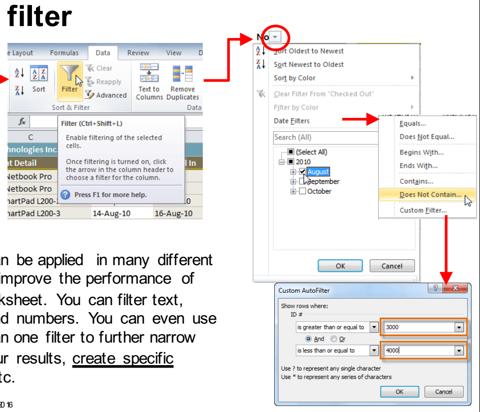
Using filter

 1

Using filter

Select row



Filters can be applied in many different ways to improve the performance of your worksheet. You can filter text, dates, and numbers. You can even use more than one filter to further narrow down your results, <u>create specific groups</u> etc.

Are means of two sets of data really different?

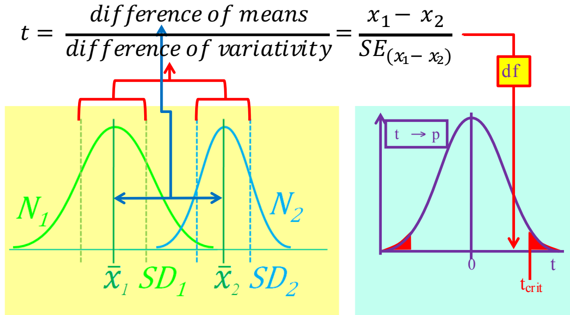
Student's t-test

The t statistic was introduced in 1908 by **<u>William Sealy Gosset</u>** (1876-1937), a statistician working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). Gosset had been hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes. Gosset devised the t-test as a way to cheaply monitor the quality of beer.

 3

t-statistik kahe keskmise võrdlemiseks

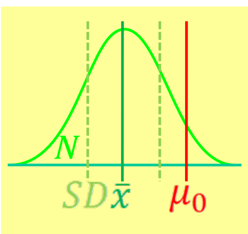
$$t = \frac{difference\ of\ means}{difference\ of\ variativity} = \frac{x_1 - x_2}{SE_{(x_1 - x_2)}}$$



 4

One-sample t-test:

Student's t-test for comparison of mean widh standard



Whether the sample mean differs from a standard μ_0?

$$t = \frac{\bar{x} - \mu_0}{\frac{SD}{\sqrt{N}}}$$

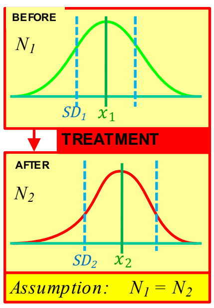
$$df = N - 1$$

t =ABS(X-M₀)/(SD/SQRT(N))
p=TDIST(t;(N-1);2)
p=TDIST(ABS(X-M₀)/(SD/SQRT(N));(N-1);2)

 5

Paired t-test:

Compares (before and after) means of two paired groups



Used to compare means on the same or related subject over time or in differing circumstances; subjects are often tested in a before-after situation.

Given two paired sets X_1 and X_2 of N measured values, the paired t-test determines if they differ from each other in a significant way.

$$t = \frac{\bar{x}_d}{\frac{SD_d}{\sqrt{N}}}$$

$$df = N - 1$$

Where:

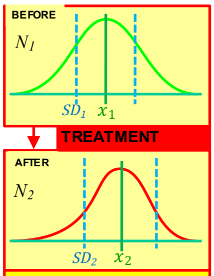
$$x_d = x_1 - x_2$$

$$\bar{x}_d = \frac{\sum x_1 - x_2}{N}$$

$$SD_d = \sqrt{\frac{(x_d - \bar{x}_d)^2}{N - 1}}$$

 6

Paired t-test:
Compares (before and after) means of two paired groups



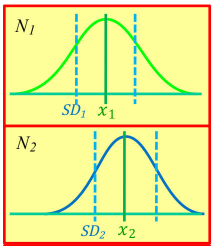
Assumption: $N_1 = N_2$

Function:
=T.TEST(arrayX_1,arrayX_2,tails,type)
=TTEST(arrayX_1,arrayX_2,tails,type)
Arguments:
- Array1 - The first data set; X_1.
- Array2 - The second data set; X_2.
- Tails - Specifies the number of distribution tails.
  - If = 1, one-tailed distribution is used.
  - If = 2, two-tailed distribution is used.
- Type = 1 (Paired)

Returns the probability (p) associated with a Student's t-Test.

7

Two-sample unpaired t-test assuming equal variance (homoscedastic t-test)



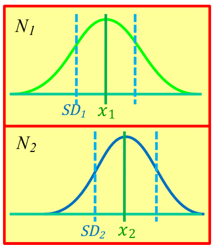
$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left[\frac{N_1 + N_2}{N_1 \times N_2}\right] \times \left[\frac{(N_1 - 1)\times SD_1^2 + (N_2 - 1)\times SD_2^2}{N_1 + N_2 - 2}\right]}}$$

$$df = N_1 + N_2 - 2$$

Assumptions:
$SD_1 = SD_2$; $N_1 \neq N_2$

8

Two-sample unpaired t-test assuming equal variance (homoscedastic t-test)



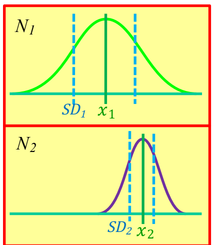
Assumptions:
$SD_1 = SD_2$; $N_1 \neq N_2$

Function:
=T.TEST(arrayX_1,arrayX_2,tails,type)
=TTEST(arrayX_1,arrayX_2,tails,type)
Arguments:
- Array1 - The first data set; X_1.
- Array2 - The second data set; X_2.
- Tails - Specifies the number of distribution tails.
  - If = 1, one-tailed distribution is used.
  - If = 2, two-tailed distribution is used.
- Type = 2 (Two-sample equal variance (homoscedastic))

Returns the probability (p) associated with a Student's t-Test.

