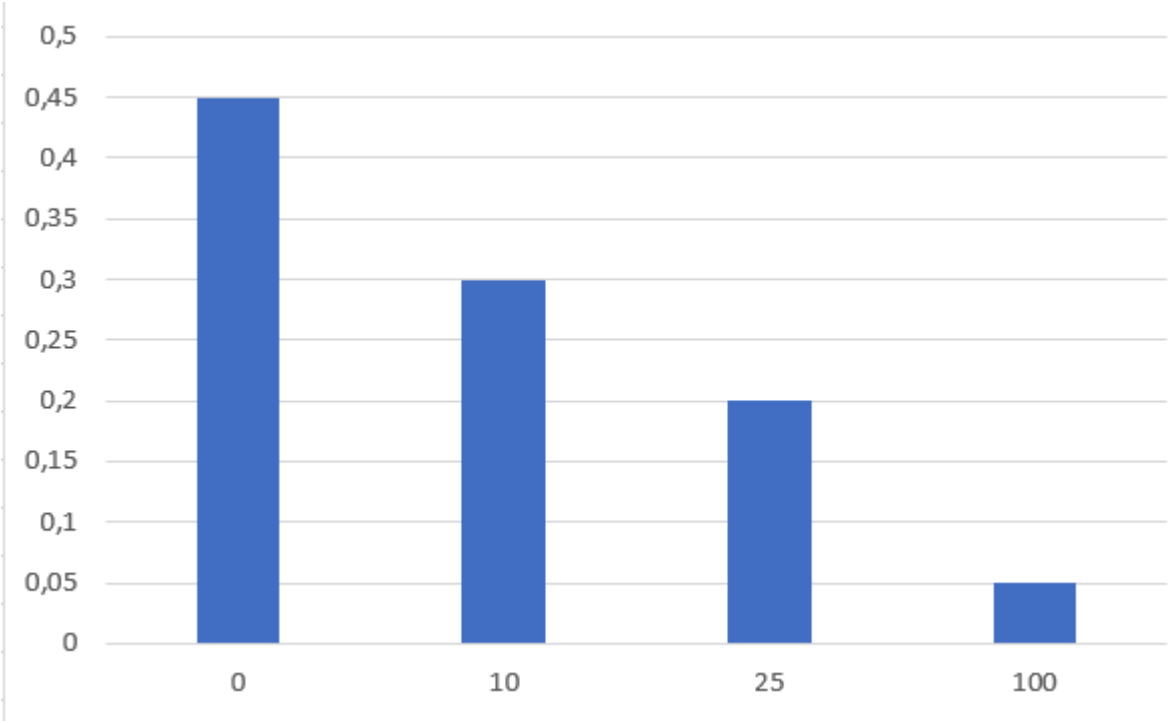
In a random phenomenon chance determines the outcome. Random variable is a variable which values are determined by chance. If we want to know the distribution of a random varible, we have to know all the values and their probabilities. This is called the density function of the variable or in the case of a discrete variable, we can call it the propability function.

Example 1

The wins in a lottery are 100 €, 25 € and 10 €. We know that 5 % of the lotterytickets hold 100 € wins, 20 % have 25 € wins and 30 % have the 10 € win. The rest have no wins. A person buys one ticket. Let X be the random variable X="the amount of the win". This variable's distribution is:

Win	Propability
100 €	0,05
25 €	0,20
10 €	0,30
0 €	0,45

The distribution can be presented as a table, as we did in example 1. On th other hand we can also present it as a graph or mathematically:
$P\left(X=100\right)=0{,}05 \qquad P\left(X=25\right)=0{,}20 \qquad P\left(X=10\right)=0{,}30 \qquad P\left(X=0\right)=0{,}45$



The cumulative distribution function tells us the probability that the value is at most x.

$F(x) = P(X \leq x)$.

Example 2

Lets find the cumulative distribution of the situation in example 1.

x	p	F
0 €	0,45	0,45
10 €	0,30	0,75
25 €	0,20	0,95
100 €	0,05	1,00

5.2. Statistics of a random variable

The statistics of a random variable are there to describe the distribution. Corresponding term for the mean is the expected value. This is usually marked with the letter E or μ.

$$EX = \mu = \sum_{i=1}^{k} p_i x_i$$

Example 3

What should be the price of the ticket in example 1, if we wish it to be the same as the variable's X expected value?

The expected value: $EX = 100 \cdot 0{,}05 + 25 \cdot 0{,}20 + 10 \cdot 0{,}30 + 0 \cdot 0{,}45 = 13$.

The price should be 13 €.

The standard deviation tells the average distribution of the values from the expected value.

$$\sigma = \sqrt{\sum_{i=1}^{k} p_i \left(x_i - \overline{x}\right)^2}.$$

Example 4

Lets find the variable's X standard deviation in example 1.

x	p	$x - \overline{x}$	$(x - \overline{x})^2$	$p\,(x - \overline{x})^2$
100	0,05	87	7569	378,45
25	0,20	12	144	28,8
10	0,30	-3	9	2,7
0	0,45	-13	169	76,05
Total				486

$\sigma = \sqrt{486} \approx 22{,}05$.

5.3. Binomial distribution

Some discrete distributions have specific features and that's why they have special names.

We throw a regular dice 5 times. Find the propability to get the number 6 twice.

Throwing a dice is so called repeated trial. We repeat the same situation over and over again and each repetition is independent from the previous ones aka. the propability stays the same.

- If we have n repetitions and
- the propability of the favourable event is p,

then the propability of having k times the favourable event is
$P\,(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$

Example 5

We throw a regular dice 5 times. What is the propability to have the number six twice?

Let X="the amount of number 6". Now

$$P(X=2) = \binom{5}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^3 = \frac{625}{3888} \approx 0{,}161.$$

The random variable in example 5 is a typical example of a variable that obeys the binomial distribution. The specific features are as follows:

- we have a repeated trial with n repetitions,
- the repetitions are independent, aka. the propabilities for the favourable event and non-favourable event stays the same.

When a random variable obeys the binomial distribution, we say it: $X \sim Bin(n,p)$. The letters n and p are the parametres, which defines the shape of the distribution. The variable in example 5 obeys the binomial distribution: $X \sim Bin(5,\frac{1}{6})$.

The statistics of the binomial distribution are

- expected value $EX = \mu = np$,
- standard deviation $DX = \sigma = \sqrt{npq}$.

Example 6

Find the expected value and the standard deviation of the random variable X in example 5.

$$EX = np = 5 \cdot \frac{1}{6} = \frac{5}{6}$$

$$DX = \sqrt{npq} = \sqrt{5 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \sqrt{\frac{25}{36}} = \frac{5}{6}$$

OBS! The binomial distribution in EXCEL works with the command =BINOM.DIST(k; n; p; cumulative). The parametres in the brackets are:

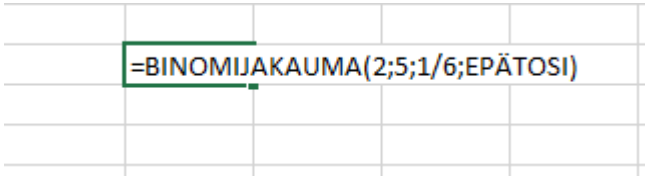
1) k, the amount of favourable trials

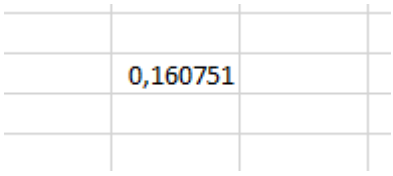
2) n, the amout of all trials

3) p, the propability od the favourable trial

4) TRUE or FALSE, wether you want to have the cumulative distribution's propability or not. TRUE means that you want the cumulative propability and FALSE will give you just the propability of this single event.

The situation in example 5 calculated with finnish EXCEL.





5.4. Poisson distribution

When we have the amount of a certain event within a fixed time range, we have a random variable that obeys the Poisson distribution. For example the amount of insurance payments done by the insurance company within a week is a discrete variable in a time range and therefore obeys the Poisson distribution. The parametre for the Poisson distribution is the expected value $X \sim Poisson(\lambda)$. What is the letter for the expected value depends on the book you're reading. In many finnish books the letter for Poisson distribution's expected value is λ.

