Principles of Biostatistics and Informatics

2nd Lecture: Descriptive Statistics

13th September 2016

Dániel VERES

Tastitsticsss? What's that?

Statistics describes **random mass** phanomenons.

- ***Data Collecting (Sampling)***
- ***Data Organization***

Descriptive Statistics

- ***Data Analysis***
- ***Conclusion***

Inferential Statistics *(Inductive)*

Tastitsticsss? What's that?

Statistics describes **random mass** phanomenons.

- Data Collecting (Sampling)
- Data Organization

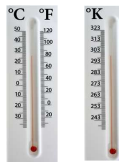
Descriptive Statistics

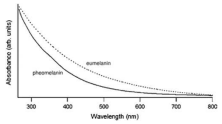
- Data Analysis
- Conclusion

Inferential Statistics (Inductive)

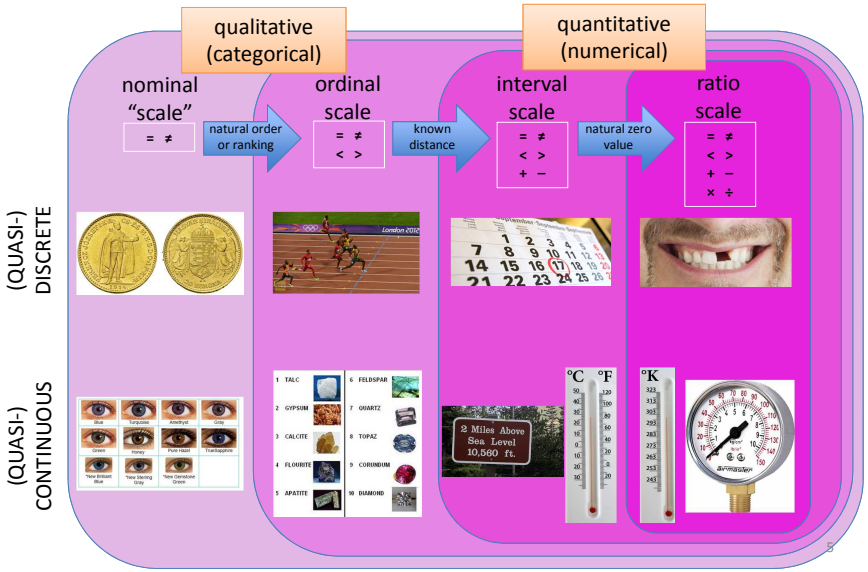
Variables, outcomes

Could be measured or observed





Variable Types:
Levels of Measurement



Description of Nominal Variables I.

Numerical (analytical)

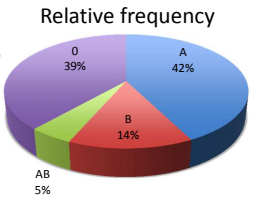
List

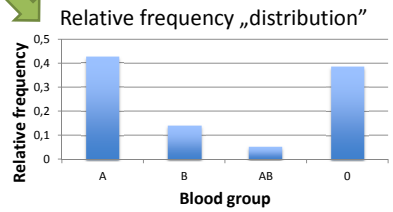
patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	147
3	B	169
4	B	159
5	B	150
6	B	167
7	A	144
8	B	158
9	AB	177

Frequency table

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Graphical





Univariate organization – without loosing information

Description of Nominal Variables II.

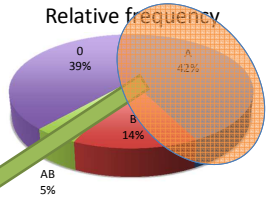
Numerical

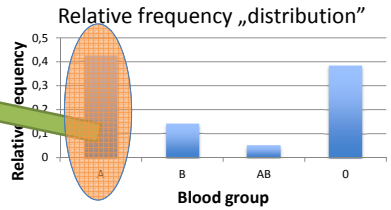
Frequency table

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Graphical

The brain and the common sense





Organization, but **loss of information**

„Typical value" (***indicator***): ~~**Mean?!**~~

Mode**:** most frequent element(s)

