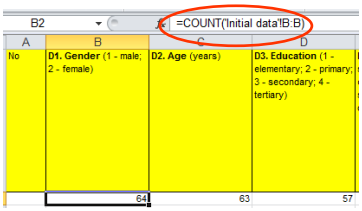
Descriptive statistics

Descriptive statistics provides simple summaries about the sample and about the observations that have been made.

1

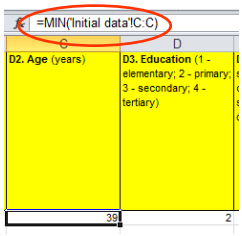
Sample size
Valimi maht



Sample size (N) – how many elements (numbers) are in sample.
For example: how many answeres are there.
Function: =COUNT(range) returns numbers in range.

2

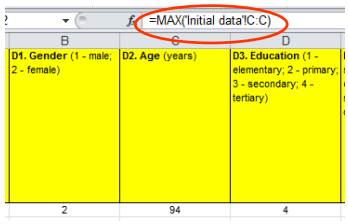
Minimum
Miinimum



Function: =MIN(range) returns the minimum value of a data set

3

Maximum
Maksimum

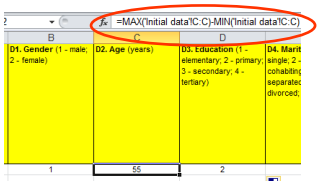


Function: =MAX(range) returns the maximum value of a data set

Recommendation:
The minimum and maximum values should be calculated routinely – values of minimum and maximum show major typing errors on data entry!

4

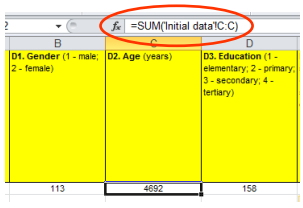
Range
Haare



Function: =MAX(range)-MIN(range) returns range of values

5

Sum
Summa



Function: =SUM(range) adds a range of cells

$$Sum = \sum x = x_1 + x_2 + \ldots + x_n$$

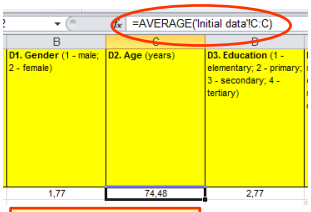
6

Descriptive statistics

Central tendency

Mean (arithmetic average)

Aritmeetiline keskmine



$$\bar{x} = \frac{\sum x}{n}$$

Sign: M, x

Function:

=AVERAGE(range)

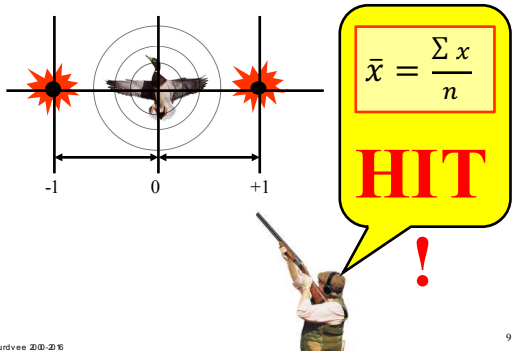
Calculates the mean (arithmetic average) of a range of cells.

The arithmetic mean:

- Depends on the value of all observations.
- Is simple to interpret.
- Is the most familiar and the most used measure.
- Is frequently used as an estimator of the mean of the population.
- **Has a value that can be falsified by the outliers.**

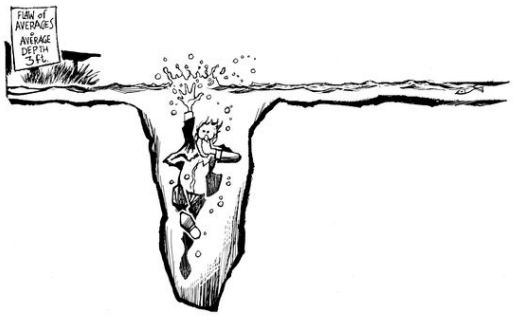
8

Hunting statistician



9

Statistician drowned in a lake with an average depth of a half meter.