9

Two-sample unpaired t-test assuming unequal variance (heteroscedastic)



Assumptions:
$SD_1 \neq SD_2$; $N_1 \neq N_2$

if the variance in the two groups are extremely different. e.g. the two samples are of very different sizes

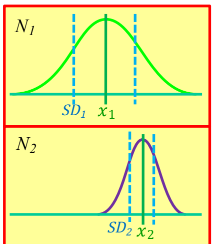
$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}}$$

$$df = \frac{\left(\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}\right)^2}{\frac{\left(\frac{SD_1^2}{N_1}\right)^2}{N_1 + 1} + \frac{\left(\frac{SD_2^2}{N_2}\right)^2}{N_2 + 1}} - 2$$

round up to integer

10

Two-sample unpaired t-test assuming unequal variance (heteroscedastic)



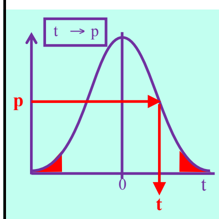
Assumptions:
$SD_1 \neq SD_2$; $N_1 \neq N_2$

Function:
=T.TEST(arrayX_1,arrayX_2,tails,type)
=TTEST(arrayX_1,arrayX_2,tails,type)
Arguments:
- Array1 - The first data set; X_1.
- Array2 - The second data set; X_2.
- Tails - Specifies the number of distribution tails.
  - If = 1, one-tailed distribution is used.
  - If = 2, two-tailed distribution is used.
- Type = 3 (Two-sample unequal variance (heteroscedastic))

Returns the probability (p) associated with a Student's t-Test.

11

Obtaining t-value from p-value

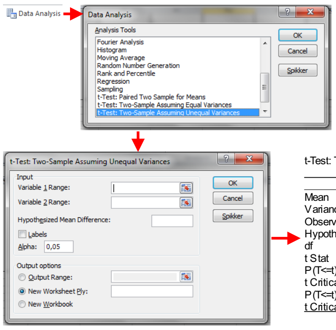


Function:
=T.INV(probability;deg_freedom)

returns the t-value of the Student's t-distribution as a function of the probability and the degrees of freedom

12

t-test using Data Analysis Package



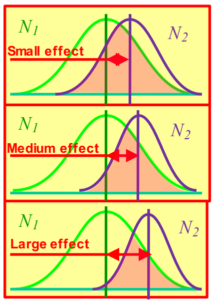
t-Test: Two-Sample Assuming Unequal Variances

	x	y
Mean	14,77	12,94
Variance	6,67	6,30
Observations	63	63
Hypothesized Mean Difference	0	
df	124	
t Stat	4,03	
P(T<=t) one-tail	0,00	
t Critical one-tail	1,66	
P(T<=t) two-tail	0,00	
t Critical two-tail	1,98	

 13

Effect size for difference of mean

The standardized effect size - Cohen's d:



Effect size is a standardized measure of the difference between two (or more) group means

$$d = \frac{x_1 - x_2}{SD_{1,2}}$$

$$SD_{1,2} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Verbal Description	*Effect Size (d)*	*Populations overlap*
Small	0,20	85%
Medium	0,50	67%
Large	0,80	53%

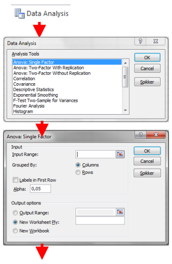
 14

ANOVA - ANalysis Of Variance

Dispersioonanalüüs

NB! Data: two or more adjacent ranges

Grupp 1	*Grupp 2*	*Grupp 3*
1	2	4
3	4	5
3	2	5
2	1	3
1	3	4
1	2	2
2	2	2
2	3	4
2	2	5
3	4	5



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Grupp 1	10	20	2,0	0,67
Grupp 2	10	25	2,5	0,94
Grupp 3	10	39	3,9	1,43

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19,4	2	9,70	9,56	0,00	3,35
Within Groups	27,4	27	1,01			
Total	46,8	29				

- Purpose of analysis of variance (ANOVA) is to test for significant differences between means.
- Excel:
  Data analysis - ANOVA

15

Scatterplot

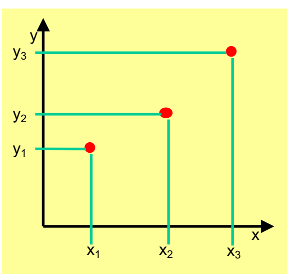
Hajuvusdiagramm

Relation of parameters

 16

Scatterplot

Hajuvusdiagramm



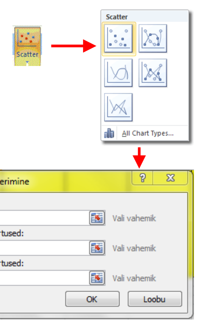
- Scatterplot graphs are used to show trends in data.
- The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis (x) and the value of the other variable determining the position on the vertical axis (y).

