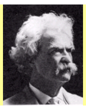


There are three kinds of lies:
lies, damned lies and statistics.

Attributed by Mark Twain to Benjamin Disraeli

Research Methods and Data Analysis - exercise

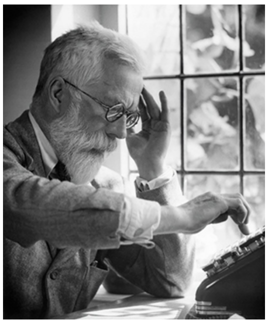
Compiled by
Mart Murdvee

 1

He uses statistics as a drunken man uses a lamppost—for support rather than illumination.
— ANDREW LANG

Giving a school man only a little, or very superficial, knowledge of statistics is like putting a razor in the hands of a baby.
— CARTER ALEXANDER

 2



To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
Ronald Fisher

 3

Прежде чем делать открытие — загляни в справочник.
К. Прутков-инженер. «Советы начинающему гению»

In modern way:
Before You apply for Nobel Price, look into the Wikipedia!

Kozma Prutkov – ingeneur. „Recommendations for start-up geniuses"

 4

5

Levels of phenomenon investigation and use of results

1. **Description:**
   definitions, facts, measures, structures, conditions, etc.
2. **Explaining:**
   regularities, relations, causality, etc.
3. **Prediction:**
   using the knowledge about of the regularities predict future changes
4. **Using (manipulation):**
   using the knowledge about of the regularities induce desired changes

 6

Statistics

- **Statistics (as scientific discipline)** is the study of the collection, organization, analysis, interpretation, presentation of data and making decisions based on data.
- **Statistics (as practice)** is a set of the methods, procedures, rules and customs for collection, organization, analysis, interpretation, presentation of data and making decisions based on data.

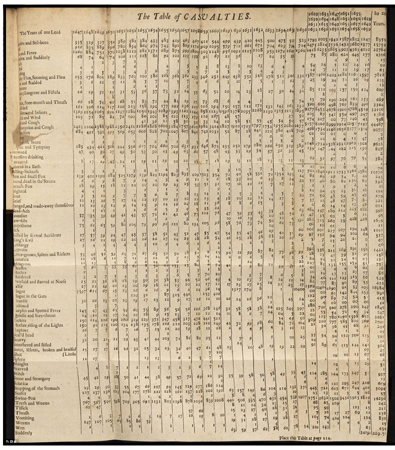
Descriptive and inferential statistics

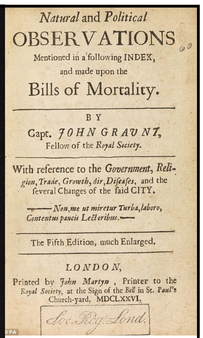
Descriptive statistics (kirjeldav statistika) – describing data in a generalized / simplified, manageable and understandable format:

- resumptive statistics (central tendency, variation, distribution, ...);
- presentation of data - graphs, charts, drawings, tables ...

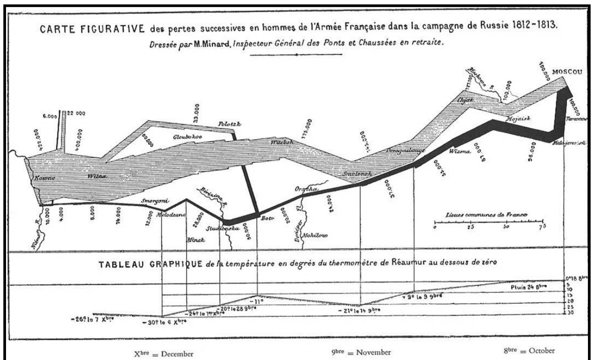
Inferential statistics (järelduslik statistika) - the probability calculations to draw conclusions about the population, to determine the accuracy and reliability of the findings and hypotheses tested:

- reliability of statistics (confidence interval...);
- probability / error of differences or correlations (statistical significance, p-value).



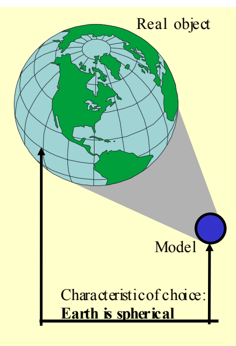


John Graunt (1679) Natural and Political OBSERVATIONS Mentioned in a following INDEX and made upon the Bills of Mortality



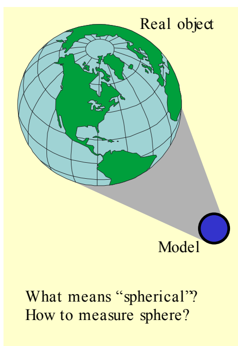
Charles Joseph Minard's (1781–1870) 1861 thematic map of Napoleon's Russian campaign of 1812-13.