Mean temperature of patients in hospital

Patients	*t°*
Patient_1	41,0
Patient_2	41,0
Patient_3	41,0
Patient_4	41,0
Patient_5	41,0
Patient_6	41,0
Patient_7	41,0
Patient_8	41,0
Patient_9	41,0
Patient_10	40,0
Patient_11	40,0
Patient_12	40,0
Patient_13	40,0
Patient_14	10,0
Patient_15	10,0
MEAN	**36,6**

11

Geometric mean

Geomeetriline keskmine

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

=GEOMEAN(range)

- n member values are multipied and the n[th]-degree root is taken.
- Used (in business and banking) for calculation of the average trends not random processes.
- More conservative than the arithmetic average whitch is used for calculation of the probability of processes.

12

Harmonic mean

Harmooniline keskmine

$$\bar{x}_{har} = \frac{n}{\sum \frac{1}{n}}$$

=HARMEAN(range)

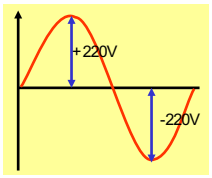
- Reciprocal value is used.
- Used for calculation of average speed, scalars, vectors in the parallel processes
- Tends strongly toward the least elements of the list, it may (compared to the arithmetic mean) mitigate the influence of large outliers and increase the influence of small values.

13

Quadratic mean

Ruutkeskmine

$$\bar{x}_{quad} = \sqrt{\frac{\sum x^2}{n}}$$

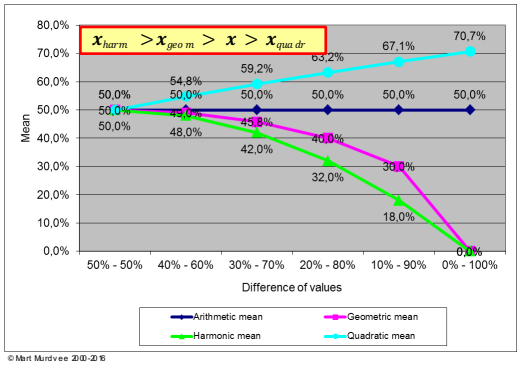


- Squares of the member's values are used in calculation.
- Used for the data sets where range of data includes data which is less than zero.

- Arithmetic mean = 0 V
- Quadratic mean = 220V

14

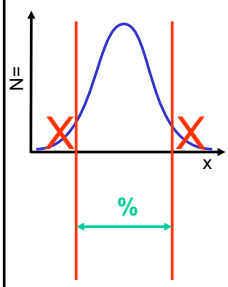
Behavior of the means when there is difference in members values



15

Truncated mean

Trimmitud keskmine



- Truncated mean is calculated from „survived" values after extreme parameter values are removed.
- Truncated mean 50% means that mean is calculated from 50% of values, the 25% of high values and 25% low values removed.
- Is used if the parameter extremes can significantly affect the mean.
- In sports: the extreme ratings of judges are removed.

=TRIMMEAN(range;percent)

16

Weighted mean

Kaalutud keskmine

- Used in cases when members of data set have different characteristics that affect the average or occur at different frequencies
- May be calculated using arithmetic or geometric mean.

	Weight	*Value*
	N	*x*
Sample 1	200	20
Sample 2	20	10
TOTAL / Weighted mean	**220**	**19,09**

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

17

Weighted mean

Kaalutud keskmine

Excel:

	Weight	*Value*
	N	*x*
Sample 1	200	20
Sample 2	20	10
TOTAL / Weighted mean	**220**	**19,09**

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

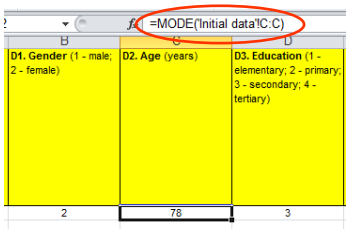
=SUMPRODUCT(range x;range N)

=SUM(range N)

=SUMPRODUCT(range X;range N)/SUM(range N)

18

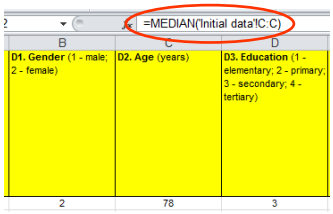
Mode
Mood



Function:
=MODE(range)
returns the most frequently occurring, or repetitive, value in an range of data.

19

Median
Mediaan



Function:
=MEDIAN(range)
Returns the median of the given range.