The propability function of the Poisson distribution is $P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$ and the properties of the distribution are

- $D^2X = \sigma^2 = \lambda$
- $DX = \sigma = \sqrt{\lambda}$

Example 7

A insurance company pays 4 large compensations in a month. Find the propability that this company has to pay

a) zero,

b) at least 5 large compensations during the next month.

Let X="the amount of large compensations the company has to pay during a month".

Now $X \sim Poisson(4)$

a) $P(X=0) = \frac{4^0}{0!} \cdot e^{-4} = e^{-4} \approx 0{,}0183$

b)

$P(X \geq 5) = 1 - P(X < 5) = 1 - P(X \leq 4) = 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$

$= 1 - \left[\frac{4^0}{0!} \cdot e^{-4} + \frac{4^1}{1!} \cdot e^{-4} + \frac{4^2}{2!} \cdot e^{-4} + \frac{4^3}{3!} \cdot e^{-4} + \frac{4^4}{4!} \cdot e^{-4}\right]$

$= 1 - \frac{103}{3} \cdot e^{-4} \approx 0{,}371$

OBS! The Poisson distribution in EXCEL works as the binomial distribution.

=POISSON.DIST(x; mean; cumulative)

1) x, the amount of wanted events

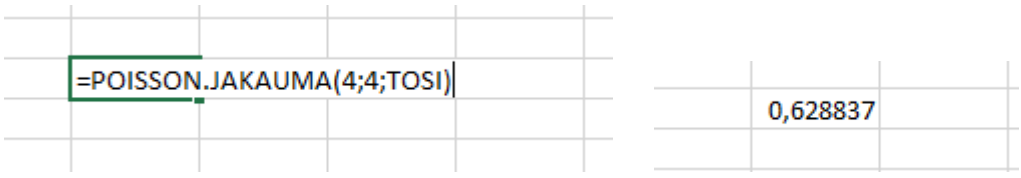
2) mean, the expected value

3) TRUE/FALSE, regarding wether you wish to count the cumulative propability or not.

The situation in example 7 done with EXCEL:

a)

b)



so the propability in question is the complement of the given value:

6. Continuous random distributions

6.1. Continuous distributions

A continuous varible can be age, time, hight etc. These kind of variables get all possible real values within a certain range (for example x>0). So the density function cannot be represented with tables or with single probabilities.

The density function of a continuous variable is $f\left(x\right)$ when

- $f\left(x\right) \geq 0,$ with all possible values of x,
- the area limited between$f\left(x\right)$ and the x-axis is 1.

So $P\left(a \leq X \leq b\right)$ is equal to the area limited between the density function of X and the x-axis when x is in range $\left[a,\ b\right]$.

Example 1

A random variable X is defined within the range $\left[0,\ 1\right]$, and its density function is $f\left(x\right) = \frac{x}{a} + \frac{a}{2}$. Find a. Find also the probability to X being within the range of $\left[0,\ \frac{1}{2}\right]$?

We know that the density function limites a area with the x-axis that has to be 1:

$$\int_0^1 \left(\frac{x}{a} + \frac{a}{2}\right) dx = 1$$

$$\int_0^1 \left(\frac{x}{a} + \frac{a}{2}\right) dx = \Big/_0^1 \left(\frac{x^2}{2a} + \frac{ax}{2}\right) = \frac{1}{2a} + \frac{a}{2}$$

$$\frac{1}{2a} + \frac{a}{2} = 1 \,||\cdot 2a$$

$$1 + a^2 = 2a \Rightarrow a^2 - 2a + 1 = 0$$

$$\left(a - 1\right)^2 = 0 \Rightarrow a - 1 = 0 \Rightarrow a = 1$$

So $a = 1$. With this value of a, the function is $f\left(x\right) = x + \frac{1}{2}$. Now the probability in question is

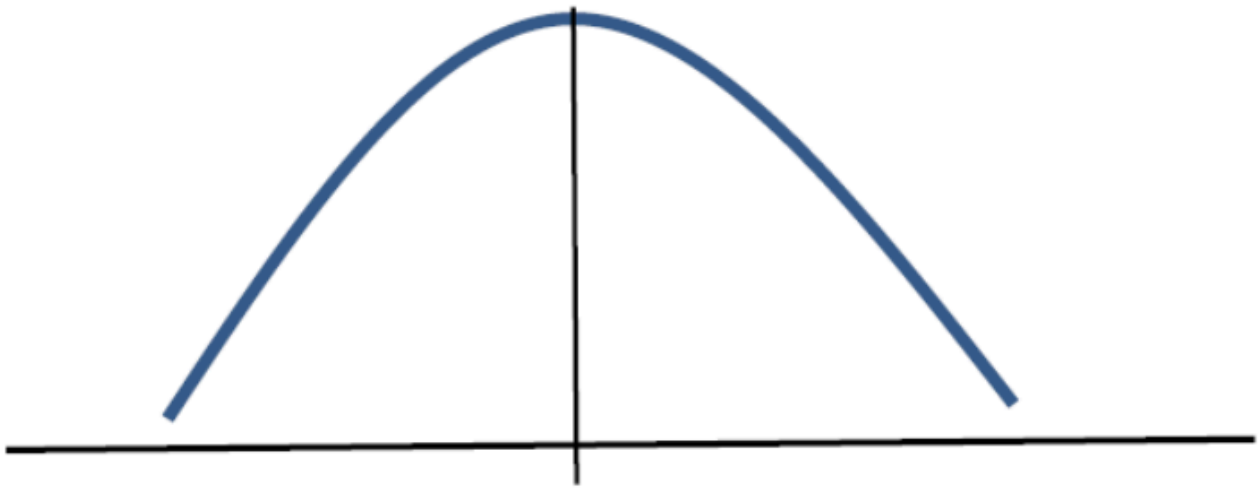
$$P\left(0 \leq X \leq \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \left(x + \frac{1}{2}\right) dx = \Big/_0^{\frac{1}{2}} \left(\frac{1}{2}x^2 + \frac{1}{2}x\right) = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{4} = \frac{1}{8} + \frac{2}{8} = \frac{3}{8}.$$

The cumulative probability function F(x) tells us the probability that the value of X is at most x. So $P\left(X \leq x\right) = F\left(x\right)$.

6.2. Standard normal distribution

The most common of the continuous distributions is the normal distribution. This is still used in Finland to score the final exams of the upper secondary school (high school). The form of the distribution is determined by the expected value and the standard deviation (some materials use also variance instead of the deviation). If the variable X follows the normal distribution with the expected value of μ and the standard deviation of σ, we write $X \sim N(\mu,\sigma)$.

The normal distribution is symmetrical according to the expected value. The expected value defines the vertex of the distribution curve. If the standard deviation is small, the values are found tight around the expected value. If the standard deviation is large, the values are more spreaded around the x-axis.



The density function and the cumulative function of the normal distribution are usually presented in matematical books. If you wish, you can find them even online. But knowing them is not really essential.

The values of the cumulative distribution are found in a table for such distribution with expected value of 0 and deviation of 1. This is called the standard normal distribution. The cumulative function of the standard normal distribution is marked with Φ in stead of F. From the table you can read the cumulative probabilities $P(Z \leq z) = \Phi(z)$.

Example 2

Let us persume that $X \sim N(0,1)$. Find $P(X \leq 1{,}25)$ and $P(X \leq -1{,}25)$.

$P(X \leq 1{,}25) = \Phi(1{,}25)$. This can be read from the table so that from the first collumn we find 1,2 and from the first row we find 0,05. Now the probability is found from the intersect of these (collumn and row). $P(X \leq 1{,}25) = \Phi(1{,}25) = 0{,}8944$.

For finding $P(X \leq -1{,}25)$ we have to use the fact that the normal distribution is symmetrical. $P(X \leq -1{,}25) = P(X \geq 1{,}25)$, meaning "what is on the right side of 1,25 is equal to the left side of -1,25.
$P(X \leq -1{,}25) = P(X \geq 1{,}25) = 1 - \Phi(1{,}25) = 1 - 0{,}8944 = 0{,}1056$.

6.3. General normal distribution

Seldom we find a varible which follows the standard normal distribution. For example the hight of a person is a continuous random variable, which has the expected value closer to the value of 170 than of 0. However we can standardize each random variable (following some normal distribution) to correspond the standard normal distribution. Lets say we wish to find the probability that a randomly chosen person has the hight of at most 160 cm. First we must find the exact same spot from the standard normal distribution (limits the same area). We mark this spot as z: $z = \frac{x - \mu}{\sigma}$.