 17

Scatterplot

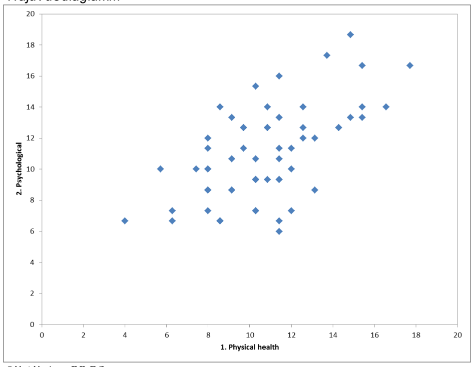
Hajuvusdiagramm

Select some data



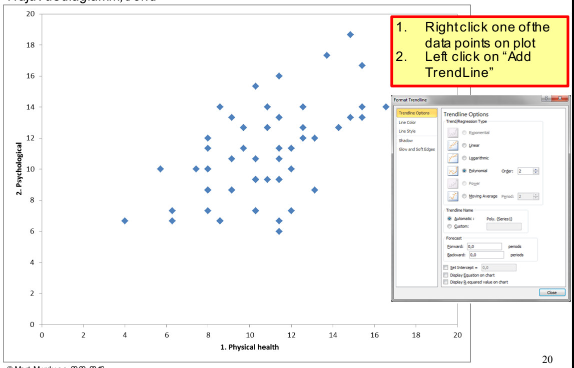
 18

Scatterplot
Hajuvusdiagramm



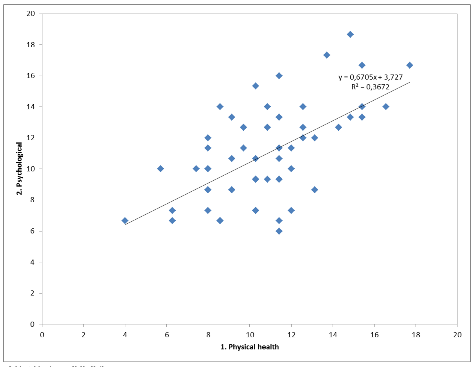
19

Scatterplot - adding trendline
Hajuvusdiagramm, trend

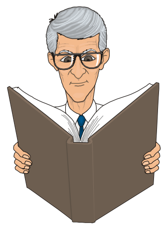


20

Scatterplot with trendline, equation of trendline and coefficent of determination

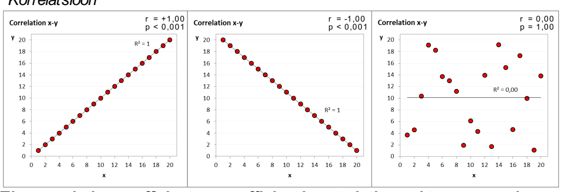


21



Correlation
relation of parameters

Correlation, Pearson r
Korrelatsioon



The correlation coefficient or r coefficient is a statistic used to measure the degree or strength of this type of relationship.

- The correlation r value represents both a magnitude and a direction of the linear relationship (positive or negative).
- It may take on a range of values from -1 to 0 to +1
- The values are absolute and nondimensional with no units involved.
- A correlation coefficient of zero indicates that no association exists between the measured variables. The closer the r coefficient approaches ± 1, regardless of the direction, the stronger is the existing association indicating a more linear relationship between the two variables.

23

Correlation, Pearson r
Korrelatsioon

$$r_{xy} = \frac{\sum_{i=1}^{N} z_x z_y}{N-1}$$

Where:

$$z_x = \frac{x - x}{SD_x} \qquad z_y = \frac{y - y}{SD_y}$$

Where:

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

$$SD_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y)^2}{N-1}}$$

Correlation between two parameters is average of multiplied z-scores of the parameters.

Function:
=CORREL(arrayX;arrayY)
=PEARSON(arrayX;arrayY)
Returns the correlation coefficient between two data sets.

Shorter way:

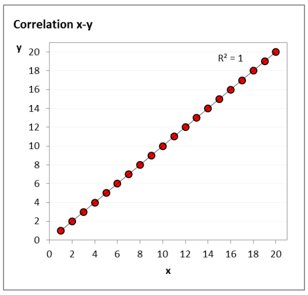
$$r_{xy} = \frac{\sum(x - x)(y - y)}{\sqrt{\sum(x - x)^2 \sum(y - y)^2}}$$

24

Correlation

No	x	y
N1	1	1
N2	2	2
N3	3	3
N4	4	4
N5	5	5
N6	6	6
N7	7	7
N8	8	8
N9	9	9
N10	10	10
N11	11	11
N12	12	12
N13	13	13
N14	14	14
N15	15	15
N16	16	16
N17	17	17
N18	18	18
N19	19	19
N20	20	20

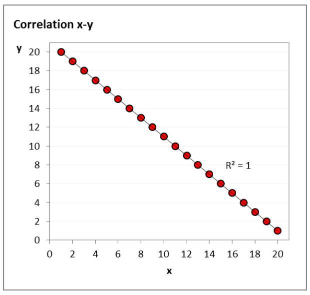
r =	1,00
p =	0,00
Mx =	10,5
My =	10,5
SDx =	5,9
SDy =	5,9



Correlation

No	x	y
N1	1	20
N2	2	19
N3	3	18
N4	4	17
N5	5	16
N6	6	15
N7	7	14
N8	8	13
N9	9	12
N10	10	11
N11	11	10
N12	12	9
N13	13	8
N14	14	7
N15	15	6
N16	16	5
N17	17	4
N18	18	3
N19	19	2
N20	20	1