The median is the number in the middle of a set of numbers; half the values in the data set are greater than the median and half are less than the median

20

Quantiles
kvantiilid

No	Ordered x	Median	Quartiles	Deciles	Percentiles
		Min = 0,32	Q0 = 0,32	Dec0 = 0,32	
1	0,32			1/10	
2	0,38			Dec1 = 0,83	
3	0,88		1/4	1/10	Perc15% = 1,53
4	1,64			Dec2 = 2,07	
5	2,18		Q1 = 2,20	1/10	
6	2,20	1/2		Dec3 = 2,35	
7	2,42			1/10	
8	2,47		1/4	Dec4 = 2,49	
9	2,50			1/10	
10	2,89	Mdn = 3,09	Q2 = 3,09	Dec5 = 3,09	
11	3,29			1/10	
12	4,85			Dec6 = 5,57	
13	6,66		1/4	1/10	
14	6,80			Dec7 = 6,98	
15	7,42	1/2	Q3 = 7,45	1/10	
16	7,56			Dec8 = 7,82	
17	8,87			1/10	
18	9,57		1/4	Dec9 = 9,57	
19	9,59			1/10	
20	9,82	Max = 9,82	Q4 = 9,82	Dec10 = 9,82	

are a set of 'cut points' that divide a sample of data into groups containing (as far as possible) equal numbers of observations.
The most frequently used quantiles are:
- **Median**, which divides a distribution of observation into two parts.
- **Quartiles**, which separate a collection of observations into four parts,
- **Deciles**, which separate a collection of observations into ten parts,
- **Centiles**, which separate a collection of observations into a hundred parts.

Median
Mediaan

No	Ordered x	Median
		Mdn = 0,32
1	0,32	
2	0,38	
3	0,88	
4	1,64	
5	2,18	
6	2,20	1/2
7	2,42	
8	2,47	
9	2,50	
10	2,89	Mdn = 3,09
11	3,29	
12	4,85	
13	6,66	
14	6,80	
15	7,42	1/2
16	7,56	
17	8,87	
18	9,57	
19	9,59	
20	9,82	Max = 9,82

is the middle score of a set if the scores are organised from the smallest to the largest.

Function:
=MEDIAN(range)
returns the median of the given range.

The median:
- Is easy to determine because only one data classification is needed.
- Is easy to understand but less used than the arithmetic mean.
- Is not influenced by outliers, which gives it an advantage over the arithmetic mean, if the series really have outliers.

22

Quartiles
kvartiilid

No	Ordered x	Quartiles
		Q0 = 0,32
1	0,32	
2	0,38	
3	0,88	1/4
4	1,64	
5	2,18	Q1 = 2,20
6	2,20	
7	2,42	1/4
8	2,47	
9	2,50	
10	2,89	Q2 = 3,09
11	3,29	
12	4,85	1/4
13	6,66	
14	6,80	
15	7,42	Q3 = 7,45
16	7,56	
17	8,87	
18	9,57	1/4
19	9,59	
20	9,82	Q4 = 9,82

are values that divide a sample of data into four groups containing equal numbers of observations.
=QUARTILE(range;q)
q = 0 (minimum)
q = 1 (lower quartile)
q = 2 (median)
q = 3 (upper quartile)
q = 4 (maximum)

23

Interquartile mean (IQM), or midmean

is insensitive to outliers.

No	Quartiles
1	
2	
3	1/4
4	
5	Q_1
6	
7	
8	1/4
9	
10	M_e
11	
12	
13	1/4
14	
15	Q_3
16	
17	
18	1/4
19	
20	Q_4

$$IQM = \frac{lower\ quartile + 2 \times median + upper\ quartile}{4}$$

24

Deciles
detsiilid

No	Ordered x	Deciles
		$Dec_0 = 0,32$
1	0,32	**1/10**
2	0,38	$Dec_1 = 0,83$
3	0,88	**1/10**
4	1,64	$Dec_2 = 2,07$
5	2,18	**1/10**
6	2,20	$Dec_3 = 2,35$
7	2,42	**1/10**
8	2,47	$Dec_4 = 2,49$
9	2,50	**1/10**
10	2,89	$Dec_5 = 3,09$
11	3,29	**1/10**
12	4,85	$Dec_6 = 5,57$
13	6,66	**1/10**
14	6,80	$Dec_7 = 6,98$
15	7,42	**1/10**
16	7,56	$Dec_8 = 7,82$
17	8,87	**1/10**
18	9,57	$Dec_9 = 9,57$
19	9,59	**1/10**
20	9,82	$Dec_{10} = 9,82$

are values that divide a sample of data into ten equal groups containing equal numbers of observations.

- The x^{th} decile indicates the value where 10x% of the observations occur below this value and (100 − 10x)% of the observations occur above this value.
- For example, the eighth decile is the value where 80% of the observations fall below this and 20% occur above it.

Percentiles
protsentiilid

No	Ordered x	Percentiles
1	0,32	
2	0,38	
3	0,88	$Perc_{15\%} = 1,53$
4	1,64	
5	2,18	
6	2,20	
7	2,42	
8	2,47	
9	2,50	
10	2,89	
11	3,29	
12	4,85	
13	6,66	
14	6,80	
15	7,42	
16	7,56	
17	8,87	
18	9,57	
19	9,59	
20	9,82	

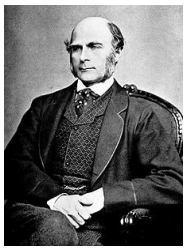
are the values that divide a distribution into 100 equal parts (each part contains the same number of observations).

Function
=PERCENTILE(range;k)
0<k<1
returns the k-th percentile of values in a range.

For example =PERCENTILE(range;0,15) gives value below which are 15% and over which are 85% of sample elements.

Descriptive statistics

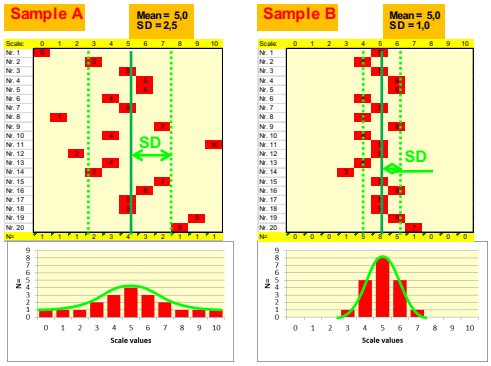
Variance, dispersion
dispersioon, hajuvus



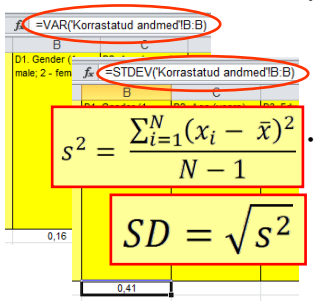
Francis Galton
1822–1911
- Use of standard questionnaire
- Idea of Central Limit Theorem
- Concepts of standard deviation, regression and correlation

Karl Pearson
1857-1936
- Histogram
- Standard deviation
- Correlation coefficent (Pearson r)
- χ^2 chi-square test
- Statistics as a general method that is applicable to all sciences

Variability of Values
Standard Deviation *(standardhälve)*



Variance and standard deviation
Dispersioon ja standardhälve



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

$$SD = \sqrt{s^2}$$

- **Variance** (s^2) is a measure of the spread of or dispersion within a set of data.
  Function:
  =VAR(range)
- **Standard deviation (SD)** is a measure of how widely values are dispersed from the average value (the mean).
  Function:
  =STDEV(range)

Standard score, Z-score

standardiseeritud väärtus, *z-skoor*

$$z = \frac{x - \bar{x}}{SD}$$

Function:
=STANDARDIZE(x;x;SD)

indicates by how many standard deviations a value of data set is above or below the mean.

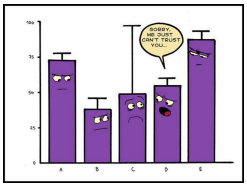
- It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.
- This conversion process is called standardizing or normalizing.
- Standard scores are also called z-values, z-scores, normal scores, and standardized variables.

31

Coefficient of variation (CV)

(variatsioonikordaja)
or **relative standard error (RSE)**

$$CV \text{ or } RSE = \frac{SD}{\bar{x}}$$

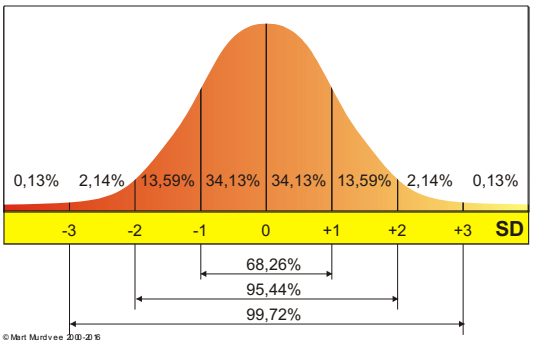


is a normalized measure of dispersion of a distribution and measures the relative dispersion of the studied variable.

- Distributions with CV < 1 are considered low-variance,
- Distributions with CV > 1 are considered high-variance.
- Relative standard error (RSE) is expressed as a percentage

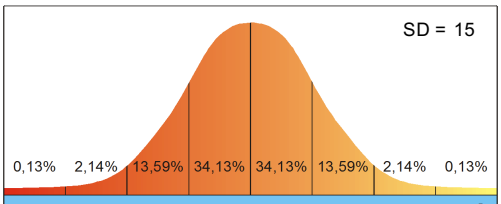
32

Standard deviaton by normal distribution



33

Example: distribution of IQ



N	-3	-2	-1	0	+1	+2	+3
20	0	0	3	7	7	3	0	0
50	0	1	7	17	17	7	1	0
100	0	2	14	34	34	14	2	0
1 340 602	1 743	28 689	182 188	457 547	457 547	182 188	28 689	1 743

34

Standard error of the mean

Standardviga

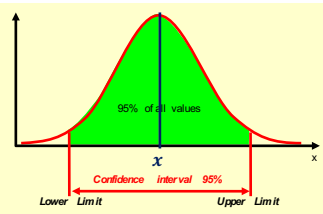
$$SE = \frac{SD}{\sqrt{N}}$$

(SE, SEM) is the standard deviation of the values of a given function of the data (parameter) over all possible samples of the same size.

35

Confidence interval for mean

Usaldusvahemik



Lower limit= $x - z_{1-p} \times SE$

Upper limit = $x + z_{1-p} \times SE$

$z_{1-0,05} = z_{95\%} = 1,95996 = 1,96$

$z_{1-0,01} = z_{99\%} = 2,57583 = 2,58$

Confidence intervals for mean are intervals that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time.

These intervals are referred to as 95% (CI95%) and 99% (CI99%) confidence intervals respectively.

Function:

=CONFIDENCE(alpha;SD;N)

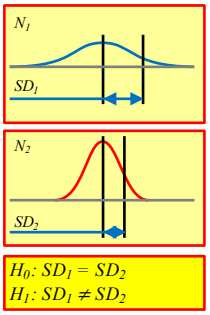
=CONFIDENCE.NORM(alpha;SD;N)

returns the confidence interval value for a population mean using a normal distribution.

=CONFIDENCE.T(alpha;SD;N)

Returns the confidence interval value for a population mean using a Student's t-distribution.

F-test



$H_0: SD_1 = SD_2$
$H_1: SD_1 \neq SD_2$

F-test is used to test if the variances of two populations are equal.

$$F = \frac{s_1^2}{s_2^2}$$

To remember:

$$SD = \sqrt{s^2}$$

Function:
=F.TEST(array1;array2)
returns the result of an F-test, the two-tailed probability (p) that the variances in array1 and array2 are not significantly different.

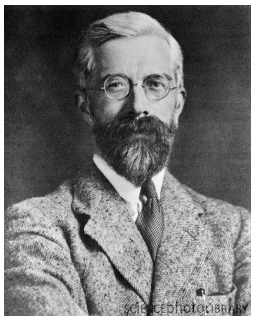
Function:
=F.INV.RT(p;N1-1;N2-1)
returns the inverse of the (right-tailed) F probability distribution

37

Probability and Statistical Significance

38

Sir Ronald Aylmer Fisher
(1890-1962)

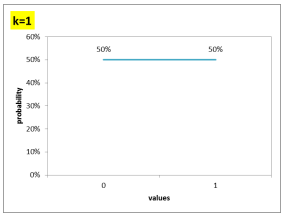


- **AN**alysis **O**f **VA**riance
- Fisher's exact test
- Fisher's z-distribution (F distribution)
- Fisher transformation of r
- Fisher's combined probability test – a meta-analysis technique
- term "variance" in statistics
- variance ratio (F-test)
- phrase „test of significance"
- principles of the design of experiments
- Books:
  - Statistical Methods for Research Workers (1925)
  - Design of Experiments (1935)

39

k=1	SUM
0	0
1	1

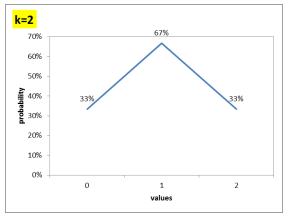
Value	Count	Probability, %
0	1	50%
1	1	50%
TOTAL	2	100%



40

k=2		SUM
0	0	0
1	0	1
0	1	1
1	1	2

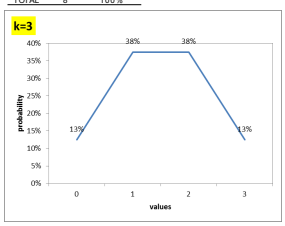
Value	Count	Probability, %
0	1	33%
1	2	67%
2	1	33%
TOTAL	3	100%



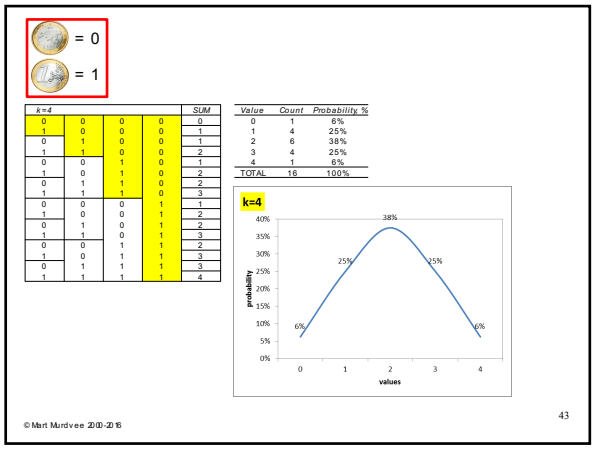
41

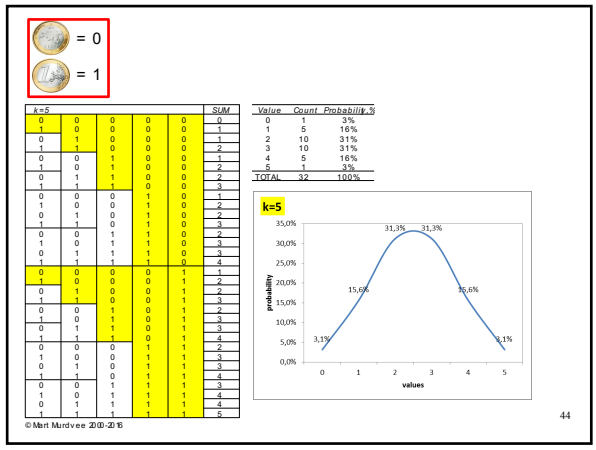
k=3			SUM
0	0	0	0
1	0	0	1
0	1	0	1
1	1	0	2
0	0	1	1
1	0	1	2
0	1	1	2
1	1	1	3

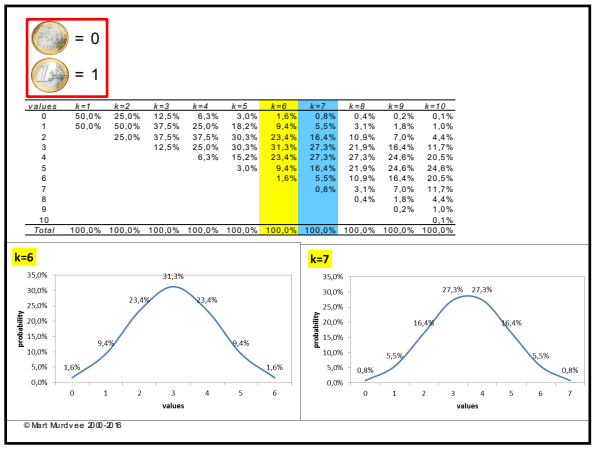
Value	Count	Probability, %
0	1	13%
1	3	38%
2	3	38%
3	1	13%
TOTAL	8	100%

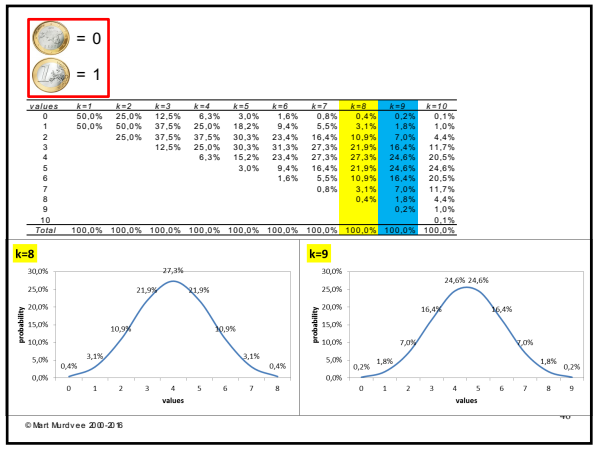


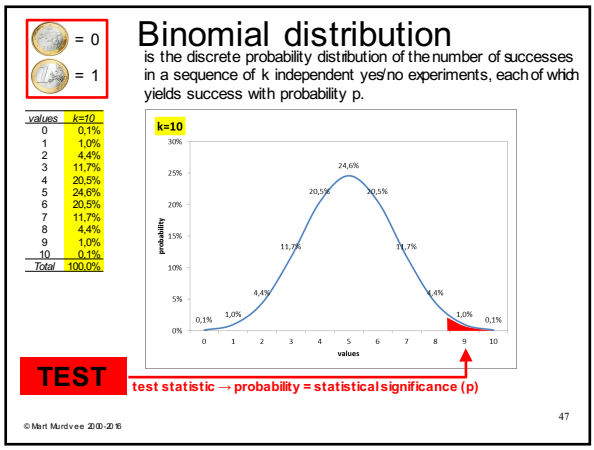
42

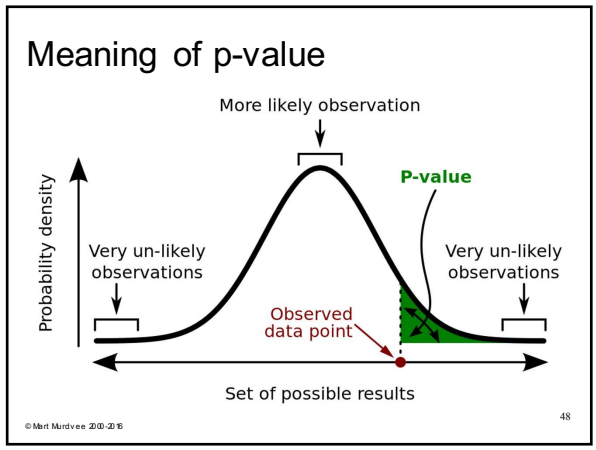




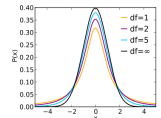




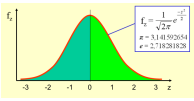




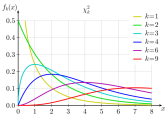
Distributions



- Student's t-distribution



Normal distribution



- Chi-square Distribution
- etc

49

Confidence level, statistical significance

- **Level of confidence *(usaldusnivoo)***
  95% - in 95 times from 100 cases the regularity found in sample is present in population.
  99% - in 99 times from 100 cases the regularity found in sample is present in population.
- **Significance level - α (olulisuse nivoo)**
- **Statistical significance, size of error, p-value *(vea suurus, statistiline olulisus)***
  p = 1 – level of confidence
  (p = 0,05 ... p = 0,01 ... p < 0,01)

NB! The results whitch size of error (p) is equal to or less than 0,05 are conventionally declared as statistically significant.

50

Levels of statistical significance

p ≤ 0,05	borderline statistically significant	*piiripealselt statistiliselt oluline*
p ≤ 0,01	statistically significant	*statistiliselt oluline*
p ≤ 0,005 **p ≤ 0,001**	highly statistically significant	*kõrgelt statistiliselt oluline*

51