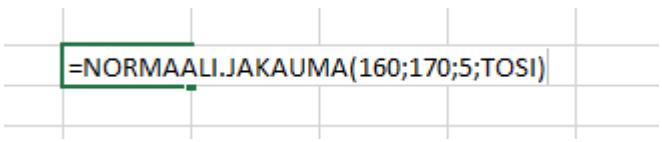
Example 3

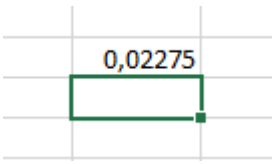
Let the hight of a person follow a normal distribution with the following parametres: $\mu = 170$ and $\sigma = 5$. Find the probability that the hight of a randomly chosen person is at most 160 cm.

$P(X \leq 160) = P\left(Z \leq \frac{160 - 170}{5}\right) = P(Z \leq -2) = \Phi(-2) = 1 - \Phi(2) = 1 - 0{,}9772 = 0{,}0228$.

So the probability is around 2,3 %.

OBS! The normal distribution is EXCEL works as the binomial and Poisson distribution in last chapter.





7. Correlation and regression

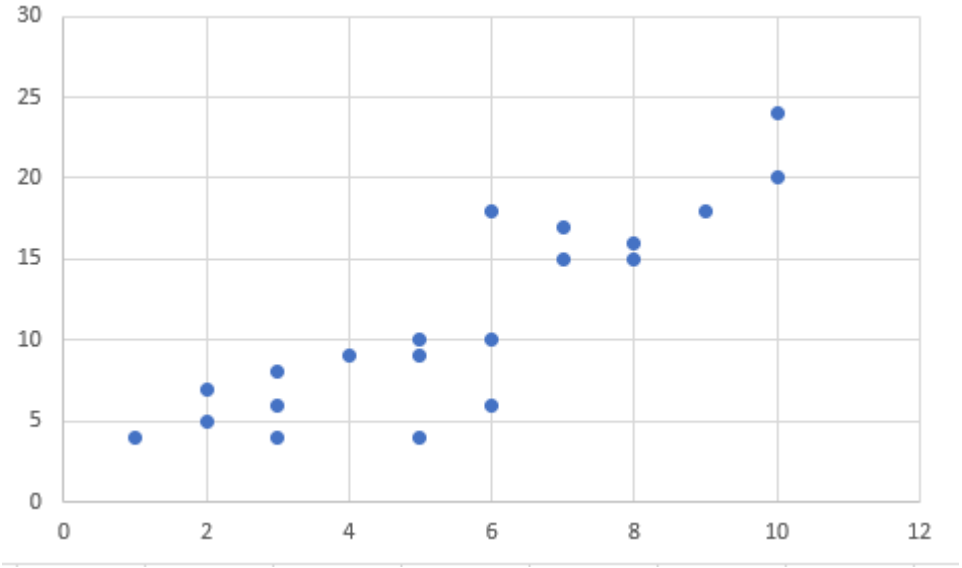
7.1. Connection between variables

In statistical analysis we are usually interested in the fact wether the two variables in question have a connection, and if so, what is the type of the connection. This can be examined by a scatter diagram. We choose one of the variables to be on the x-axis and the other to the y-axis. Then we place every statistical units anwers to the coordinate system. The connection and it's type can be found from the picture.

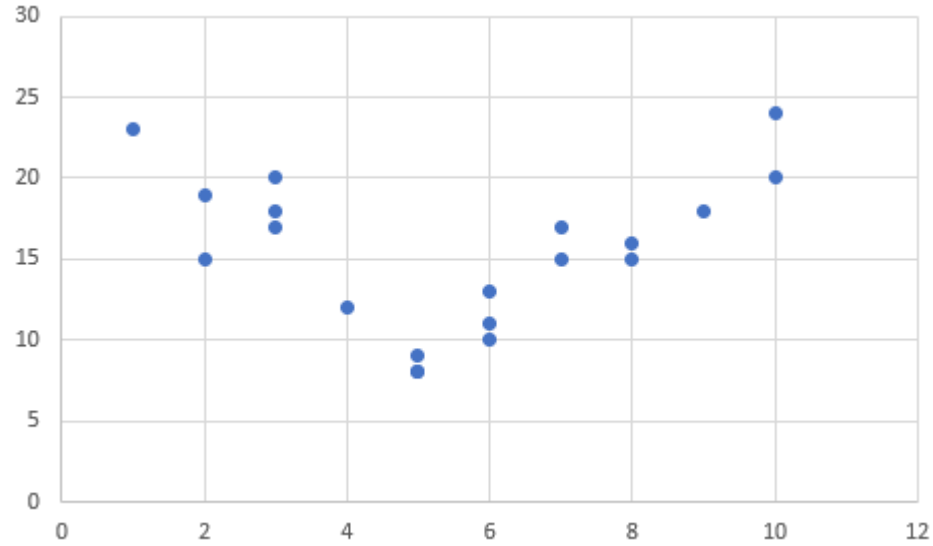
- If the dots in the scatter diagram from a regular mathematical graph, there is a connection.
- The type of the connection is defined according to the mathematical graph.

Different types are presented down below:

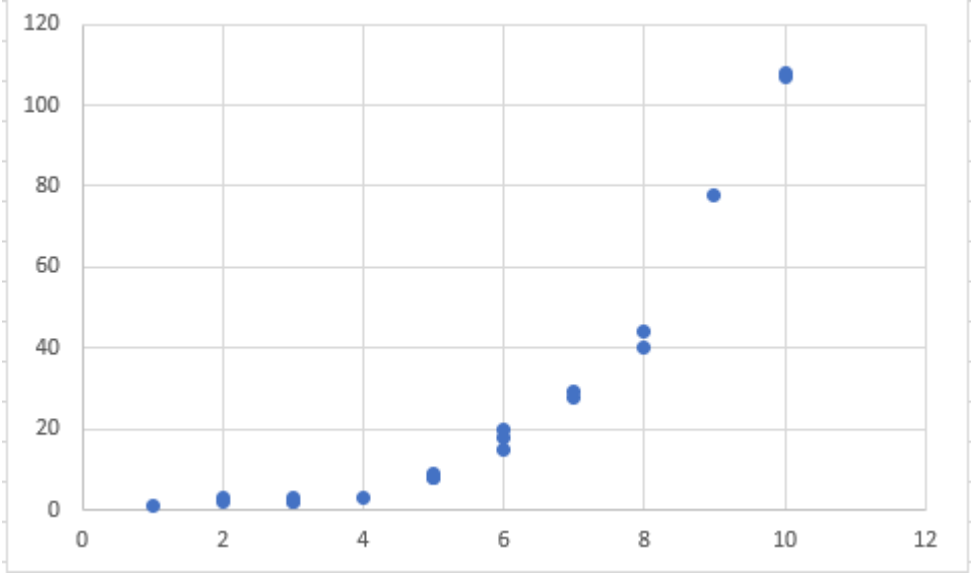
Linear connection



Polynomial connection



Exponential connection



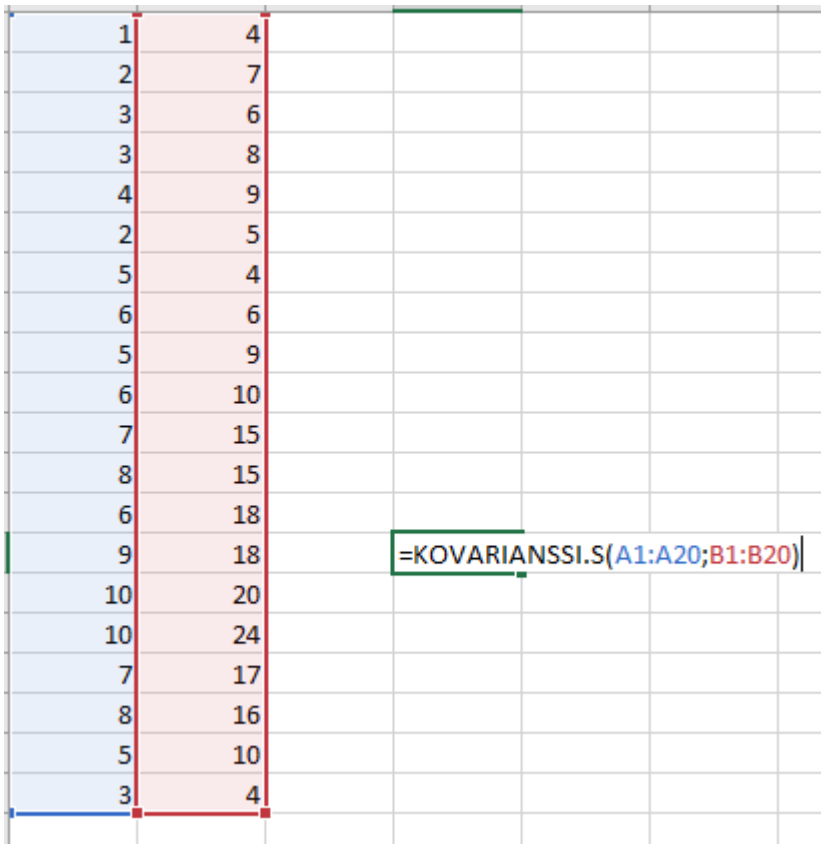
So we have to consider each diagram separately. We have to look the diagram and consider what kind of mathematical graph fits into it. We still have to consider that all the points will never fit perfectly to the graph. The will differ from it slightly. The graph will describe the average situation.

7.2. Correlation coefficient

How strong the connection is? Can there be a statistic to show how strong the connection between two variables is? Actually there have been several. We have begun from the linear connection. First statistic to describe the linear connection was the covariance. This is positive, if the linear graph is a rising line and negative if the graph is a decending line. If the covariance is around zero, then there is no linear connection between these variables.

The covariance can be calculated $\sigma_{xy} = \frac{\sum_{i=1}^{n} [(x_i - \overline{x})\,(y_i - \overline{y})]}{n-1}$.

Using EXCEL this can be done =COVARIANCE.S(), where the material is placed inside the brackets: x-variable; y-variable.



The problem with the covariance is that it's not comparable with other covariances calculated from different materials. The covariance is effected by the unit of measure and the order of magnitude. But the aim was to have a statistic that is comparable. This is the correlation coefficient.

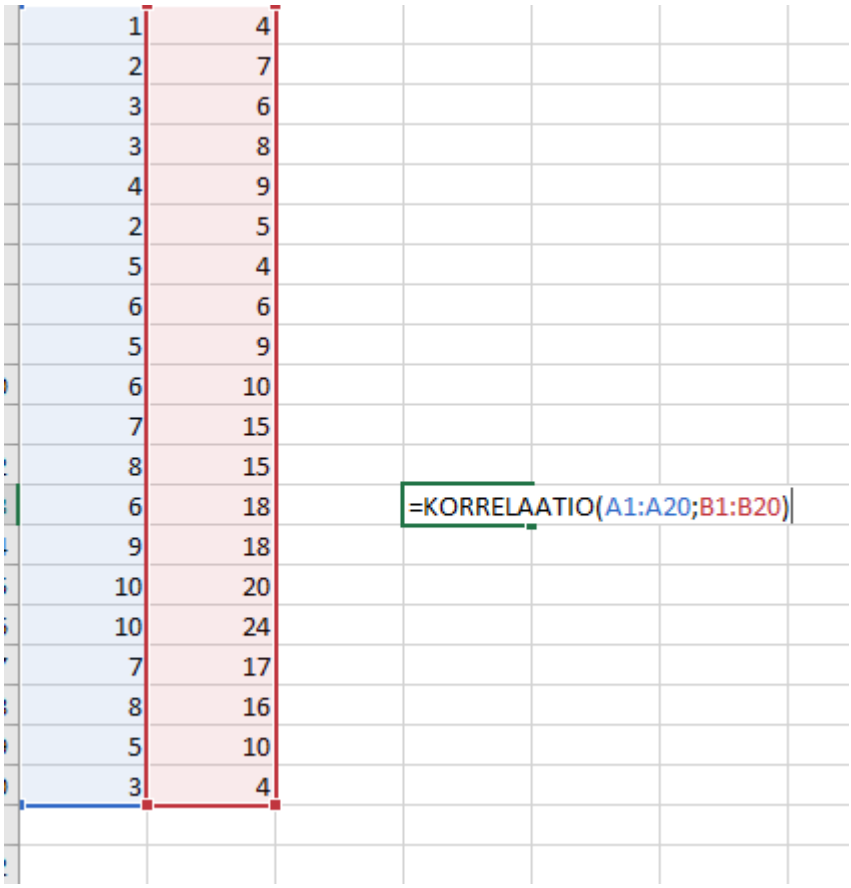
The correlation coefficient is like the covariance but it is standardiced to the range [-1, 1]. Wether the correlation coefficient is positive, the line is rising and negative correlation coefficient indicates a decending line. If the coefficient is around zero, then there is no linear connection between the two variables. OBS! If the coefficient is around zero, there is no LINEAR connection. This does not mean that there is no connection whatsoever. The type can be some other kind. This must be looked from the scatter diagram.

The correlation coefficient can be calculated by dividing the covariance with the product of the standard deviations:

$$r_{xy} = \frac{\sigma_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x}) (y_i - \overline{y}) \right]}{(n-1) s_x s_y} = \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x}) (y_i - \overline{y}) \right]}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2 \right) \cdot \left(\sum_{i=1}^{n} (y_i - \overline{y})^2 \right)}} = \frac{n \cdot \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right)}}$$

.

With EXCEL:



We can classify the correlation coefficients according to its value.

- The connection is meaningless if $|r|$ is within the range 0 - 0,3.
- The connection is moderate, when $|r|$ is within the range 0,3 - 0,5.
- The connection is mediocre, when $|r|$ is within the range 0,5 - 0,7.
- The connection is strong, when $|r|$ is within the range 0,7 - 0,9.
- The connection is very strong, when $|r|$ is within the range 0,9 - 1.
- The connection is perfect, when $|r|$ is 1.

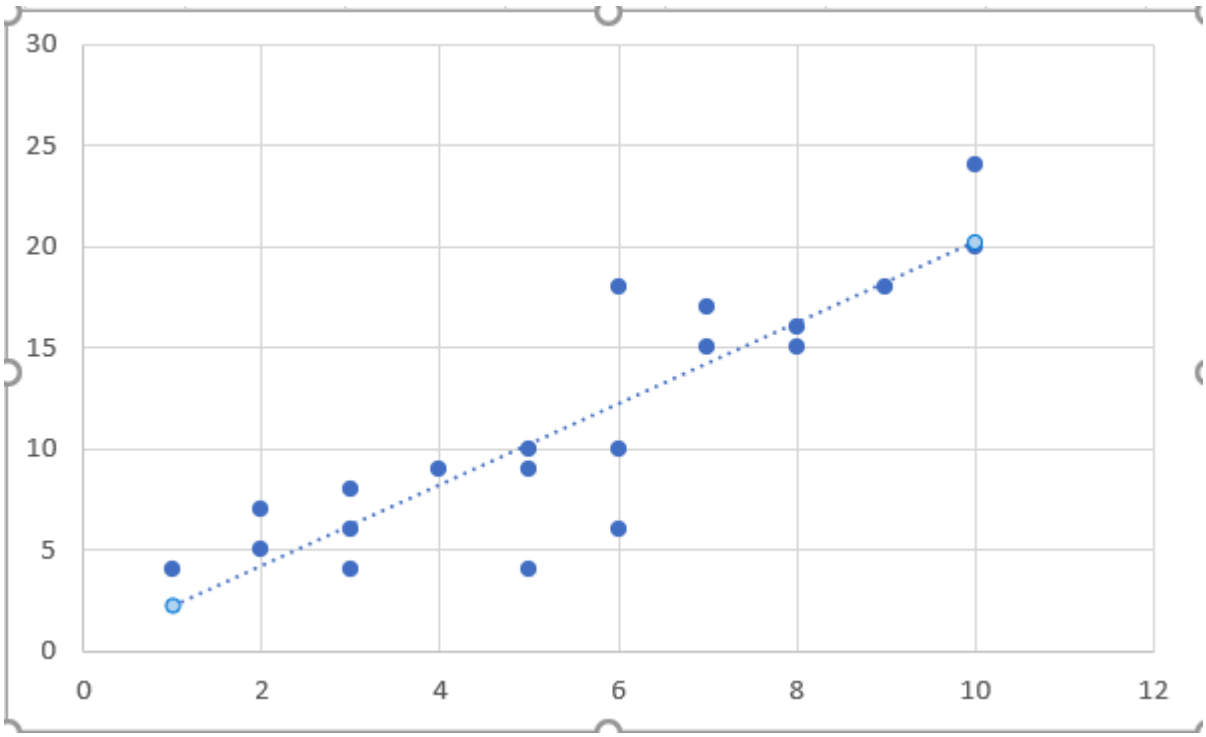
7.3. Pearson's product-moment coefficient

Usually when we speak of the correlation coefficient, we mean the parametric correlation coefficient, Pearson's product-moment coefficient. This statistic describes the linear connection of two parametric variables. In the last example (previous page) the EXCEL gives us Pearson's product-moment coefficient.