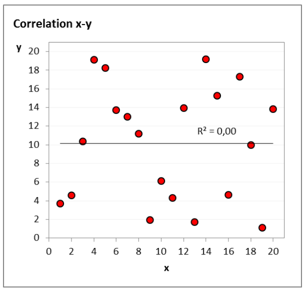
r =	-1,00
p =	0,00
Mx =	10,5
My =	10,5
SDx =	5,9
SDy =	5,9



Correlation

No	x	y
N1	1	3,7
N2	2	4,6
N3	3	10,3
N4	4	19,1
N5	5	18,3
N6	6	13,7
N7	7	13,0
N8	8	11,2
N9	9	1,9
N10	10	6,1
N11	11	4,3
N12	12	13,9
N13	13	1,7
N14	14	19,2
N15	15	15,3
N16	16	4,6
N17	17	17,3
N18	18	10,0
N19	19	1,1
N20	20	13,8

r =	0,00
p =	1,00
Mx =	10,5
My =	10,2
SDx =	5,9
SDy =	6,2

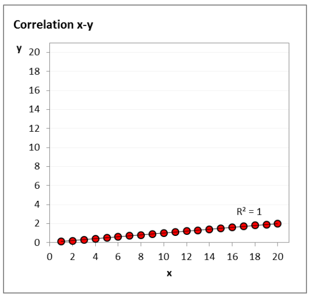


Correlation

NB! Correlation describes relative not absolute change!

No	x	y
N1	1	0,1
N2	2	0,2
N3	3	0,3
N4	4	0,4
N5	5	0,5
N6	6	0,6
N7	7	0,7
N8	8	0,8
N9	9	0,9
N10	10	1,0
N11	11	1,1
N12	12	1,2
N13	13	1,3
N14	14	1,4
N15	15	1,5
N16	16	1,6
N17	17	1,7
N18	18	1,8
N19	19	1,9
N20	20	2,0

r =	1,00
p =	0,00
Mx =	10,5
My =	1,05
SDx =	5,9
SDy =	0,6

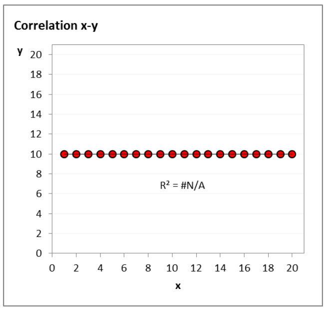


Correlation

NB! Correlation describes change! Therefore correlation with constant is indeterminable

No	x	y
N1	20	10,0
N2	19	10,0
N3	18	10,0
N4	17	10,0
N5	16	10,0
N6	15	10,0
N7	14	10,0
N8	13	10,0
N9	12	10,0
N10	11	10,0
N11	10	10,0
N12	9	10,0
N13	8	10,0
N14	7	10,0
N15	6	10,0
N16	5	10,0
N17	4	10,0
N18	3	10,0
N19	2	10,0
N20	1	10,0

r =	#DIV/0!
p =	#DIV/0!
Mx =	10,5
My =	10,0
SDx =	5,9
SDy =	0,0

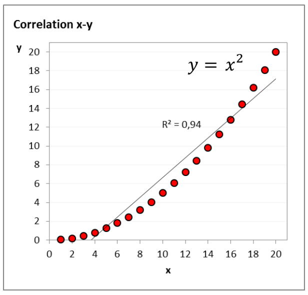


Correlation

NB! Correlation describes linear relationship! Therefore correlation is not the best way for detecting non-linear relationships.

No	x	y
N1	1	0,1
N2	2	0,2
N3	3	0,5
N4	4	0,8
N5	5	1,3
N6	6	1,8
N7	7	2,5
N8	8	3,2
N9	9	4,1
N10	10	5,0
N11	11	6,1
N12	12	7,2
N13	13	8,5
N14	14	9,8
N15	15	11,3
N16	16	12,8
N17	17	14,5
N18	18	16,2
N19	19	18,1
N20	20	20,0

r =	0,97
p =	0,00
Mx =	10,5
My =	7,2
SDx =	5,9
SDy =	6,4



Correlation

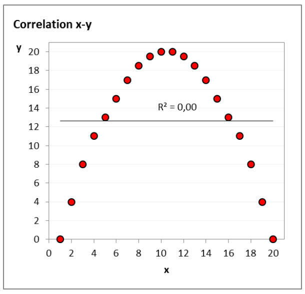
NB! Correlation describes linear relationship! Therefore correlation is not the best way for detecting non-linear relationships.

No	x	y
N1	1	0
N2	2	4
N3	3	8
N4	4	11
N5	5	13
N6	6	15
N7	7	17
N8	8	18,5
N9	9	19,5
N10	10	20
N11	11	20
N12	12	19,5
N13	13	18,5
N14	14	17
N15	15	15
N16	16	13
N17	17	11
N18	18	8
N19	19	4
N20	20	0

r = 0,00
p = 1,00

Mx = 10,5
My = 12,8

SDx = 5,9
SDy = 6,4



Recommendation: use scatterplot.

