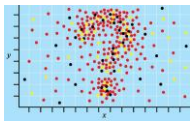
MUNI
PHARM

Introduction to Statistics

Biophysics

21

- **Data**
  - Data is a set of values of qualitative or quantitative variables



- **Population**
  - The group we want to find information about

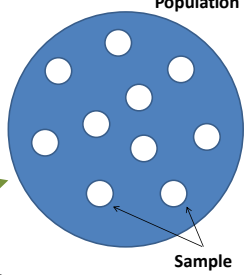
  E.g. the nation, batch produced, people with a certain disease that we want to treat
- **Sample**
  - Part of the population from which we want to find information about the population

22

Population/ Sample

Analyze data across population would be expensive – it is not possible tested whole population. Therefore, only a part of the population is analyzed (Sample).

Selection of subjects (Sample) from basic file (Population) should be random.



Example: Determining of voter preferences.
There is the effect of time and place of data collection:
Morning of working day near the supermarket – old age pensioners have impact – estimation of voter preferences will be inaccurate.

23

Data

- **Qualitative - categorical data**
  - e.g. eye color, breed of dog, ...
- **Quantitative - continuous data**
  - Continuous data can take any value
  - e.g. temperature, weight, height, time, ...

Level of measurement:

cardinal – numbers with which we can perform mathematical operations
ordinal – order (1st, 2nd, 3rd, etc.) by which data can be sorted
nominal – differentiates between items or subjects based only on their names or categories and other qualitative classifications.
The ordinal and nominal measurement is more suitable for eg. sociological research.

24

Quantitative data properties

- **Minimum and maximum**
  - the smallest and the largest value
- **Arithmetic mean (Average)**
  - represents the central tendency in a data set

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \dots)}{n}$$

- **Median**
  - middle value that divides the data set into two halves
  - simply sort values by size and take a value that is in the middle

● odd number

$$\mathrm{Med}(x) = x_{\frac{n+1}{2}}$$

1, 3, 3, **6**, 7, 8, 9
Median = 6

● even number

$$\mathrm{Med}(x) = \frac{x_{n/2} + x_{(n+2)/2}}{2}$$

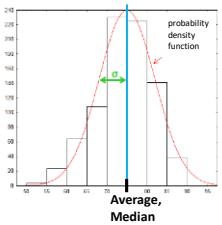
1, 2, 3, **4**, **5**, 6, 8, 9
Median = (4 + 5) ÷ 2 = 4.5

26

Symmetrical data

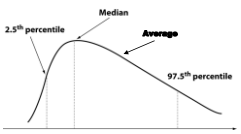
(Gauss normal distribution)

- Average and median correctly describe the data
- Data frequency is the highest around the center

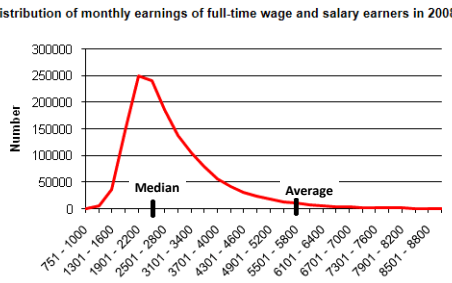


Nonsymmetrical data

- Average can not be used to describe the data
- Often at the medical data, pharmacokinetic studies, etc.
- Verification of normality by tests or graphically (statistical software)



27



- **The median** is the true indicator of salary size (it is a robust parameter).
- **The average** is shifted to higher values; It is significantly affected by high salaries of a small group of people.

28

Quantitative data properties

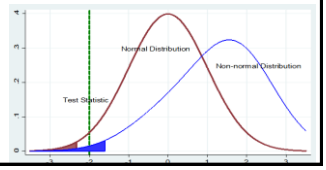
- **Data transformation**

Linear transformation (e.g. multiplication by a constant): **weighting**. The process of weighting involves emphasizing the contribution of some aspects of a set of data to a final effect or result.

Weighted arithmetic mean $\bar{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n}$

Non-linear transformation

-changes the structure of the data



29

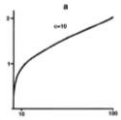
Quantitative data properties

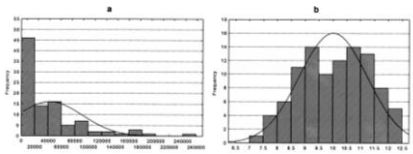
- **Non-linear data transformation**

Logarithmic transformation

Measured values are replaced by their logarithms

$x'_{ij} = \log_e x_{ij}$





30

Quantitative data properties

- **Non-linear data transformation**

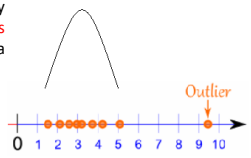
Square root transformation

Exponential transformation

Box-Cox transformation

The reasons for the transformation: data are not normally distributed (nonsymmetrical data) or data contain outliers.

1. Data not normally distributed => Data transformation (or median)
2. Data normally / not normally distributed; data contain outliers which can not be excluded => Data transformation (or median)



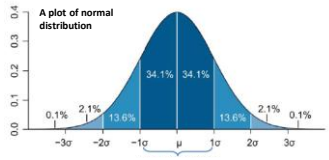
31

Quantitative data properties

- **Standard deviation (σ)**
  - Represents the variability in the data set
  - For a sample is called the standard deviation of sample, see function in Excel (**STDEV**)

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots}{n-1}}$$

Set of random observation tends to be situated around the average value.



68.3% of all measurement

2σ = 95.6% of all measurement
3σ = 99.7% of all measurement

32

Quantitative data properties

- **Confidence interval (CI)**
  - The interval which contain the true measured parameter (e.g. mean)
  - CI are constructed using a given confidence level (usually 95%)
  - For a sample is called the **CONFIDENCE.T**, see function in Excel

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

1.96 is valid for 95% confidence interval (n ≥ 30).

Simplified evaluation of CI (for small samples):

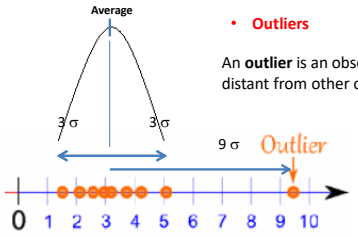
$\bar{x} \pm (K_n \times R)$

RANGE (R) = MAXIMUM - MINIMUM

n	K_n (95% confidence level)
2	6.4
3	1.3
4	0.72
5	0.51
6	0.40

33

Quantitative data properties



- **Outliers**

An **outlier** is an observation point that is distant from other observations.

An outlier may be due to **variability** in the measurement or it may indicate **experimental error**; the are sometimes excluded from the data set **if they are greater than 3*σ**.

Outliers can not be excluded e.g. for medical data (or other data, where outliers are an integral part of the sample)!

34

Outliers

- **Identification of outliers for small samples: Dixon's Q-test**

Tests the first and last value results against neighboring value.

$$Q_n = \frac{x_n - x_{n-1}}{R} \qquad Q_1 = \frac{x_2 - x_1}{R}$$

RANGE (R) = MAXIMUM - MINIMUM

The calculated value is compared with tabulated values. If Q_n or Q_1 > $Q_{tab.}$, than is the extreme result excluded.

n	$Q_{tab.}$ (α = 0.05)
2	-
3	0.941
4	0.765
5	0.642
6	0.560
7	0.507

35