The parametric coefficient means that the variables have to obey enough the normal distribution and their scale have to be interval or ratio level. If either one of these conditions are not met, Pearson's product-moment coefficient cannot be calculated. Instead we must calculate Spearman's rank correlation coefficient which is explained on the next page.

The parametric correlation coefficient is very sensitive for divergent observations, because a single divergent observation moves the average line. Pearson's product-moment coefficient

explains how well the line fits to the scatter diagram. The picture from the example on the previous page is below:



7.4. Spearman's rank correlation coefficient

When atleast one of the variable's scale is ordinal level or the variable doesn't obey the normal distribution, we have to calculate the non-parametric coefficient, Spearman's rank correlation coefficient. Calculating this requires only that the scale of the variable's has to be atleast ordinal level. When we calculate this coefficient, the values are replaced with the ordinal number according to the values. In the case of even values, we give the ordinal number as a mean of real ordinal numbers.

Usually the coefficient can be calculated with the formula $r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$, where d_i is the difference of the ordinal numbers of x and y variables.

EXCEL doesn't have ready function for calculating this but we can go round this. We just have to form the ordinal numbers.

x	y	järjestys x	järjestys y
1	4	1	2
2	7	2,5	6
3	6	5	4,5
3	8	5	5
4	9	7	5,5
2	5	2,5	3
5	4	9	1,5
6	6	12	2
5	9	9	2
6	10	12	2,5
7	15	14,5	3,5
8	15	16,5	3
6	18	12	5,5
9	18	18	5
10	20	19,5	5
10	24	19,5	5
7	17	14,5	4
8	16	16,5	3
5	10	9	2
3	4	5	1

This can be done with RANK.AVG function:

=RANK.AVG(number; ref; [order])

- number = the cell whose rank you wish to know
- ref = are all the cells in which the rank is calculated
- [order] = optional, values 0 or 1
    - 0, if you wish to arrange the order from larger to smaller
    - 1, if you wish to arrange the order from smaller to larger.

With EXCEL you can define all the ranks at once. In the example the variable in the column A has ranks in column C:

You write to the cell C2:

=RANK.AVG(A2; $A$2:$A$21; 1)

Now press ENTER and you get to the cell C2 the cells A2 rank in the material written in the cells A2-A21. Now clicking the cell C2 and pulling from the right downcorner to the cell C21 you all the ranks for the cells A2-A21. $-marks around cells means that this is to be held constant even though the first written cells alters.

Now =CORRELATION(C2:C21; D2:D21) you can calculate the Spearman's rank correlation coefficient.

7.5. Monotonical non-linear connections

The correlation coefficient tells us usually the strenght of the linear connection between the two variables. But the Spearman's rank correlation coefficient gives us connections between such variables, that have non-linear connections but are strictly monotonic.

A strictly monotonic connection is such a connection, which is rising all the way or decending all the way. This rising or decending doesn't have to happen with a constant speed.

A good way to notice that the connection is a nonlinear connection, is to find the Spearman's rank correlation coefficient to be much larger in an absolute value than the Pearson's product-moment coefficient's absolute value. When the connection is linear, these two are almost equal.

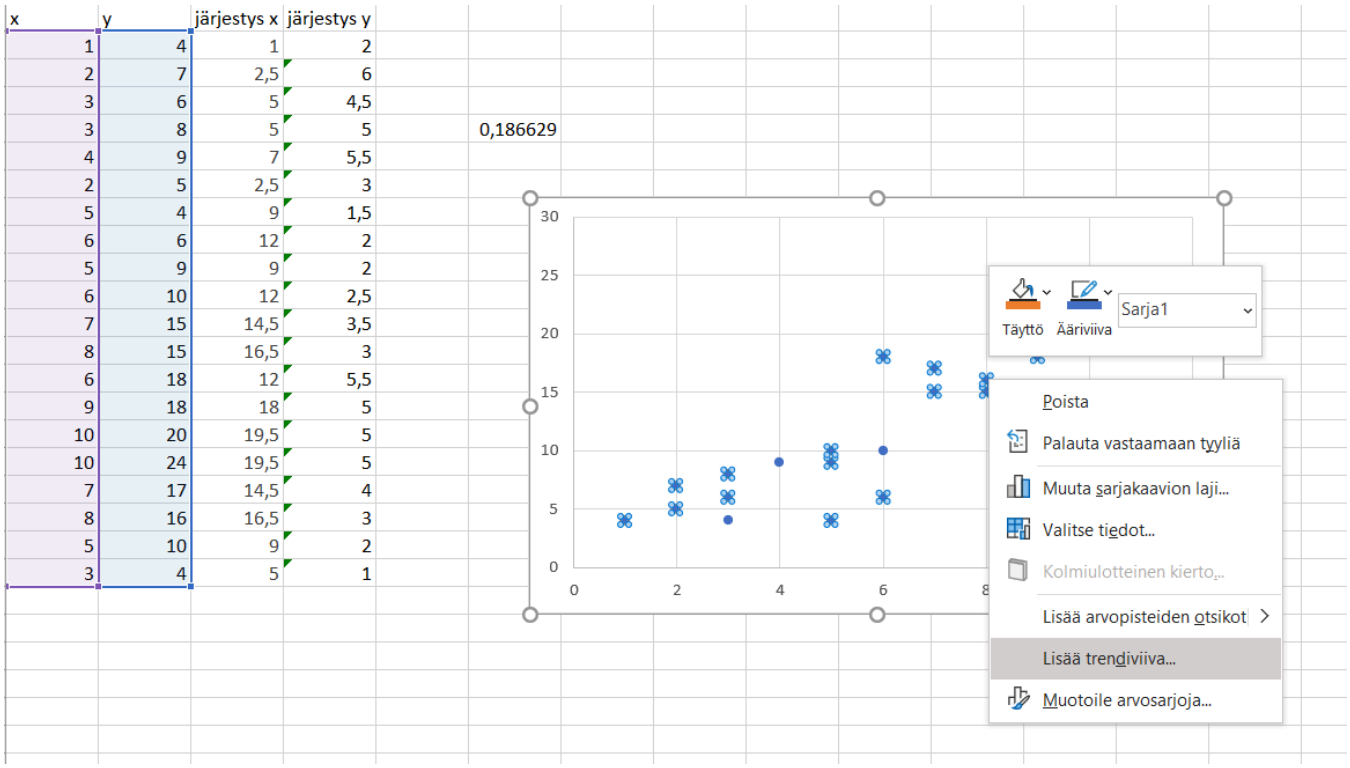
7.6. Regression

The correlation coefficient measures the strength of the linear connection between two variables. This does not however form any mathematical model between these variables. The correlation coefficient does not take part in causality. It it the same if the x and y variables are vice versa.

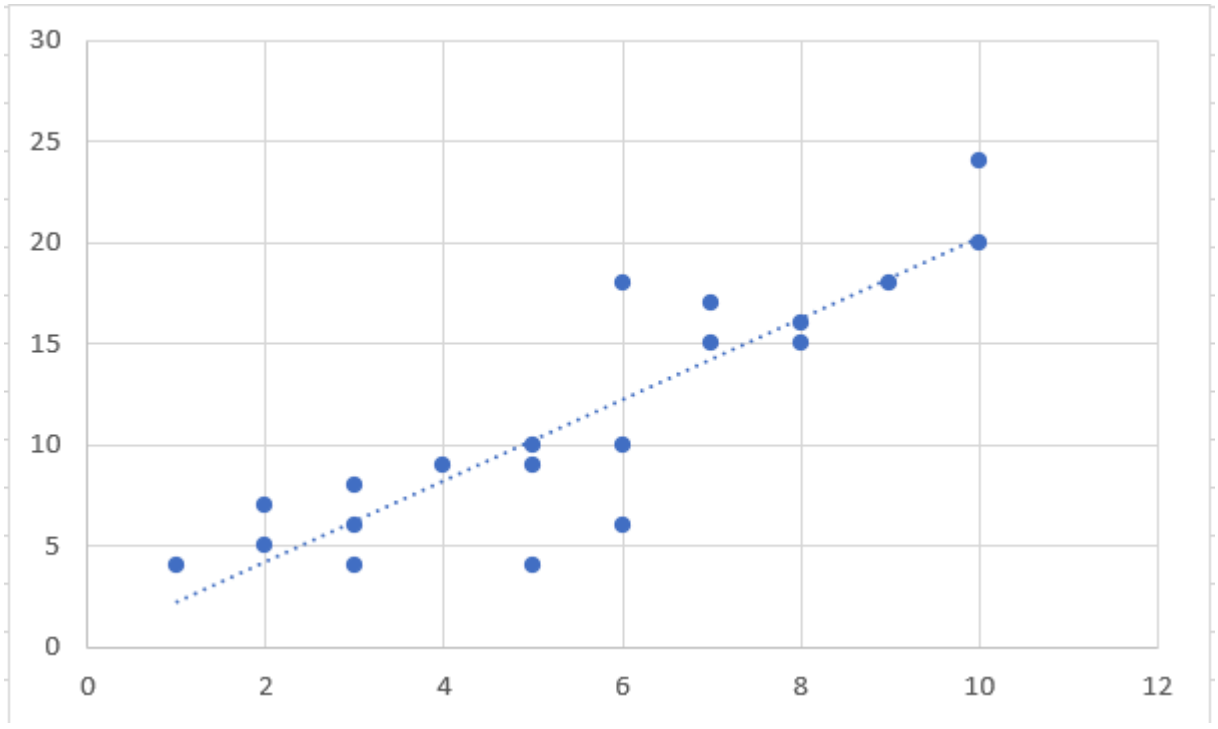
If we want to find the causality, we must do a correlative study. In this study we alter the x variables values and we measure it's effects on the y variable. This way we can find the connection between these variables and create a "trendcurve" into the scatter diagram. This curve can be a line, a parabola etc. In general this process is called regression analysis. With the help on this analysis we can find a mathematical model to describe the connection and with the help of this model we can forecast things by counting. For example a icecream sales person can calculate the estimated amount of icecream needed for the shop next week. This means less icecream to be thrown away.

In the most simple models we try to explain the behaviour of the y variable with the help of one x variable. The y variable is called the explainable variable or the dependent variable, whereas the x variable is the Explanatory variable or the independent variable. The trendcurve is easiest to form, when it's a line. With computer programs, this is done automatically.

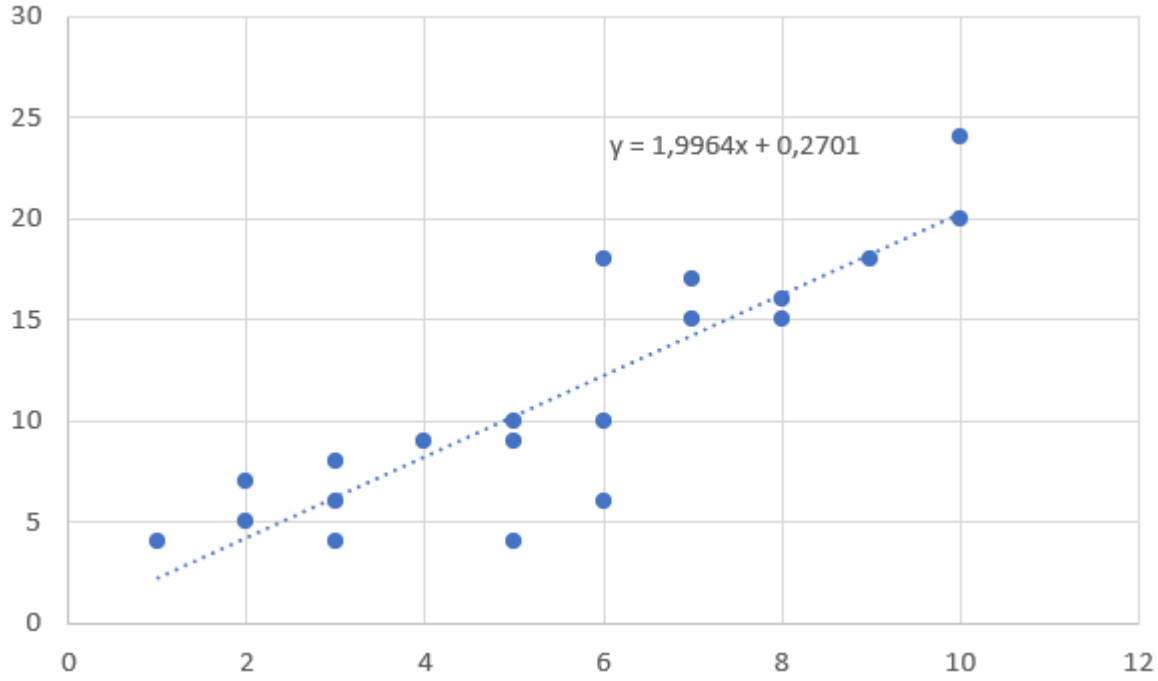
Previously we calculated the correlation coefficient between variables x and y. Now we want to fit a line to the scatter diagram.



By painting the cells in use, we choose the scatter diagram from the menu. No click a single point with the right button of the mouse and from that menu, click "add trendline".



From the menu on the right you can choose the type of the curve and cross the box "show the equation".



Now the variable y depends on the variable x according to the equation $y = 1{,}9964x + 0{,}2701$. We can forecast now with the help of this equation.

Example 1

Use the equation above and calculate

a) y when x is 7.

b) x when y is 12.

a) $y = 1{,}9964 \cdot 7 + 0{,}2701 = 14{,}2449 \approx 14$

b) $12 = 1{,}9964 \cdot x + 0{,}2701$
$11{,}7299 = 1{,}9964x \;||:1{,}9964$
$x = 5{,}8755... \approx 6$

You can form the coefficients of the regressionline yourself. When the line's equation is in the form $y = b_0 + b_1 x$, then

$$b_1 = \frac{n \cdot \sum_{i=1}^{n}(x_i y_i) - (\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i)}{n \cdot (\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

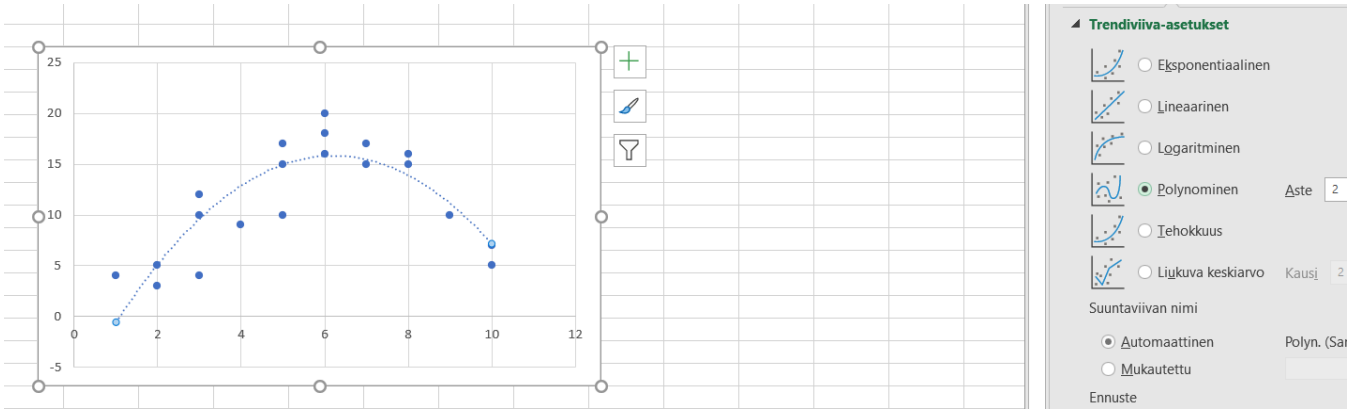
$$b_0 = \overline{y} - b_1 \overline{x}$$

From here, we get that the line always goes through the mean point $(\overline{x}, \overline{y})$.

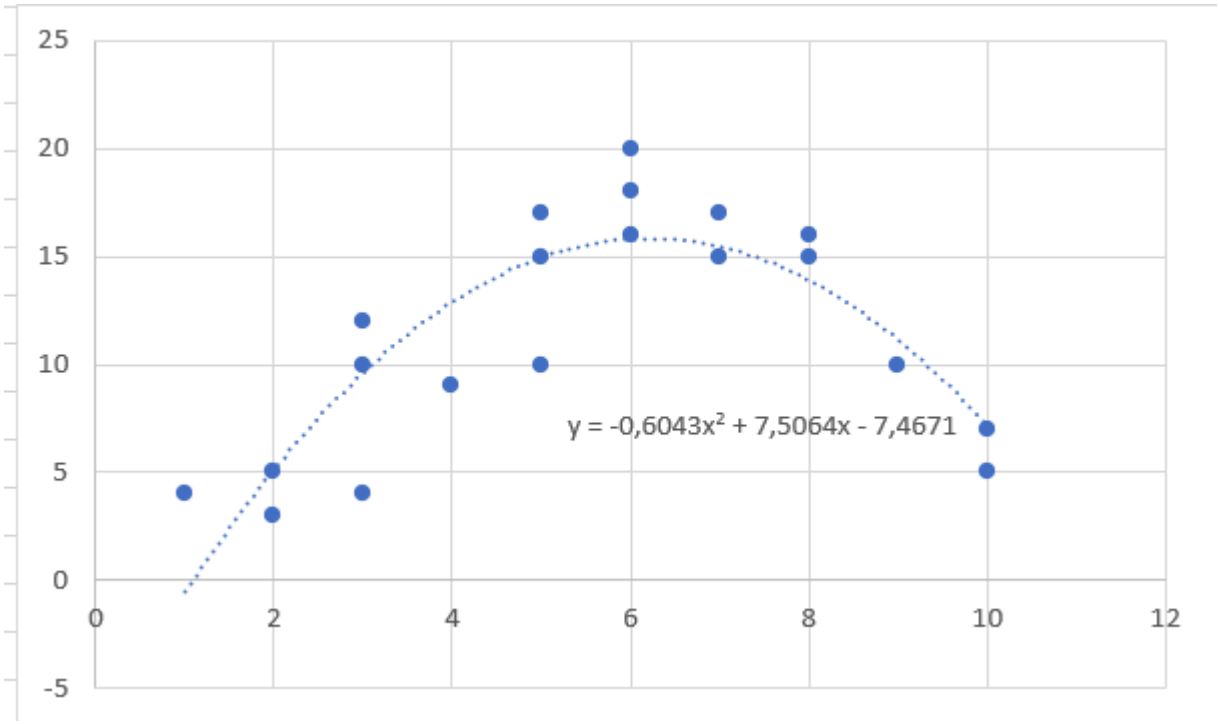
7.7. Other models

As we discovered looking at the scatter diagrams, we cannot always form a line to describe the dots. If we see, that the connection is polynomial, we must use the polynomial model. We can choose this model from the menu.

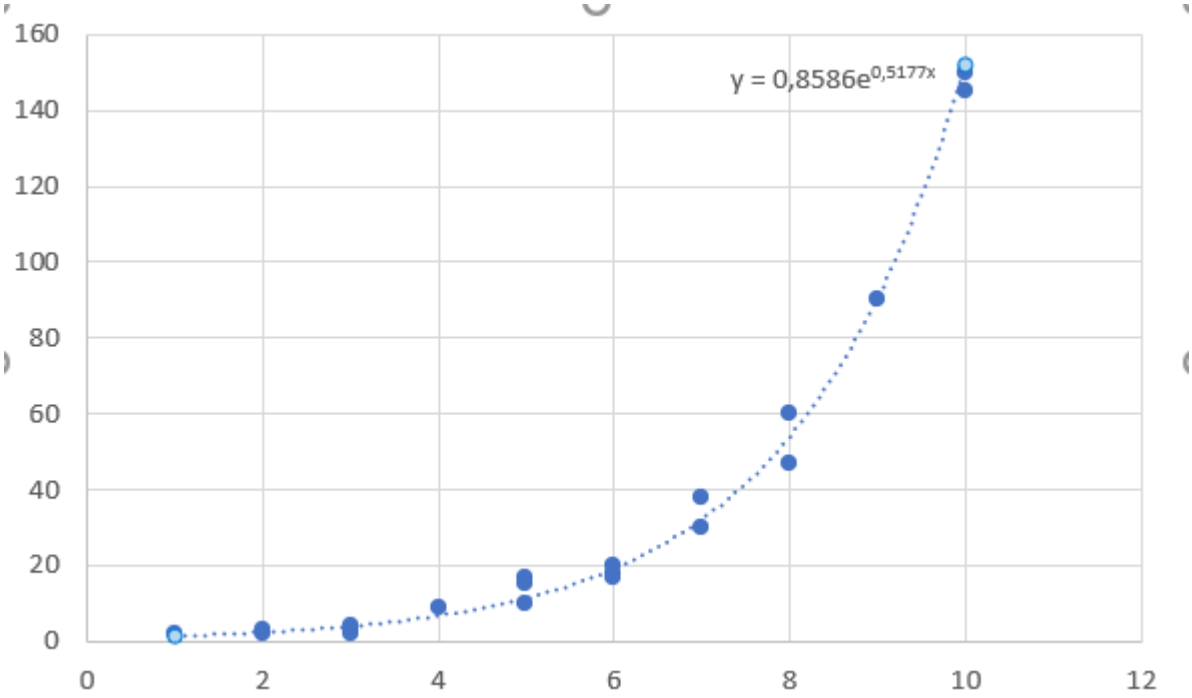
The polynomial model for quadratic function is $y = Ax^2 + Bx + C$.



From here you can choose to see the equation: $y = -0{,}6043x^2 + 7{,}5064x - 7{,}4671$.



The exponential model is in the form of $y = A \cdot B^x$, where A and B are constants.



In this material, the connection is $y = 0{,}8586 \cdot e^{0{,}5177x}$.

7.8. Coefficient of determination

The fact how well you can describe the behaviour of the y variable with the help of the x variable, is measured with the coefficient of determination. This tells you that proportional share of the y variables fluctuation that can be explained with the help of the x variable.

The coefficient of determination is correlation coefficient squared: $R^2 = r^2 \cdot 100\ \%$. This can be presented as a decimal form but there is no established form.

The greater this coefficient is, the better the curve fits to the diagram. This also means more accurate forecasts.

7.9. General linear model

We can form the linear model with more than just one x variable. Now we choose several x variables to explain the behaviour of the y variable. This means that the model is written $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$, where b_i is the slope of the i's x-variable.

Each b_i tells you, how the y variable changes when this x variable is increased by one (as the other x's stay constant). We can form these models with the help of computer programs.

	A	B	C	D	E	F	G	H	I	J	K
1	x1	x2	x3	y							
2	1	3	10	2							
3	2	4	19	3							
4	3	4	20	4							
5	3	12	40	2							
6	4	7	32	3							
7	2	6	10	5							
8	5	5	17	10							
9	6	4	24	19							
10	5	1	27	13							
11	6	3	38	20			=LINREGR(D2:D21;A2:C21;TOSI;EPÄTOSI)				
12	7	8	40	25							
13	8	9	45	28			0,043805	-0,08412	3,954395	-6,71124	
14	6	10	64	22							
15	9	10	28	30							
16	10	6	10	32							
17	10	5	18	34							
18	7	3	50	22							
19	8	2	32	25							
20	5	1	19	12							
21	3	9	20	5							

EXCEL gives us using the LINEST function $y = -6{,}71124 + 0{,}043805x_1 - 0{,}08412x_2 + 3{,}954395x_3$. Note that the constant b_0 is given last.

8. Statistical testing

8.1. Basics

In statistics we usually try to make conclusions based on the population. If we generalize the phenomenon from the sample to the whole population, we'll have to calculate the propability of a false conclusion. For this, we need statistical testing. These are based on propability distributions.

Parametrical tests rely on normal distribution or distributions like the normal distribution, whereas non-parametrical tests are independent from distributions. These aren't as specific as the parametrical ones, but they come in hand when we cannot perform the material according to the parametrical test.

In the most simple statistical deductioning we estimate statistics. We try to estimate the population's statistics based on the statistics from the sample.

8.2. Confidence interval of sample mean

One way to estimate the interval where the population's statistics lie, is to calculate the confidence interval. Lets calculate the confidence interval of sample mean. The confidence interval is an interval that lies symmetrically on both sides of the sample meanm and into

this the population mean goes with a given propability. For this we must define the confidence level and so on the propability of false conclusion. this propability is marked with α. If $\alpha = 0{,}01$, the propability of a false conclusion is 1 %, and the confidence level is 99 %. Most commonly used confidence levels are 95 %, 99 % and 99,9 %, but there are also others. The more sure we wish to be, the wider the confidence level has to be. The accuracy also increases when the sample size increases.

You can calculate the confidence interval by adding and subtracting the marginal of error from the sample mean. For the marginal of error we must consider, how the different sample means are distributed. So if would take every single sample we possibly can from the population, and we would count all the sample's means, how would they be distributed?

This is examined with the help of the central limit theorem. Based on it, it is proven that if the original variable obeys normal distribution with the expected value of μ and standard deviation of σ, then the distribution of the sample means is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, where n is the sample size.

Now we can find the marginal of error: $e = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. So we multiply the deviation of sample means with the critical value $z_{\frac{\alpha}{2}}$, which is defined based on the confidence level.

Finding the critical value

The critical value can be found when we place the confidence level symmetrically around the expected value. Now on both edges there is equally large "tails" (=half of the error's propability). Finding that z that separates the confidence level from the tail, we get the critical value. This can be done with the help of the normal distribution table.

- 95 % $z_{\frac{\alpha}{2}} = 1{,}96$
- 99 % $z_{\frac{\alpha}{2}} = 2{,}58$
- 99,9 % $z_{\frac{\alpha}{2}} = 3{,}30$

So the confidence interval is: $\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \ \leq \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

It is ok to use normal distribution for this, when we are doing a census study. When we have a sample (, which sample size is below 30), we use Student's t-distribution to find the critical values.

Student's t-distribution is like normal distribution, symmetrical, but it's more accurate with small samples than the real normal distribution. The form of the Student's t-distribution is

based on the degrees of freedom, df. If the sample size is n, then the degrees of freedom in the sample is $n-1$. You can find the table for Student's t-distribution from the table above.

Confidence interval, when the sample size is under 30: $\overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$.

Example 1

A statistical researcher has made a small study (n=10) with a certain IQ-measuretest. From the material, the researcher has calculated that the mean is 110,5 and the deviation is 16,06. The researcher wishes to calculate the 95 % confidence interval.

From the table, we get the critical value, $df = 10 - 1 = 9$, And the error propability is 5 %, so the critical value is 2,262.

Confidence interval:

$$e = 2{,}262 \cdot \frac{16{,}06}{\sqrt{10}} = 11{,}4878$$
$$110{,}5 - 11{,}4878 \leq \mu \leq 110{,}5 + 11{,}4878$$
$$99{,}0122 \leq \mu \leq 121{,}9878$$
$$99 \leq \mu \leq 122$$

So this means that the mean of the whole population is in the interval [99, 122] with the propability of 95 %.

OBS! When the sample size is at least 30, the difference between the sample and population standard deviations isn't statistically significant. This is why we can use normal distribution table to find the critical value whith bigger samples.

8.3. Testing hypotheses

In a statistical testing, we always test a certain presumption. This is called hypothesis. It means that the researcher has a presumption, what the state is (the null-hypothesis), and it holds until proven otherwise. Testing the hypothesis goes forward always according to a certain protocol or "recipe":

1) Set the hypothesis,

2) Make the sample and collect data,

3) analyze data and test it with a suitable statistical test,

4) evaluate the state of the hypothesis and make the final conclusion.

There should be two hypotheses: the null-hypotheses, which is true until proven otherwise, and the alternative hypotheses, which is true, if the null-hypotheses is proven to be false. Only one of these hypotheses can be true.

8.4. Errors in statistical testing

We won't ever get a 100 % certainty of things. Even considering the confidence interval we set the confidence level, leaving there a small chance of error. This means that we'll have to endure always a certain amount of uncertainty. Each conclusion is made with a specific chance for error. There are two kinds of errors in statistical testing.

Type 1 error = false positive solution – so the alternative hypotheses is true even though the situation is supporting the null-hypotheses.

Type 2 error = false negative solution – so the null-hypotheses is true even though the situation supports the alternative hypotheses.

From these we are trying to avoid the type 1 error. In the case of type 2 error, the harm is not that bad as in the case of type 2 error. This is because "the situation will not get worse", the situation stays the same as it was. In the case of type 1 error, a new concept or perception is adapted. If this new perception is faulty, it can create even dangerous situations.

The propability for type 1 error is expressed with p-value. This is considered to be a numerical approximation of the state of the hypotheses. p-value is usually used to express the propability of error in statistical testing. According to its definition, the p-value tells us the propability to have equally deviant or more deviant result as in the sample while the null-hypotheses is true.

- If the p-value is great → the material is in-line with the null-hypotheses. If the null hypotheses is rejected, the propabilty for error would be great.
- If the p-value is small → the material is not in-line with the null-hypotheses and it supports the alternative hypotheses. The propability for error while rejecting the null-hypotheses is very small.

OBS!!!! Neither conclusion does ever mean that one or the other hypotheses in right/wrong. The conclusion only tells us, which hypotheses is supported by this material.

8.5. Comparing means

The purpose of this test is to see, wether the mean of the sample differs significantly from the mean of the population. While looking into the confidence interval, we already met the

distribution of sample means. Now we'll make a test, which states where the sample mean lies.

Testing is done by calculationg a testvalue. This value is the sample mean standardized. Now with the help of this value and the Student's t-distribution we can calculate the p-value for the test. The test value:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}.$$

In general the testing happens while using computer programs. Most programs calculate the p-value straight. You'll just have to compare the p-value to the chosen propability of error. The testing can also be done by finding the critical value for the test from the table, and comparing the test value to this critical value. The area which supports the null-hypotheses is between the positive and negative critical values, and the area which supports the alternative hypotheses is outside the critical values (the "tails").

The null-hypotheses is: $H_0 : \overline{x} = \mu$, aka. the sample mean doesn't differ from the population mean significantly. The alternative hypotheses is that there is a significant difference: $H_1 : \overline{x} \neq \mu$.

Example 2

A researcher has collected a small sample (n=10) by using a specific IQ-test. From this sample he has calculated that the sample mean is 110,5 and sample standard deviation is 16,06. He wishes to discover wether this sample mean differs from the population mean (100) significantly, when the propability for error can be only 5 %.

$H_0 : \ \overline{x} = \mu$
$H_1 : \ \overline{x} \neq \mu$

Test value:

$$t = \frac{110{,}5 - 100}{\frac{16.06}{\sqrt{10}}} = 2{,}067$$

This obeys the Student's t distribution with degrees of freedom of 10-1=9. The critical value is now 2,262 (from the table).

Because $t = 2{,}067 < 2{,}262 = t_{\frac{\alpha}{2}}$, The test value supports the null-hypotheses.

This can be made with EXCEL by finding the p-value straight from student's t-distribution. We need the test value: 2,067. Now T.DIST.2T(x,deg_freedom) we can calculate the p-value.

We write =T.DIST.2T(2,067; 9) (so x is the test value, and deg_freedom is the degrees of freedom). We get 0,068697. This means that if we reject the null hypotheses, the propability for error is around 6,9 %. This is greater than the allowed 5 %. So the material supports the null-hypotheses.

More info! Testing can be done with 2-tails or with 1-tail. With 2-tailed testing the confidence level is set like counting the confidence interval, symmetrically around the population mean. This leaves us two rejection areas, the tails on both sides. In the 2-tail testing we don't take part in the fact wether the sample mean is greater or smaller than the population mean. We just look, is there a difference. In the 1-tail testing the rejection area is set only to one side. This means that we'll have to say, if the sample mean is greater or smaller.