Elements of correlation matrix

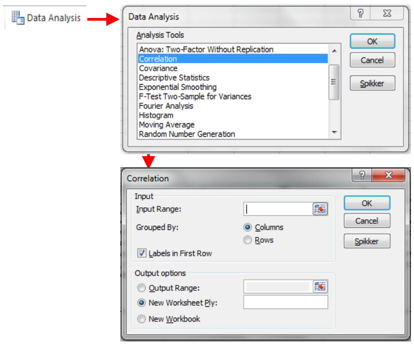
Parameter names

	a	b	c	d	e	f
a	r_{aa}=1	r_{ab}	r_{ac}	r_{ad}	r_{ae}	r_{af}
b	r_{ba}	r_{bb}=1	r_{bc}	r_{bd}	r_{be}	r_{bf}
c	r_{ca}	r_{cb}	r_{cc}=1	r_{cd}	r_{ce}	r_{cf}
d	r_{da}	r_{db}	r_{dc}	r_{dd}=1	r_{de}	r_{df}
e	r_{ea}	r_{eb}	r_{ec}	r_{ed}	r_{ee}=1	r_{ef}
f	r_{fa}	r_{fb}	r_{fc}	r_{fd}	r_{fe}	r_{ff}=1

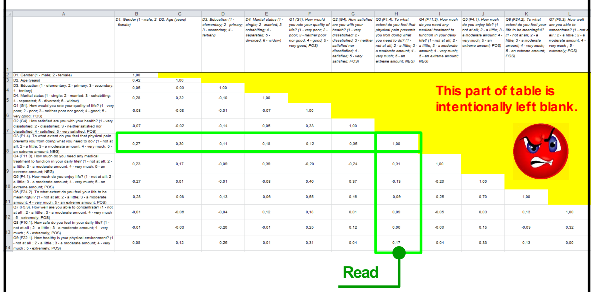
Symmetrical (r_{cb}= r_{bc}; r_{db}= r_{bd}; etc)

Correlation coefficent for correlation between parameters f and a

Correlation



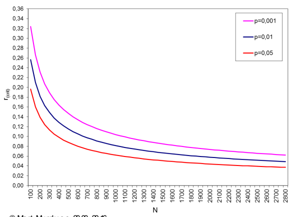
Correlation matrix



Statistical significance of correlations

N=	r_{crit} (p=0,05)
5	0,8784
10	0,6319
50	0,2788
100	0,1966
1000	0,0620
2000	0,0439

- The statistical significance of the correlation depends on the absolute value of the correlation and sample size.
- For each sample sizes there is critical value for the correlation coefficient, above which the absolute value of the correlations is considered statistically meaningful.



Statistical significance of correlations - calculations

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{N-2}}}$$

Using t-test:
t=ABS(r)/SQRT((1-POWER(r;2))/(N-2))
p=TDIST(t;(N-1);2)
p=TDIST((ABS(r)/SQRT((1-POWER(r;2))/(N-2));(N-1);2)

$$t = |r|\sqrt{\frac{N-2}{1-r^2}}$$

t=ABS(r)*SQRT((N-2)/(1-POWER(r;2)))
p=TDIST(t;(N-1);2)
p=TDIST(ABS(r)*SQRT((n-2)/(1-POWER(r;2)));N-2;2)

$$z = r\sqrt{n-1}$$

Using z-test:
z=r*SQRT(N-1)
p=NORM.S.DIST(z;FALSE)
p=NORM.S.DIST(r*SQRT(N-1);FALSE)

Calculating t-value from r

B2 f_x =ABS('r'!B2)/(SQRT((1-POWER('r'!B2;2))/(65-2)))

	A	B	C	D
1		*D1. Gender (1 - male; 2 - female)*	*D2. Age (years)*	*D3. Educatio elementary; 2 - pri secondary; 4 - t*
2	D1. Gender (1 - male; 2 - female)	#DIV/0!		
3	D2. Age (years)	3,71		
4	D3. Education (1 - elementary; 2 - primary; 3 - secondary; 4 - tertiary)	0,39		
5	D4. Marital status (1 - single; 2 - married; 3 - cohabiting; 4 - separated; 5 - divorced; 6 - widow)	2,35		

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{N-2}}}$$

Calculating p-value from t

B2 f_x =TDIST(t!B2;65-2;2)

	A	B	C
1		*D1. Gender (1 - male; 2 - female)*	*D2. Age (years)*
2	D1. Gender (1 - male; 2 - female)	#DIV/0!	0,00
3	D2. Age (years)	0,00	#DIV/0!
4	D3. Education (1 - elementary; 2 - primary; 3 - secondary; 4 - tertiary)	0,37	0,38

Filtering significant correlations

B2 f_x =IF(p!B2>Correlations!A1;"-";'r'!B2)

	A	B	C
1	*0,05*	*D1. Gender (1 - male; 2 - female)*	*D2. Age (years)*
2	D1. Gender (1 - male; 2 - female)	#DIV/0!	0,42
3	D2. Age (years)	0,42	#DIV/0!
4	D3. Education (1 - elementary; 2 - primary; 3 - secondary; 4 - tertiary)	-	-
	D4. Marital status (1 - single; 2		

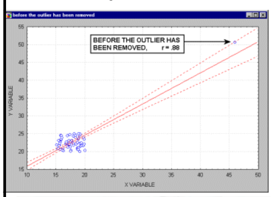
Generating text

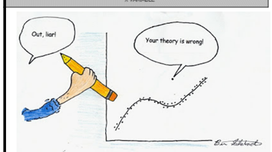
B2 f_x =IF(Correlations!B2="-";"-";CONCATENATE("r = ";FIXED('r'!B2;2);" (p = ";FIXED(p!B2;2);")"))