Model



- **The model is a description of real world object or phenomenon, which describes the major structures and/or properties of explored object(s).**
- Models are immanently incomplete - model includes only some aspects (characteristics) of the real-world object.
- To build the model, some assumptions about the essential and important structures and relations of real-world objects or events must be made.

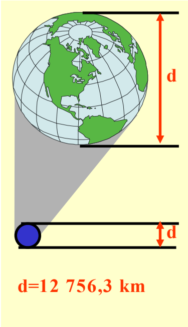
Operationalization



is the process of defining a fuzzy concept so as to make the concept clearly distinguishable or measurable and to understand it in terms of empirical observations.

- First step: defining of the concept
- Operationalization specifies methods for measuring

Types of Data



Ratio data – values can be ranked, interval size is known and is same for the whole range of scale. Scale has absolute zero point. For example - height, weight, age, salary, distance.

Interval data - values can be ranked, interval size is known and is same for the whole interval of scale. For example - IQ, SAT score...

Ordinal data – values can be ranked, but the interval is unknown. For example - education, satisfaction, military rank ...

Nominal data - values can not be ranked and used directly in analysis. For example - gender, nationality, place of residence... Nominal data can be used only for sample grouping.

Dichotomic nominal data (such as gender) can be used directly for the statistical analysis (eg, correlation).

 13

Using Excel

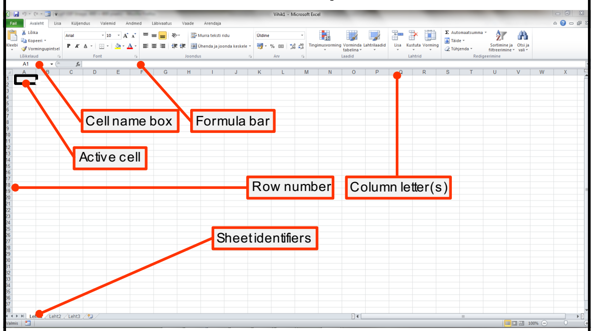
 14

Some useful Windows shortcuts

Press this key	*To do this*
Ctrl+C (or Ctrl+Insert)	Copy the selected item
Ctrl+V (or Shift+Insert)	Paste the selected item
Ctrl+Z	**Undo an action!!!**
Ctrl+Y	Redo an action
Ctrl+A	Select all items (in a document)
Esc	Cancel the current task
Left Alt+Shift	Switch the input language when multiple input languages are enabled

 15

Excel 2010 screen parts



 16

Excel: Managing screen and data

Widening a column
Place your cursor over the column headings (A, B, C etc.) You can drag the edges of the column heading (when the cursor becomes <-|->) to widen the column. If you double-click the edge of the column heading, the column widens to fit the widest piece of data in it.

Inserting a row or column
Select row or column. Right-clic and add. Rows are inserted above your cursor. Columns are inserted to the left of your cursor.

Copying and moving data
Select the cells containing the data with the mouse. Right-clic and Copy (Ctrl-C) to copy the data. Move your cursor to where you want the data to start. Paste (Ctrl-V) or select paste mode. Going to another location and pressing Ctrl-V will copy the data there as well.
The last time you insert the data you can also use Enter instead of Ctrl-V.

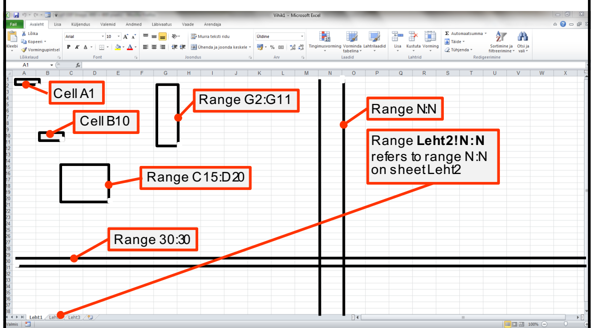
Clearing the data from a row or column
Select the row(s) or column(s) by clicking the numbers to the left and the letters above respectively. Press the Delete key.

Deleting a row or column
Select the row(s) or column(s). Right-click and select Delete.

Keeping columns and rows visible as you scroll
This is useful for keeping column and row labels etc. visible as you scroll around the data.
To "freeze" the top row(s) and left row(s), select the cell where the scrolling starts. Go to View | Freeze Panes.

 17

Excel: cell address, or "cell reference"



 18

Excel: Simple calculations

- **Cells containing numbers for calculations should not contain anything else!**
- Click in the cell where you want the answer to appear.
- Insert an = and then enter numbers or click cells and insert operands:
  - \+ for addition [=2+3]
  - \- for subtraction [=2-1]
  - * for multiplication [=4*5]
  - / for division [=4/2]

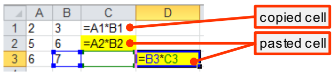
or use equations with cell references →

	A	B	C
1	2	3	=A1+B1
2	2	1	=A2-B2
3	4	5	=A3*B3
4	4	2	=A4/B4

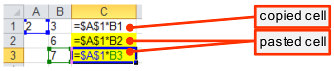
- There is a quick way to add — highlight the data to be added up and click the AutoSum icon (Σ) and select cell range.

Relative and absolute references

- References to cells in Excel are by default "relative references". This means if you copy the cells somewhere else, any formulae in them will still work, but references are relative.



- Using $ sign, the references can be „freezed":
  - $ before column letter „freezes" column
  - $ before row number „freezes" row



Series

	A	B	C
1	1	Q1	esmaspäev
2	2	Q2	teisipäev
3	3	Q3	kolmapäev
4	4	Q4	neljapäev
5	5	Q5	reede
6	6	Q6	laupäev
7	7	Q7	pühapäev
8	8	Q8	esmaspäev
9	9	Q9	teisipäev
10	10	Q10	kolmapäev
11	11	Q11	neljapäev
12	12	Q12	reede
13	13	Q13	laupäev
14	14	Q14	pühapäev
15	15	Q15	esmaspäev
16	16	Q16	teisipäev
17	17	Q17	kolmapäev
18	18	Q18	neljapäev
19	19	Q19	reede
20	19	Q20	laupäev
21	20	Q21	pühapäev

Excel is good at expanding series automatically. Apart from simple numerical series (e.g. 1, 2, 3 ...), Excel knows the days of the week etc.

- Type the first two members of the series in a column or row, select the cells and drag the small black square at the bottom left corner of the selection to the right or down respectively.

Series

	A	B	C
1	1	Q1	esmaspäev
2	1	Q1	esmaspäev
3	1	Q1	esmaspäev
4	1	Q1	esmaspäev
5	1	Q1	esmaspäev
6	1	Q1	esmaspäev
7	1	Q1	esmaspäev
8	1	Q1	esmaspäev
9	1	Q1	esmaspäev
10	1	Q1	esmaspäev
11	1	Q1	esmaspäev
12	1	Q1	esmaspäev
13	1	Q1	esmaspäev
14	1	Q1	esmaspäev
15	1	Q1	esmaspäev
16	1	Q1	esmaspäev
17	1	Q1	esmaspäev
18	1	Q1	esmaspäev
19	1	Q1	esmaspäev
20	1	Q1	esmaspäev
21	1	Q1	esmaspäev

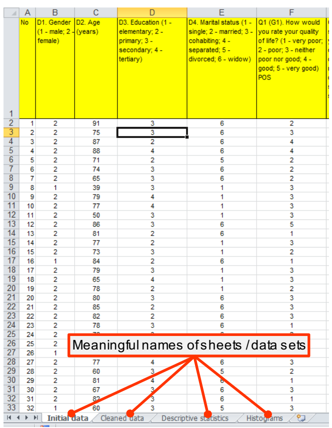
If You does not like series:

- Copy the cell(s) (CTR+C)
- Select range
- Paste cells (CTR+V)

Keep Your research (literature, data, analysis) in order and meaningful!

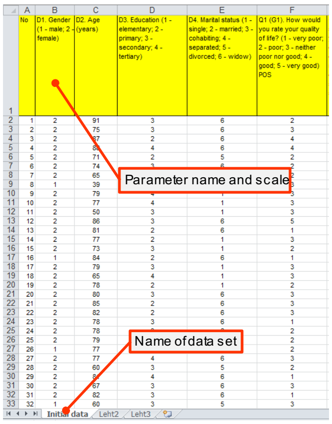
- Folder system and folder names
- File names
- Sheet names
- Parameter / variable names
- Apply a system in names

Record important manipulations and changes!



Practical recommendation for analysis:

- „Clean" data
- Use different sheets.
- Use meaningful labels for variables, sheets, tables, and graphs.



Organising and feeding data

- First row(s): variable names and scales
- First column: case identifier / number
- **Cells used for calculations / analysis must contain ONE NUMBER or ONE CHARACTERISTIC - letter, word(s)**
- ***All variables used in analysis on ONE SHEET OR one data set on one sheet.***

25

When things go wrong:

Questionnaire

35. Kuidas käitute Teie, kui eakas käitub provotseerivalt?

Valige oma kogemustele põhinedes vastav number ja kirjutage see käitumisviisi ees olevale tühjale reale. Kirjutage number iga variandi ette.

1 = väga tihti
2 = tihti
3 = mõnikord
4 = harva
5 = mitte kunagi

____ viin eaka tema tuppa
____ küsin, mis tal viga on
____ ei tee midagi
____ suhtun asjasse huumoriga
____ täpsustan õendus-/hooldusplaanist eaka tausta ja võimalikke juhiseid
____ kaotan kannatuse ja kasutan eaka vaigistamiseks füüsilist jõudu
____ panen eaka istuma geriaatrilisse või tavalisse tooli ning tõkestan tema liikumise nt lauaplaadi või rihmadega
____ kannatan olukorra ära, sest eakal on õigus vihastuda ja vihane olla
____ ei tee midagi kuulma, nägema ega märkama
____ pakun eakale tegevust, nt muusika kuulamist, jalutamist
____ kaotan kannatuse ja käsutan eakat
____ pühendan eakale aega, vestlen temaga, puudutan teda, olen tema lähedal
____ annan rahustavat ravimit
toimite mõnel muul viisil (millisel?)__________________________

Data input

35. Kuidas käitute Teie, kui eakas käitub provotseerivalt?
1,2,5,3,2,5,3,4,4,1,2,2

Remember:

ONE CELL – ONE NUMBER

26

Excel: Elements of formulas

=FUNCTION_NAME(cell;range;number;“text“;function)

- formula begins with equal sign =
- followed by FUNCTION_NAME
- formula operators are in brackets
- formula operators are separated by semicolon sign ;
- formula operators:
  - + ; - ; * ; / ; ^ ; = ; > ; < ; >= ; <= ; <> ;
  - cell „address“
  - range of cells (cell:cell)
  - number
  - „text“ (between quotation marks)
  - function operators
  - nesting functions

27

Arithmetic operators

To perform basic mathematical operations such as addition, subtraction, or multiplication; combine numbers; and produce numeric results, use the following arithmetic operators.

Arithmetic operator	*Meaning (Example)*
+ (plus sign)	Addition (3+3)
– (minus sign)	Subtraction (3–1) Negation (–1)
* (asterisk)	Multiplication (3*3)
/ (forward slash)	Division (3/3)
% (percent sign)	Percent (20%)
^ (caret)	Exponentiation (3^2)

28

Comparison operators

You can compare two values with the following operators. When two values are compared by using these operators, the result is a logical value either TRUE or FALSE.

Comparison operator	*Meaning (Example)*
= (equal sign)	Equal to (A1=B1)
> (greater than sign)	Greater than (A1>B1)
< (less than sign)	Less than (A1<B1)
>= (greater than or equal to sign)	Greater than or equal to (A1>=B1)
<= (less than or equal to sign)	Less than or equal to (A1<=B1)
<> (not equal to sign)	Not equal to (A1<>B1)

29

Text concatenation operator

Use the ampersand (&) to join, or concatenate, one or more text strings to produce a single piece of text.

Text operator	*Meaning (Example)*
& (ampersand)	Connects or concatenates two values to produce one continuous text value (="North"&"wind")

Reference operators

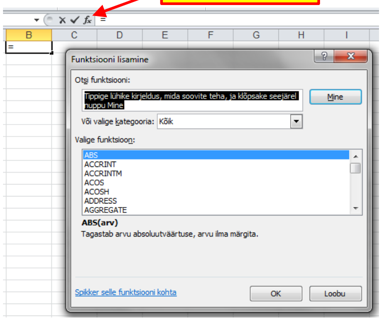
Combine ranges of cells for calculations with the following operators.

Reference operator	*Meaning (Example)*
: (colon)	Range operator, which produces one reference to all the cells between two references, including the two references (B5:B15)
; (semicolon)	Union operator, which combines multiple references into one reference (=SUM(B5:B15;D5:D15)) Separator between operators in formulas (=COUNTIF(B:B;2))

30

Entering formulas

function wizard button



1. Click on a cell where you will paste a function
2. Write formula or Click function wizard button and use function wizard menu.

© Mart Murdvee 2000-2016

31

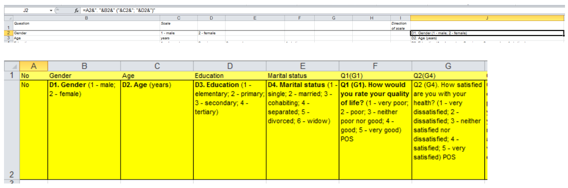
Text compilation

Operator &

or

Function

=CONCATENATE(cell;"text";…)



© Mart Murdvee 2000-2016

32

Making sense & Cleaning data

© Mart Murdvee 2000-2016

33

Common sources of error in databases are:

- missing data coded as "999"
- 'not applicable' or 'blank' coded as "0"
- typing errors on data entry (use Data Validation)
- column shift - data for one variable column was entered under the adjacent column
- fabricated data - 'made up' or contrived
- coding errors
- measurement and interview errors'

Most errors will be detected using three procedures:

- descriptive statistics (MIN; MAX)
- scatterplots
- histograms

© Mart Murdvee 2000-2016

34

Coding missing data

2	2	2
3	3	3
1	1	1
3	3	3
3	3	3
	:	x
4	4	4
5	5	5
3	3	3
3	3	3
2	2	2
2	2	2
1	1	1
3	3	3
3	3	3
3	3	3
4	4	4

- **Recommendation for use of „999" and other „meaningless" number for coding missing data IS FOR FOOLS!**
- Missing values in the database must be marked, and **not left blank**.
- Use "X", ":", etc. - signs, which does not represent anything calculable.
- **NB! Excel plotting and regression functions treat missing values (empty cells) as zeroes.**

© Mart Murdvee 2000-2016

35

Problem: missing data

- Missing data can be meaningful, while it contains information about:
  - questionnaire quality (eg poorly worded question).
  - respondents - some questions can be unpleasant, too sensitive, and therefore, less answered.
- **The person's data set can be considered usable, when more than 80% of the questions are answered.**
- Missing data (in order to obtain a sufficiently large sample for analysis) may be replaced with the parameter average, median, mode, or the parameter value can be derived from other parameters by using regression equation.

© Mart Murdvee 2000-2016

36

How many?

Function	*Does*
=COUNT(range)	Counts the number of cells containing numbers in a range
=COUNTA(range)	Counts the number of non-blank cells within a range
=COUNTBLANK(range)	Counts the number of blank cells within a range
=COUNTIF(range;value)	Counts the number of cells in range that are the same as value / criteria.
=COUNTIFS(criteria_range1; criteria1;[criteria_range2, criteria2];…)	Counts the number of times all criteria are met.

© Mart Murdvee 2000-2016 37

Logical operation IF:
Function:
=IF(logical_test;value_if_true;value_if_false)

Counting empty cells:
Function:
=COUNTBLANK(range)

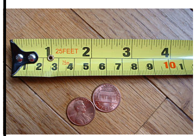
Replace empty cells with mode of a range and let full cells with initial value:
=IF(COUNTBLANK(reference cell)=1;MODE(range);reference cell)

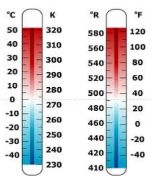
Counting „:“-marked cells:
Function:
=COUNTIF(range;“:“)

Replace „:“-marked cells with mean of a range and let not-„:“-marked cells with initial value:
=IF(COUNTIF(reference cell;“:“)=1;AVERAGE(range);reference cell)

© Mart Murdvee 2000-2016 38

Scale





a gradation pattern of measurement (rule), which determines the numeric value of the result of measurement.

Recommendation:
Use standard and controlled scales!

© Mart Murdvee 2000-2016 39

Direction of scale

Confusing Scale!

Country	Freedom on the Net Status	Freedom on the Net Total 0-100 Points
Estonia	Free	10
USA	Free	13
Germany	Free	16
Australia	Free	18
UK	Free	25
Italy	Free	26
South Africa	Free	26

Freedom House
FREEDOM ON THE NET 2011
A Global Assessment of Internet and Digital Media

It is desirable that the numerical values of the scale corresponds to the concept which is studied. A higher rate or more of something should correspond to the bigger numbers. Otherwise, the results are difficult to comprehend.

- For example, the concept of measuring the "exhaustion rate" with question, "I feel used up at the end of the workday," should have numeric scale direction:
  1 - never, 2 - sometimes, 3 - always
  NOT: 3 - never, 2 - sometimes, 1 - always

© Mart Murdvee 2000-2016 40

Inversion of the scale values:

Satisfaction (5 - completely dissatisfied, 4 - dissatisfied, 3 - do not know, 2 - satisfied, 1 - completely satisfied)	Satisfaction (1 - completely dissatisfied, 2 - dissatisfied, 3 - do not know, 4 - satisfied, 5 - completely satisfied)
initial value	*inverted value*
1	5
2	4
3	3
4	2
5	1

$X_{inverted} = (Min_{scale} + Max_{scale}) - X$

- when the numeric value of scale does not meet the direction of the concept meaning,

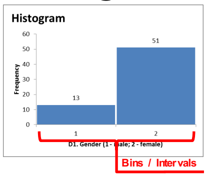
  or
- the value of the parameter must be included in to the factor with inverted value

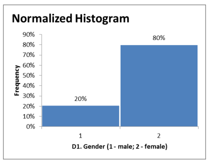
© Mart Murdvee 2000-2016 41

Distribution of values
Histogram

© Mart Murdvee 2000-2016

Histogram





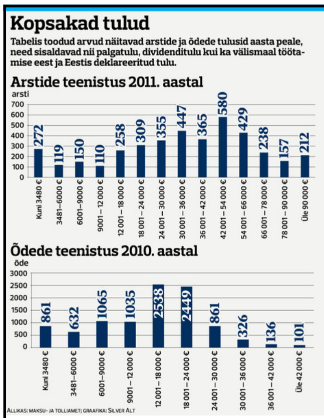
is a graphical representation showing a visual impression of the distribution of data.

First introduced by Karl Pearson.

- A histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval.
- The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval.
- The total area of the histogram is equal to the number of data.
- The categories are specified as **non-overlapping intervals** of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the **same size**.
- A histogram may also be normalized - displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1.

43

© Mart Murdvee 2000-2016

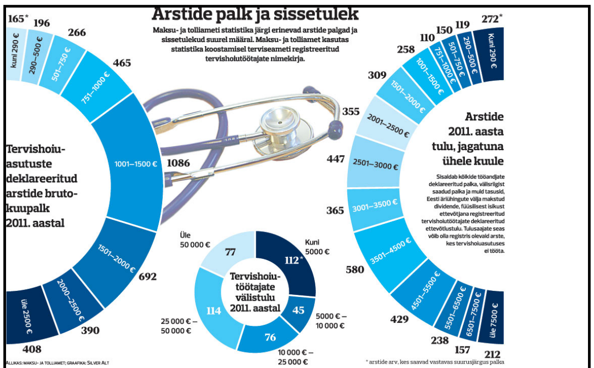


What to think about this kind of „analysis"?

http://www.tarbija24.ee/992036/tipparsti-teenistus-algab-7500-eurost-kuus

44

© Mart Murdvee 2000-2016



What to think about this kind of „analysis"?

45

© Mart Murdvee 2000-2016

Interval widths (bin widths), some "rules of thumb„

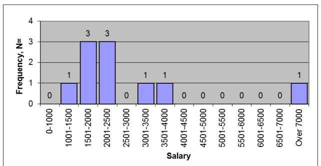
- Sturgis's rule: the number of intervals is close as possible to $1 + \log_2(N)$, where $\log_2(N)$ is the base 2 logarithm of the number of observations. The formula can also be written as $1 + 3.3 \log_{10}(N)$ where $\log_{10}(N)$ is the logarithm base 10 of the number of observations.
  =1+LOG(N;2) or =1+3,3*LOG(N;10)
- Rice rule, which is to set the number of intervals to twice the cube root of the number of observations.
  =2*POWER(N;1/3)

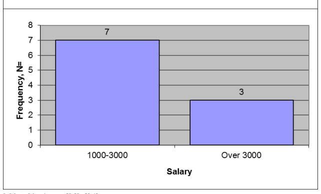
Basic rule:
Experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

46

© Mart Murdvee 2000-2016

Playing with interval widths





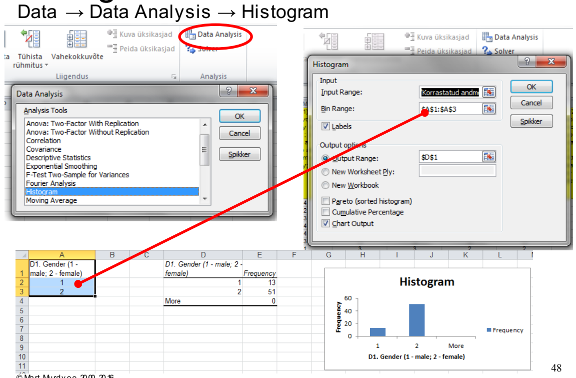
Same sample presented using different interval widths.

47

© Mart Murdvee 2000-2016

Histogram

Data → Data Analysis → Histogram



48

© Mart Murdvee 2000-2016

Histogram

Function:

=FREQUENCY(data_array;bins_array)

- Calculates how often values occur within a range of values, and then returns a vertical array of numbers.
- FREQUENCY is entered as an array formula (Shift+CTR+Enter) after you select a range of adjacent cells into which you want the returned distribution to appear.

C2 {=FREQUENCY('Korrastatud andmed'!B:B;B2:B3)}

	A	B	C
1	Gender	Bin	N
2	1 - male	1	13
3	2 - female	2	51

Histogram

Function:

=COUNTIF(range;criteria)

Counts the number of cells in range that are the same as value.

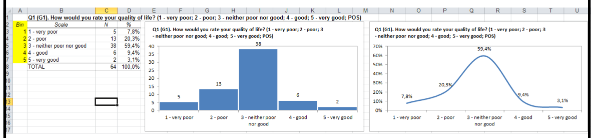
TREND =COUNTIF('Korrastatud andmed'!B:B;A3)

	A	B	C	D
1		D1. Gender (1 - male; 2 - female)		
2	Bin	Scale	N	%
3	1	1 - male	!B:B;A3)	20,3%
4	2	2 - female	51	79,7%
5		TOTAL	64	100,0%

Histogram

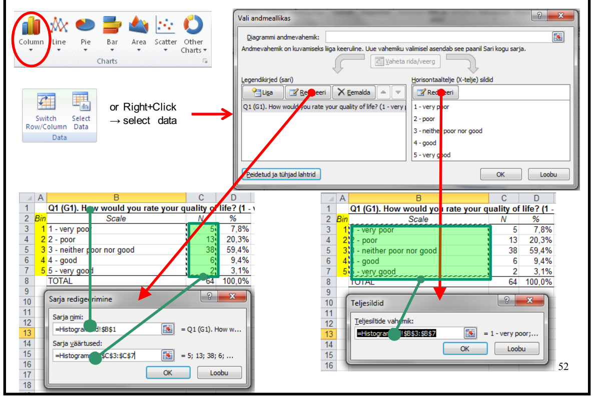
Function:

=COUNTIF(range;criteria)



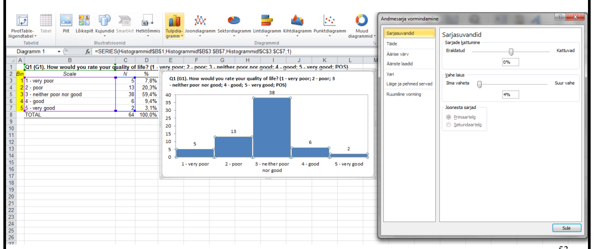
	A	B	C	D
1		='Korrastatud andmed'!F$1		
2	Bin	Scale	N	%
3	1	1 - very poor	=COUNTIF('Korrastatud andmed'!F:F;Histogrammid!A3)	=C3/C8
4	2	2 - poor	=COUNTIF('Korrastatud andmed'!F:F;Histogrammid!A4)	=C4/C8
5	3	3 - neither poor nor good	=COUNTIF('Korrastatud andmed'!F:F;Histogrammid!A5)	=C5/C8
6	4	4 - good	=COUNTIF('Korrastatud andmed'!F:F;Histogrammid!A6)	=C6/C8
7	5	5 - very good	=COUNTIF('Korrastatud andmed'!F:F;Histogrammid!A7)	=C7/C8
8		TOTAL	=SUM(C3:C7)	=C8/C8

Chart

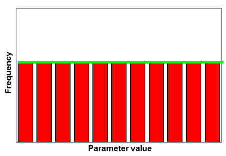


Designing Chart Elements

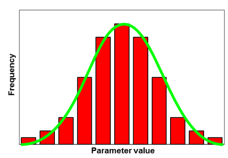
1. Select chart element
2. RightClick → Format …
3. Use menu



Interpreting histogram

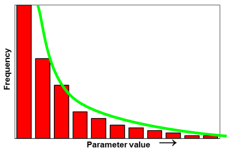


- Parameter values are uniformly distributed. Ideal random process. No trend.

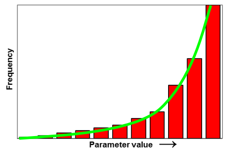


- There is a central tendency. Normal distribution.

Interpreting histogram



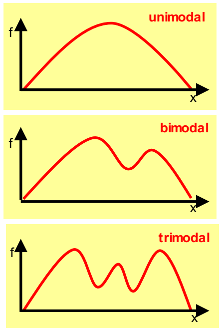
- a rare phenomenon
  - *Maybe the range of the scale is too small?*
  - *Logarithmic distribution?*



- common phenomenon
  - *Maybe the range of the scale is too small?*
  - *Square (power) distribution?*

55

Distribution of parameter
Modality
modaalsus

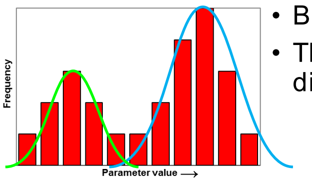


refers to the number of distinct peaks, or areas of cluster, that appear within a distribution, with each such peak being spoken of as a mode.

- A distribution with only one distinct peak is described as unimodal *(unimodaalne)*;
- if it has two distinct peaks it is spoken of as bimodal *(bimodaalne)*;
- three peaks, trimodal *(trimodaalne)*;
- and so on.

56

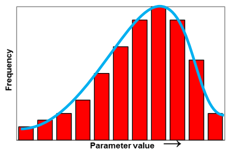
Interpreting histogram



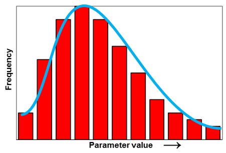
- Bimodal distribution.
- There can be two distinctive groups.

57

Interpreting histogram



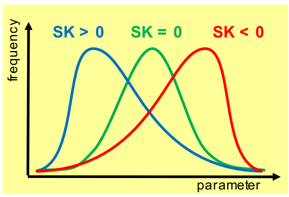
- Distribution skewed toward greater values



- Distribution skewed toward smaller values

58

Distribution of parameter
Skewness
Asümeetriakordaja



Pearson's 1. skewness coeficent:

$$\boxed{SK = \frac{\bar{x} - M_o}{SD}}$$

M_o - mode

Pearson's 2. skewness coeficent:

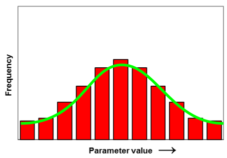
$$\boxed{SK = \frac{3(\bar{x} - M_d)}{SD}}$$

M_d - median

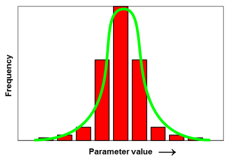
- **a measure of symmetry, or more precisely, the lack of symmetry of distribution:**
  - **SK > 0** – data that are skewed left, more smaller values;
  - **SK = 0** – symmetrical;
  - **SK < 0** – data that are skewed right, more greater values.
- =SKEW(range)

59

Interpreting histogram



- „Flat“ distribution – lot of small and large values (heavy and long tails – *(jaotuse)sabad*). Weak central tendency



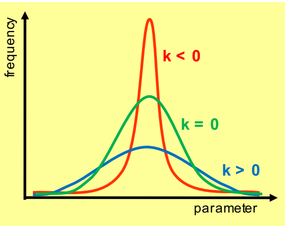
- „Peaked“ distribution – few small and large values (thin and short tails). Strong central tendency

60

Distribution of parameter

Kurtosis

järskuskordaja, ekstsess



$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3$$

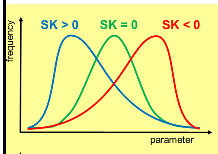
the peakedness of a graph of the distribution compared to normal distribution:

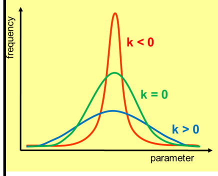
- **k > 0** – extreme values, "long tails"
- **k = 0** – normal distribution;
- **k < 0** – few extreme values, „short tails"

• =KURT(range)

61

Rule of thumb about normality of variable





PROBLEM! Many of the statistical methods (Z test, t-test, ANOVA, regression, covariance, correlation, chi-square test, F test) require the assumption that a variable or variables are normally distributed. Otherwise the statistical methods will produce totally invalid results. (QUESTIONABLE!)

A variable is reasonably close to normal if its skewness and kurtosis have values between –1,0 and +1,0.

- When a variable is not normally distributed, we can create a transfomed variable and test it for normality. If the transformed variable is normally distributed, we can substitute it in our analysis.
- Three common transformations are: the logarithmic transformation, the square root transformation, and the inverse transformation.
- OR: use non-parametric tests

62

Chi-squared (χ^2) test

Hii-ruut test

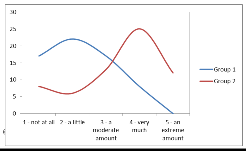
tests a 0-hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical or other distribution.

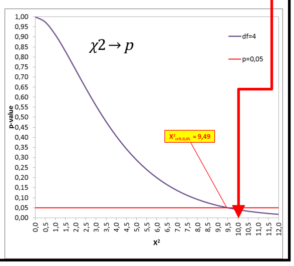
NB! Chi-square procedures can be applied only if all expected values of are equal to or greater than 5.

$$X^2 = \sum_{i=1}^{n} \frac{(O_{bserved} - E_{xpected})^2}{E_{xpected}}$$

df = (number of categories)-1

Distributions	*Observed*	*Expected*
Scale	*Group 1*	*Group 2*
1 - not at all	17	8
2 - a little	22	6
3 - a moderate amount	17	13
4 - very much	8	25
5 - an extreme amount	0	12





Chi-squared (χ^2) test

Hii-ruut test

Obs.	j = 1	j = 2	j = …	j = c
i = 1	i1j1	i1j2	i1j…	i1jc
i = 2	i2j1	i2j2	i2j…	i2jc
i = …	i… j1	i… j2	i… j…	i… jc
i = r	irj1	irj2	irj…	irjc

Exp.	j = 1	j = 2	j = …	j = c
i = 1	i1j1	i1j2	i1j…	i1jc
i = 2	i2j1	i2j2	i2j…	i2jc
i = …	i… j1	i… j2	i… j…	i… jc
i = r	irj1	irj2	irj…	irjc

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:
O_{ij} = observed frequency in the i-th row, j-th column
E_{ij} = expected frequency in the i-th row, j-th column

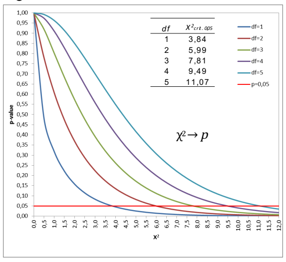
$$df = (r - 1)(c - 1)$$

r = number of rows
c = number of columns

Function:

=CHITEST(observed_range;expected_range)

returns the p-value from the chi-squared (χ^2) distribution for the statistic and the appropriate degrees of freedom.



64