Notation: *Mod*, x_{mod}

Other parameters:

data count (*n*)***, count of categories***

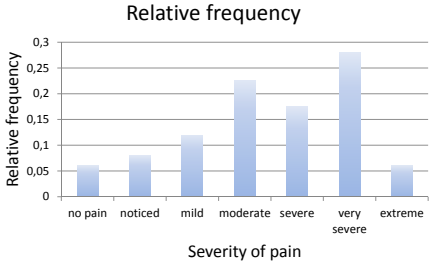
Description of Ordinal Variables I.

Numerical

Frequency table

Severity of pain	Relative frequency	Cumulative relative frequency
no pain	0,06	0,06
noticed	0,08	0,14
mild	0,12	0,26
moderate	0,225	0,485
severe	0,175	0,66
very severe	0,28	0,94
extreme	0,06	1
Σ	1	

Graphical



Indicator:
Mode

Other parameters:
data count (*n*)***, count of categories***

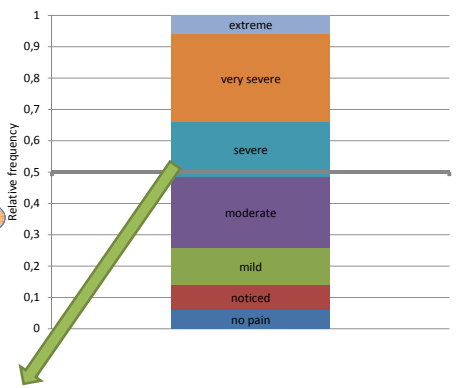
Description of Ordinal Variables II.

Numerical

Frequency table

Severity of pain	Cumulative relative frequency
no pain	0,06
noticed	0,14
mild	0,26
moderate	0,485
severe	0,66
very severe	0,94
extreme	1
Σ	

Graphical



New indicator:
Median: „middle" element(s)
Notation: Me, Med, x_{med}

Description of Quantitative Variables I.

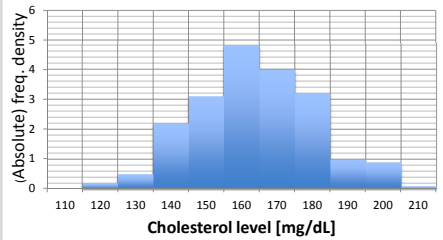
Numerical (analytical)

Frequency tables

frequency distributions (differential discrimination functions)				
bins (classes, intervals)	(absolute) frequency (FREQUENCY)	relative frequency	(absolute) frequency density	relative frequency density
x ≤ 100	0			
100 < x ≤ 110	0	0	0	0
110 < x ≤ 120	2	0,01	0,2	0,001
120 < x ≤ 130	5	0,025	0,5	0,0025
130 < x ≤ 140	22	0,11	2,2	0,011
140 < x ≤ 150	31	0,155	3,1	0,0155
150 < x ≤ 160	48	0,24	4,8	0,024

Graphical

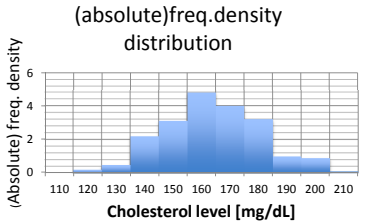
(absolute)freq.density distribution

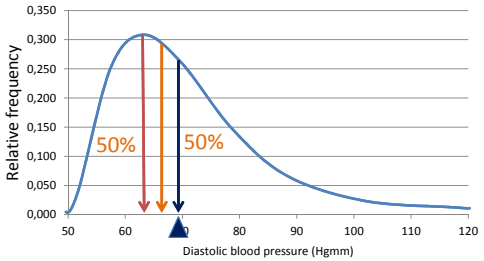


Organizing data – with **loss of information**

Determination of bin width:
- technical and aesthetic concerns
- statistical concerns

Description of Quantitative Variables II.