	A	B	C	D	E
1		*D1. Gender (1 - male; 2 - female)*	*D2. Age (years)*	*D3. Education (1 - elementary; 2 - primary; 3 - secondary; 4 - tertiary)*	*D4. Marita 2 - marriec separate*
2	D1. Gender (1 - male; 2 - female)	#DIV/0!	r = 0,42 (p = 0,00)	-	r = 0
3	D2. Age (years)	r = 0,42 (p = 0,00)	#DIV/0!	-	r = 0
4	D3. Education (1 - elementary; 2 - primary; 3 - secondary; 4 - tertiary)	-	-	#DIV/0!	
5	D4. Marital status (1 - single; 2 - married; 3 - cohabiting; 4 - separated; 5 - divorced; 6 - widow)	r = 0,28 (p = 0,03)	r = 0,32 (p = 0,01)	-	

NB! The expression $p = 0{,}00$ must be later replaced with $p < 0{,}01$

Outlier effect
Eemalasuja efekt



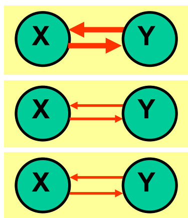


An outlier is a value that is very different from the other data in data set.

Outliers have a profound influence on the slope of the regression line and consequently on the value of the correlation coefficient.

- Identify outliers by examining a scatterplot of each important correlation.
- Useful: exclude values that are outside the range of ±2 standard deviations
- Question: who or what is outlier?
  - Error?
  - Real very different person or event?

Strenght of correlation



Perfect	*Perfektne*	± 1,0
Strong	*Tugev*	± 0,8
Moderate	*Mõõdukas*	± 0,6
Weak	*Nõrk*	± 0,2
No relation	*Seos puudub*	± 0,0

Or

Strong	+/- 0.7 … +/- 1.0
Moderate	+/- 0.5 … +/- 0.70
Weak	0 … +/- 0.5

In social sciences the correlation is considered meaningful if the absolute value of correlation is ≥0,2 (Disputable!)

Importance of small statistical effects

It is important to avoid the error of assuming that small statistical effects necessarily translate into small practical or public health effects. There are many circumstances in which statistically small effects have large practical consequences, especially when small effects accumulate over time and over large proportions of the relevant population.

Huesmann, Taylor 2005

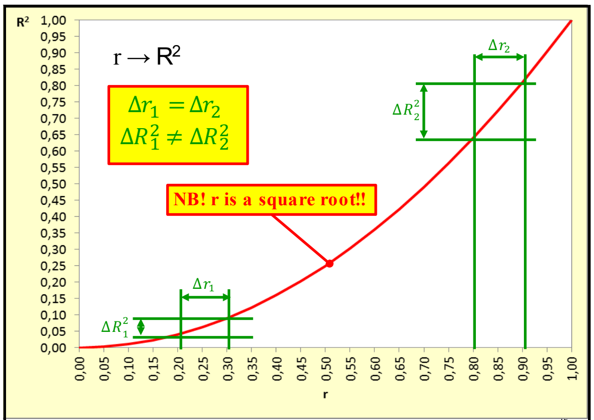
 43

HELDUR MEERITS,
investor ja LHV nõukogu liige

Kui valimiskulud ja saadikukohad on kenasti ritta saanud, siis nendevahelise korrelatsiooni arvutamine on tänapäeval ülilihtne. Tulemused on tabelis näha – ja need tulemused on jahmatavad, ehmatavad ning hirmutavad. Teatavasti võivad korrelatsioonikordaja arvutamisel tulemused olla vahemikus -1st kuni +1ni. Kui tulemus on +1 (ehk 100%), siis liiguvad kaks numbririda täielikult samas «rütmis». … 1995. aastal oli korrelatsioon väga tagasihoidlik, kõigest 67%.

Heldur Meerits: kas raha teeb riigikogu? Postimees. 16.12.2012 http://arvamus.postimees.ee/1076188/heldur-meerits-kas-raha-teeb-riigikogu/

 44



Coefficient of Determination

Determinatsioonikordaja

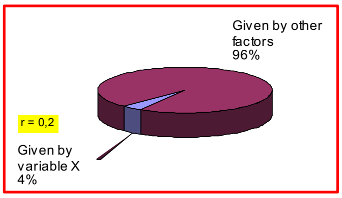
r	R^2
1	100,0%
0,9	81,0%
0,8	64,0%
0,7	49,0%
0,6	36,0%
0,5	25,0%
0,4	16,0%
0,3	9,0%
0,2	4,0%
0,1	1,0%

=RSQ(arrayX;arrayY) returns the square of the Pearson product moment correlation coefficient

is the square of the correlation between two variables (R²).

It expresses the amount of common variation between the two variables.

Special case of effect size.



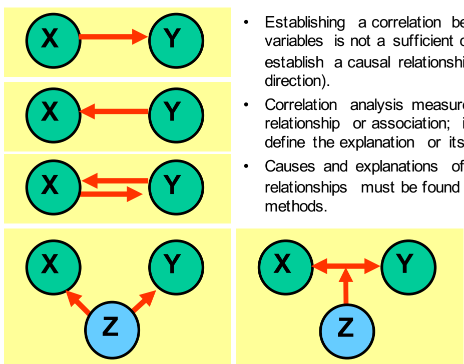
Effect size

R^2	*Character of size of effect*
$0{,}01 < R^2 < 0{,}09$	Small
$0{,}09 < R^2 < 0{,}25$	Moderate
$0{,}25 \leq R^2$	Strong

Cohen, 1988

 47

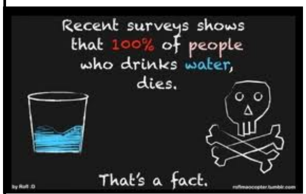
Correlation does not imply causation!

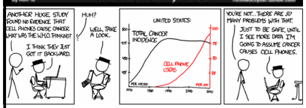


- Establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).
- Correlation analysis measures a relationship or association; it does not define the explanation or its basis.
- Causes and explanations of the relationships must be found using other methods.

48

Correlation does …



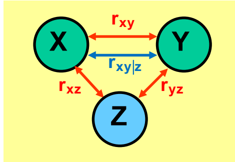


- It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest.
- Children with bigger feet spell better.
- Hight of humans is negatively correlated with hair lenght.
- IMPORTANT WARNING for those who have been drawn unsuspectingly into the use of bread:
  - More than 98 percent of convicted felons are bread users.
  - More than 90 percent of violent crimes are committed within 24 hours of eating bread.
  - Fully HALF of all children who grow up in bread-consuming households score below average on standardized tests.

 49

Partial correlation

Osakorrelatsioon



$$r_{xy|z} = \frac{r_{xy} - (r_{xz} \times r_{yz})}{\sqrt{(1 - r_{xz}^2) \times (1 - r_{yz}^2)}}$$

- Partial correlation is a method used to describe the relationship between two variables (x and y) whilst taking away the effects of another variable (z).
- Possible to find:
  - correlations due to the influence of third variables;
  - hidden correlations, ie, relationships that are masked by the impact of third variables.
- If $r_{XY} \neq r_{XY|Z}$ and the difference of correlation is statistically significant, then the parameter Z is influencing the relationship of XY.

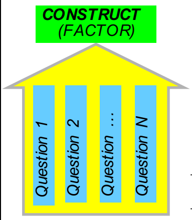
 50

Cronbach´s alpha

Cronbach´i alfa

 51

Cronbach´s alpha



is a coefficient of internal consistency of measures. It is commonly used as an estimate of the reliability of a psychometric test.

Cronbach's alpha indicates how well some features (questions) fit into a single construct:

- $\alpha = 1$, practically same questions
- $\alpha = 0$, totally incompatible questions

Cronbach's alpha	*Internal consistency of construct*
$\alpha \geq 0.9$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

 52

Cronbach´s alpha

Question 1 - s_1^2
Question 2 - s_2^2
Question ... - $s_{...}^2$
Question N - s_N^2

CONSTRUCT *(FACTOR)* s_{sum}^2

Cronbach's α

$$\alpha = \frac{N}{N-1} \times 1 - \frac{\sum_{i=0}^{N} s_i^2}{s_{sum}^2}$$

N – number of questions
s^2 - variance

 53

Standardized Cronbach´s alpha

	K1	K2	K ..	K N
Question 1				
Question 2	r			
Question...	r	r		
Question N	r	r	r	

CONSTRUCT *(FACTOR)*

Standardized α

$$\alpha = \frac{N \times \bar{r}}{1 + (N-1) \times \bar{r}}$$

N – number of correlations
r – arithmetical mean of correlations

 54

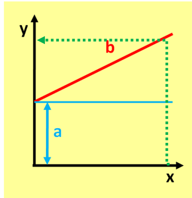
Linear Regression

lineaame regressioon

relationship of parameters

Linear regression analysis

Lineaame regressioon



is an approach to modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted X. The case of one explanatory variable is called simple regression. More than one explanatory variable is multiple regression.

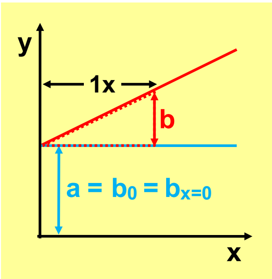
Equation:

y = a + bx

x → y

The relationship between x and y is described using linear equation, although there are other options: square, polynomial, exponential, logarithm, etc equations.

Linear regression equation



y = a + bx

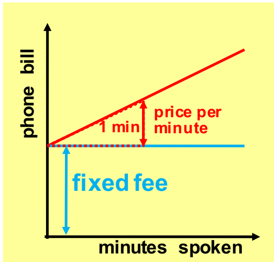
y – dependent variable;

x – independent variable;

a – constant, value of y when x = 0;

b – strenght of relationship, when x changes 1 unit, changes y on amount of b (change of y - Δy), slope of line.

Example: phone bill



y = a + bx

where:

y - the size of a phone bill;

x - the number of minutes spoken;

a - a fixed fee (minimum bill), phone bill, if the phone is not spoken at all;

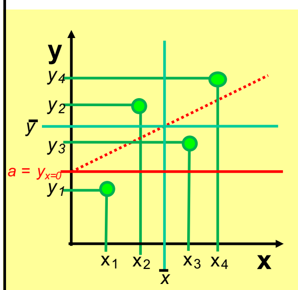
b - price per minute.