„Typical values" – ***central tendencies*** (special ***measures of location***):
- Mode: most frequent element(s) ?
- Median: „middle" element(s?)
- ***Mean*** (arithmetic mean): „gravity center" , sensitive to „outliers"?
  Notation: x_{mean}, $\overline{x}$

Advantage: compact, ***could be determined from few data***

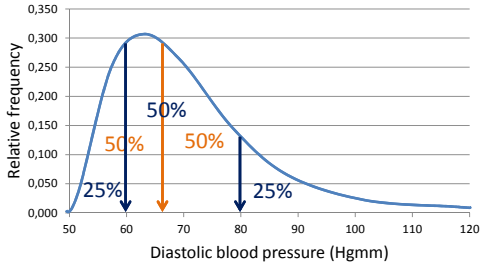
Formulas: in the formula collection...

Digress I.

Average ≠ Mean

In statistics the average could means:
mode,
median,
means – arithmetic, geometric, harmonic... mean

Quantiles I.



Other measures of location:

- Median: 50-50% (Q_2)
- ***Quartile***: lower quartile(Q_1): 25-75%; upper quartile (Q_3): 75-25%

General

p-quantile(s): is the number to which the count of data are smaller is maximum *n*p* and to which the count of data are larger is maximum *n*(1 – p)*,

where p is between 0 and 1, and *n* is the count of data

Digress II.

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	7,72
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	8,31
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	20		

Median, quantiles could differ in theory and practice.
Mean is sensitive to the outliers, but quantiles not (...).
Mode?

Digress III.

$$\frac{1}{n}\sum\left|x_i - x^*\right|$$

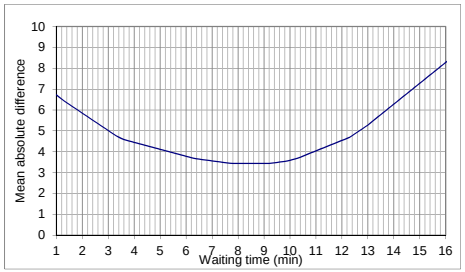
Minimal if:

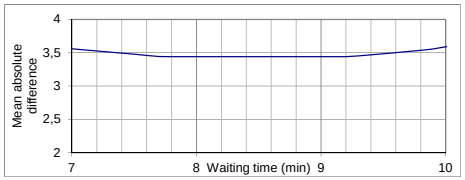
$$x^* = Median$$

$$\frac{1}{n}\sum\left(x_i - x^*\right)^2$$

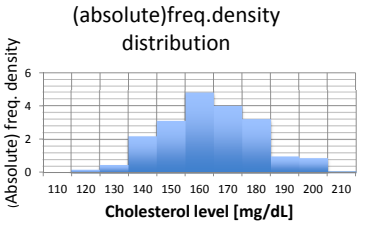
Minimal if:

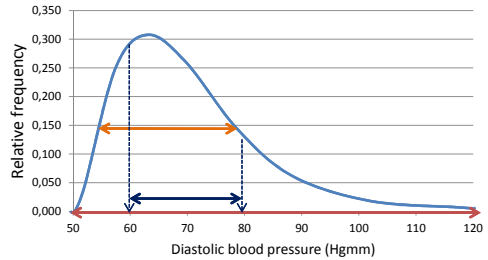
$$x^* = Mean$$





Description of Quantitative Variables III.



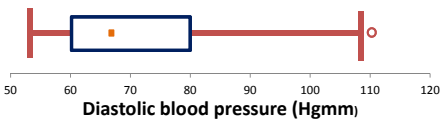


Measures of spread:

- ***Range***: the difference between the maximum and the minimum
- ***Variance (s^2):*** the average of the squared distance from the mean (corrected - sample, uncorrected - population)
- ***Standard deviation (s, sd, SD)***: the square root of the variance the width of the curve
- ***Interquartile range (IQR)*** : the difference between the upper and the lower quartile – not sensitive to the „outliers”

Description of Quantitative Variables IV.

Graphical: Box plot



Middle point: mean, or *median*
Box: 2*standard deviation, or *interquartile range,* p-quantile range
Whisker: 3*SD, minimum and maximum, 0.05 and 0.95 quantiles, p-quantiles, *1.5*IQR*...
out of whiskers: ***outliers***

Trimmed mean: mean calculated without outliers

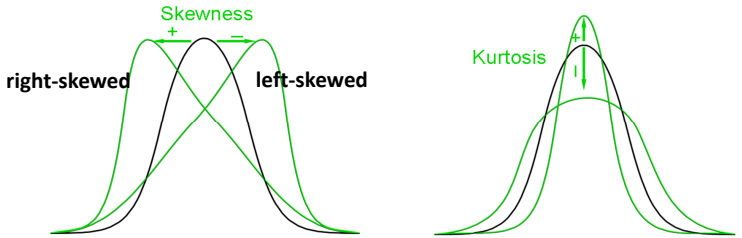
Description of Quantitative Variables V.

Other parameters:

- ***moment***:
  the k-th moment: $\Sigma(x_i)^k/n$
- ***central moment***:
  the k-th central moment: $\Sigma(x_i-\mu)^k/n$

- ***skewness***,
- ***kurtosis***

} ***measures of shape***

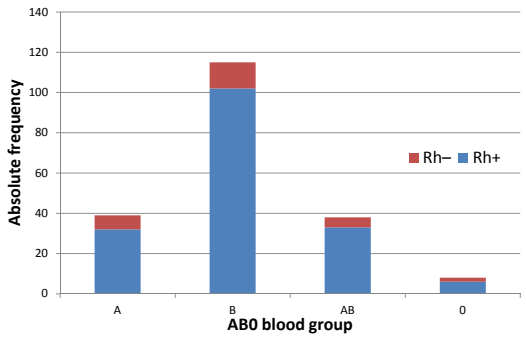


Qualitative Bivariate Description

Numerical: ***contingency*** table

	A	B	AB	0	Σ
Rh+	32	102	33	6	173
Rh–	7	13	5	2	27
Σ	39	115	38	8	200

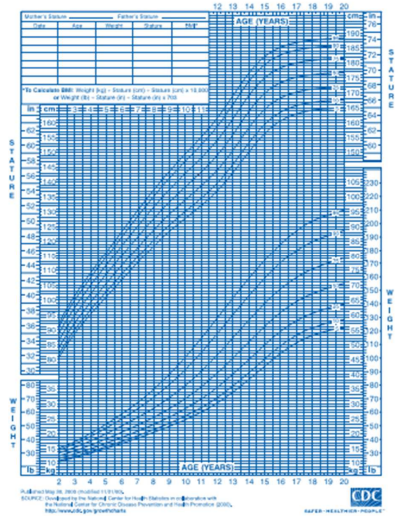
Graphical: ***stacked bar chart***

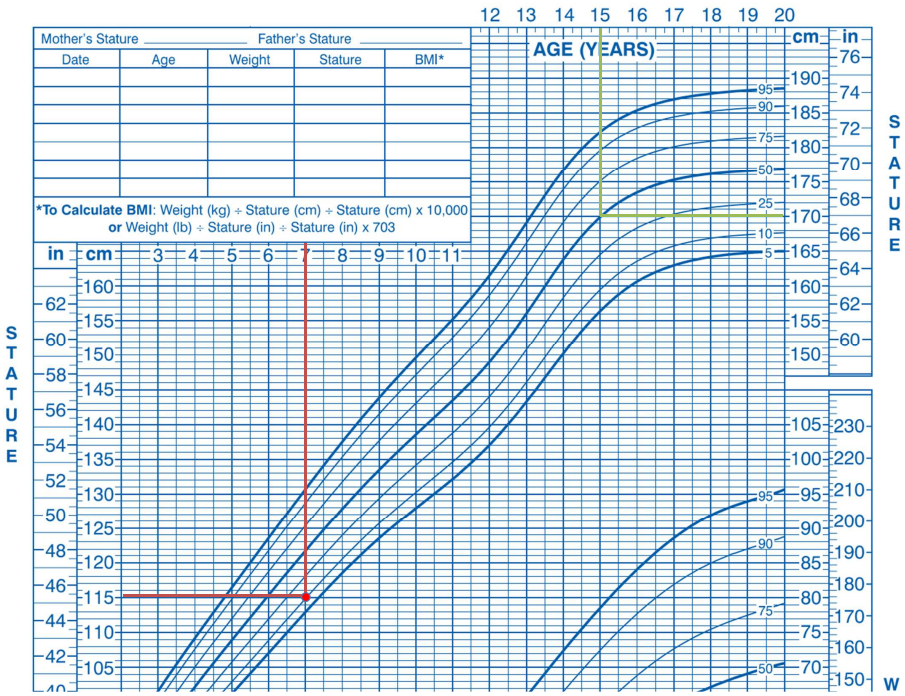


Quantitative Bivariate Description

Graphical: ***percentile curves***

Percentile: quantile expressed as percentage





Data Collection Thoughts

Data collection is motivated by a goal and not by a variable.
Use the highest measurement level as possible.

Record data

- use a form that is easy to organize and convert - *excel*
- *variables in separated cells*
- *Coding have to be clear* (type of variable, categories)

Test Questions #1

- Give the four actions of statistics.
- Give the two part of descriptive statistics.
- Give the two part of inferential statistics.
- Name some ordinal variables, scales.
- Name some discrete numerical variables, scales.
- Name some continuous numerical variables, scales.
- What is the substantial difference between a nominal and an ordinal scale?
- Give example for interval scale.
- What is the substantial difference between an ordinal and an interval scale?
- Give examples for ratio scale.
- What is the substantial difference between an interval and a ratio scale?
- Why is it important to define a statistical variable properly?
- What are the two way as we could describe a variable?
- What are the indicators that we can use to describe a nominal variable?
- What are the indicators that we can use to describe an ordinal variable?
- What are the indicators that we can use to describe a numerical variable?
- Define the mode(s) of a dataset.
- What is the notation of mode?
- Define the median(s) of a dataset.
- What is the notation of median?
- In which type of measurement scale do we loose information usually?
- How we can determine the bin width?
- What is the equation we have to use to determine the bin width?
- What are the central tendencies in case of a numerical variable?
- What is the „meaning" of the mode in a diagram?
- What is the „meaning" of the median in a diagram?
- What is the „meaning" of the mean in a diagram?
- Define the mean of a dataset.
- What is the notation of mean?
- Which central tendency sensitive to outliers?
- What is the advantage of indicators versus distribution functions?
- What is the difference between average and mean?
- What are the measures of location?
- Define the p-quantile.
- Define the lower quartile.
- What is the difference between the second quartile and the median?
- Show how we could calculate the lower quartile of a dataset in theory and in practice.
- What is the value that for the sum of the absolute differences are minimal?
- What is the value that for the sum of the squared differences are minimal?

Test Questions #2

- What are the measures of spread?
- What are the measures of shape?
- Define the variance.
- Define the standard deviation.
- Define the skewness.
- Define the kurtosis.
- Define the interquartile range.
- What is the notation of interquartile range?
- What is a box plot?
- What are the parts of a box plot?
- What we could use as a middle point of a box plot?
- What we could use as a box of a box plot?
- What we could use as a whisker of a box plot?
- What is recommended middle point in a box plot if we have a non symmetrical distribution with outliers?
- What is recommended box boundary in a box plot if we have a non symmetrical distribution with outliers?
- What is recommended box boundary in a box plot if we used a median as a middle point?
- What is the trimmed mean?
- How we define the outlier range commonly?
- What are the moments?
- What are the central moments?
- What is the first central moment?
- What is the first moment?
- What is the second central moment?
- What are the percentiles?
- What we could read out from a percentile curve?