Equation:
size of a phone bill = fixed fee + price per minute * number of minutes spoken

Least Squares Method

Vähimruutude meetod

Johann Carl Friedrich Gauss (1777–1855), A.D. 1794



- Solution:

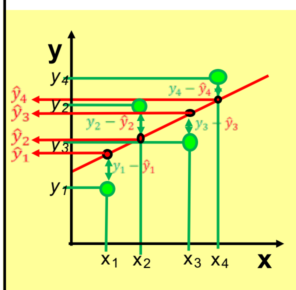
$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- EXCEL functions:
  b = SLOPE(Y:Y;X:X)
  returns the slope of the linear regression line
  a = INTERCEPT(Y:Y;X:X)
  returns the intercept of the linear regression line

How precise is the model?

Standard error of regression



1. Compute values of y from from real x-values using regression equation:

   $\hat{y}_i = a + bx_i$

2. Compute differences of real y and computed y:

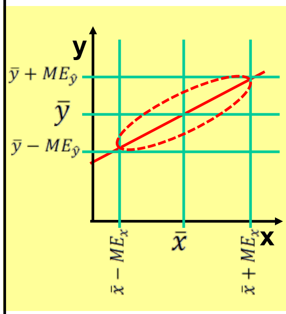
   $\varepsilon_i = y_i - \hat{y}_i$

3. Compute quadratic mean of differences = standard error of regression:

   $$SE_{\hat{y}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

- Excel: SE=STEYX(Y:Y;X:X) returns the standard error of the predicted *y*-value for each *x*-*value* in a regression

Confidence interval of x and y



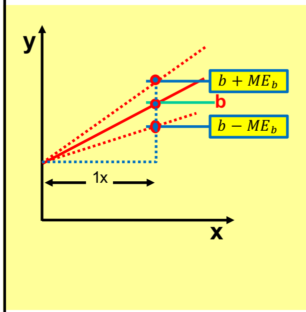
In what range are 95% or 99% of all values of x and y:

$$ME = Z_{crit} * SE$$

ME - margin of error

$$Z_{crit\ 95\%} = 1{,}96$$
$$Z_{crit\ 99\%} = 2{,}58$$

Confidence interval of b



In which interval is b with probability of 95% or 99%:

$$SE_b = \frac{SE_{\hat{y}}}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2}}$$

or

$$SE_b = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2}}$$

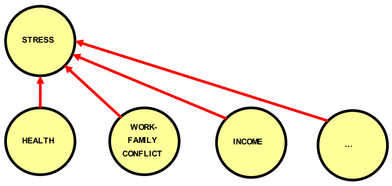
$$ME_b = Z_{crit} * SE_b$$

where:

$Z_{crit\ 95\%} = 1{,}96$
$Z_{crit\ 99\%} = 2{,}58$

Multiple Regression

Mitmene regressioon



- When parameter y is inluenced by multiple parameters: $x_1, x_2, x_3 \dots x_n$
- $y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots b_nx_n$

Multiple Regression

Mitmene regressioon

Example: phone bill

$$\mathbf{y = a + b_1x_1 + b_2x_2}$$

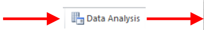
where:

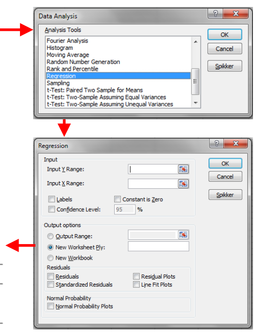
y - the size of a phone bill;
a - a fixed fee (minimum bill), phone bill, if the phone is not used at all;
b_1 - per-minute rate
x_1 - the number of minutes used;
b_2 - call setup fee
x_2 – number of calls

64

Multiple Regression

Y	X1	X2	X3
1	2	4	1
3	4	5	2
3	2	5	4
2	1	3	2
1	3	4	1
1	2	2	1
2	2	2	3
2	3	4	1
2	2	5	1
3	4	5	4





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,87
R Square	0,76
Adjusted R Square	0,64
Standard Error	0,49
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	4,55	1,52	6,27	0,03
Residual	6	1,45	0,24		
Total	9	6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,06	0,60	-0,10	0,92	-1,52	1,40
X1	0,09	0,20	0,45	0,67	-0,40	0,58
X2	0,25	0,16	1,52	0,18	-0,15	0,65
X3	0,43	0,14	3,21	0,02	0,10	0,76

65

Correlation and regression

$$b = r_{xy} \frac{SD_y}{SD_x}$$

$$r_{xy} = \frac{b}{\frac{SD_y}{SD_x}}$$

$$a = \bar{y} - b\bar{x}$$

The standardized slope of linear regression (b) has the same value as the correlation coefficient (r).

- If you are interested in simply characterizing the magnitude of the relationship between two variables, use correlation
- If you are interested in predicting or explaining your results in terms of particular values, use regression.

Useful:

- Collection of statistical terms and definitions:
  - http://www.stats.gla.ac.uk/steps/glossary/index.html
- Internet sources that may be useful for various aspects of statistics:
  - http://www.amstat.org/ (American Statistical Association),
  - http://www.stat.ufl.edu (University of Florida statistics department)
  - http://www.statsoft.com/textbook/ (covers a wide range of topics, the emphasis is on techniques rather than concepts or mathematics)
  - http://www.york.ac.uk/depts/maths/histstat/welcome.htm (some information about the history of statistics)
  - http://www.isid.ac.in/ (Indian Statistical Institute),
  - http://www.isi-web.org/30-statsoc/statsoc/282-nsslist (The International Statistical Institute)
  - http://www.rss.org.uk/ (The Royal Statistical Society),
  - http://lib.stat.cmu.edu/ (an index of statistical software and routines)

Recommendations: