Teaching Mathematics and Statistics in Sciences, IPA HU-SRB/0901/221/088 - 2011

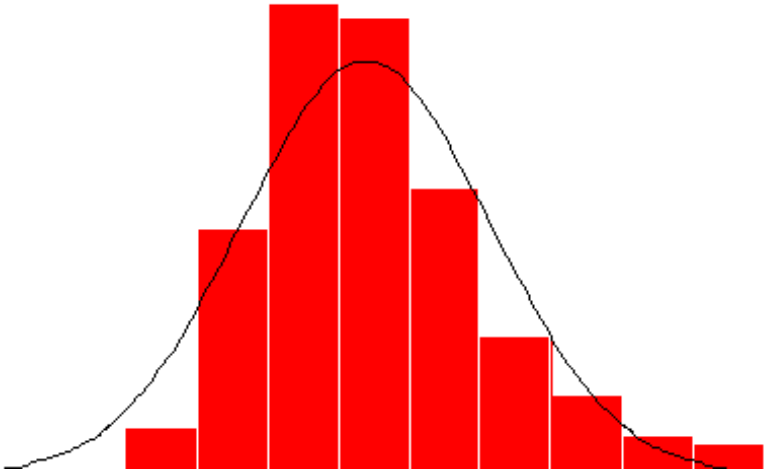
Biostatistics

Author: **Krisztina Boda** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

What is biostatisics?
Basic statistical concepts



Introduction

- All of us are familiar with statistics in everyday life. Very often, we read about sports statistics; for example, predictions of which country is favored to win the World Cup in soccer..
- Regarding the health applications of statistics, the popular media carry articles on the latest drugs to control cancer or new vaccines for HIV. These popular articles restate statistical findings to the lay audience based on complex analyses reported in scientific journals.
- Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.
- Biostatistics (or biometrics) is the application of mathematical statistics to a wide range of topics in biology. It has particular applications to medicine and to agriculture.

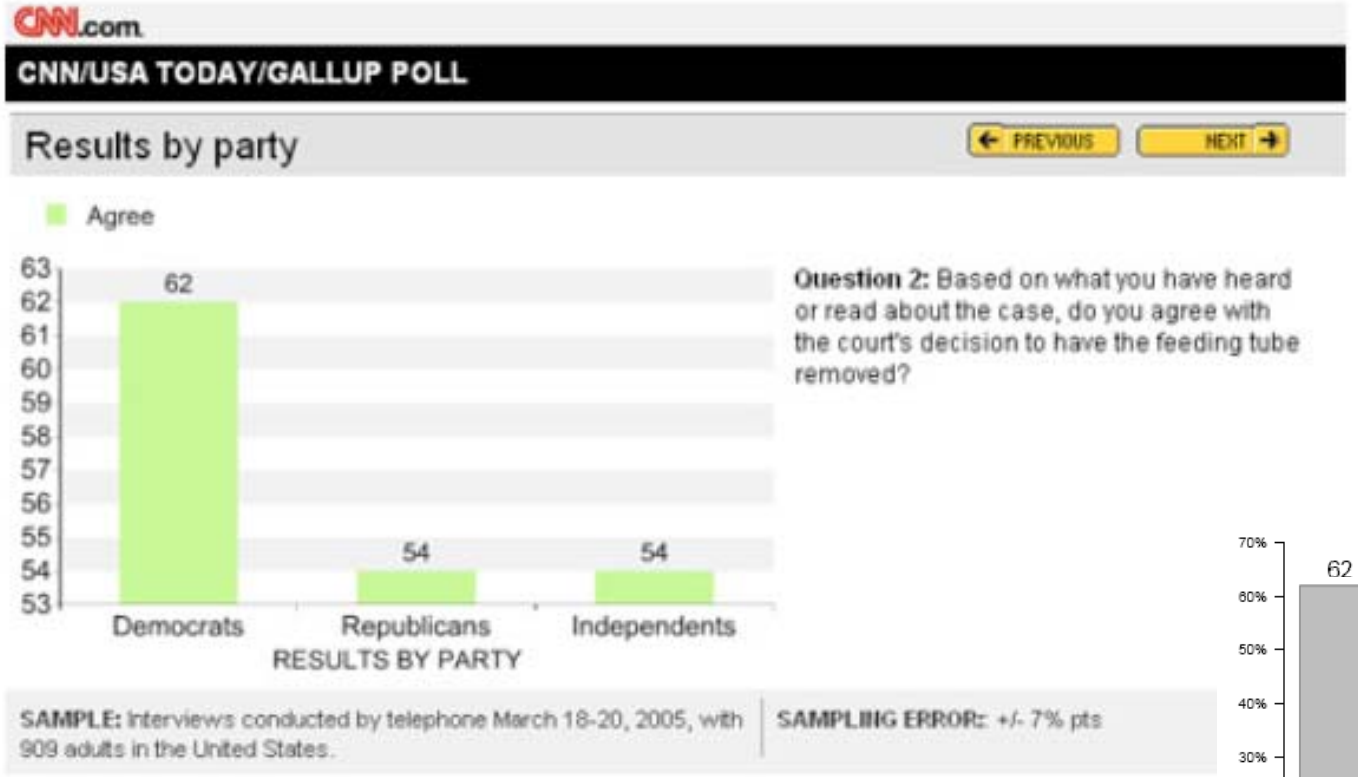
Why study statistics?

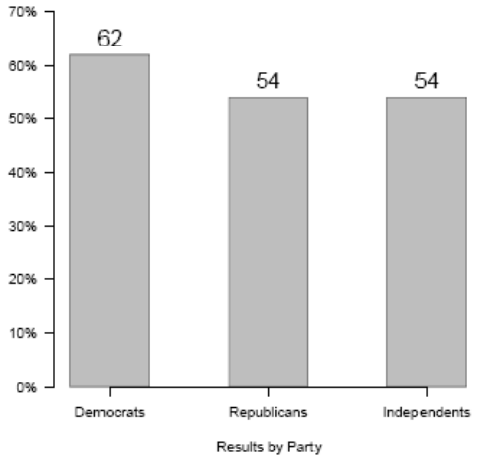
- Understand the statistical portions of most articles in medical journals
- Avoid being bamboozled by statistical nonsense.
- Do simple statistical calculations yourself, especially those that help you interpret published literature.
- Use a simple statistics computer program to analyze data.
- Be able to refer to a more advanced statistics text or communicate with a statistical consultant (without an interpreter).

Misuse of statistics

- **Children with bigger feet spell better?**
- Quite astonished? Don't be! This was the result of a survey about measuring factors affecting the spelling ability of children. When the final analysis came about, it was noted that children with bigger feet possessed superior spelling skills! Upon further analysis you will find that older children had bigger feet and quite certainly, older children would normally possess better spellings than their younger counterparts!

How to lie with statistics





http://www.stats.ox.ac.uk/~konis/talks/HtLwS.pdf

Why study statistics? (ctd)

- Understand the statistical portions of most articles in medical journals
- Avoid being bamboozled by statistical nonsense.
- Do simple statistical calculations yourself, especially those that help you interpret published literature.
- Use a simple statistics computer program to analyze data.
- Be able to refer to a more advanced statistics text or communicate with a statistical consultant (without an interpreter).

About this course

- Medical physics and statistics
- The **Biostatistics** lecture course provides students with an advanced practical knowledge in biostatistics. With conceptual understanding of data and data collection, we introduce techniques of data processing, representation and interpretation. We cover topics of trend analysis, use of hypotheses, frequently used statistical tests and their applications.
- Knowledge of elementary mathematics is required. The main purpose is teaching students how to find the most appropriate method to describe and present their data and how to interpret results.

- There is a five-grade **written exam** at the end of both semesters.

- Lecture notes can be downloaded:
- http://www.szote.u-szeged.hu/dmi/

- **For a better understanding, we suggest the attendance of the compulsory elective practical course, Biostatistical calculations (2 hours/week) accompanying the 1 hour/week Biostatistics lecture.**

Biostatistical calculations
Compulsory elective practical course

- **Practice:** 2 lessons per week
  Form of examination: practical mark
  Year/semester: 1st year, 1. semester
  Credits: 2
- The subject is designed to give basic biostatistical knowledge commonly employed in medical research and to learn modelling and interpreting results of computer programs (SPSS). The main purpose is to learn how to find the most appropriate method to describe and present their data and to find significant differences or associations in the data set.
  Attendance of the course facilitates the accomplishment of the obligatory course "Medical physics and statistics".

- Data sets
  - Data about yourself
  - Real data of medical experiments

- **Forms of testing:** The students have to perform two tests containing practical problems to be solved by hand calculations and by a computer program (EXCEL, Statistica or SPSS). During the tests, use of calculators, computers (without Internet) and lecture notes are permitted. Final practical mark is calculated from the results of the two tests.

Application of biostatistics

- Research
- Design and analysis of clinical trials in medicine
- Public health, including epidemiology,
- …

Biostatistical methods

- Descriptive statistics
- Hypothesis tests (statistical tests)
  - They depend on:
    - the type of data
    - the nature of the problem
    - the statistical model

Descriptive statistics, example

BMJ

Objectively monitored patching regimens for treatment of amblyopia: randomised trial

Catherine E Stewart, David A Stephens, Alistair R Fielder and Merrick J Moseley

BMJ published online 13 Sep 2007;
doi:10.1136/bmj.39301.460150.55

Table 1| Baseline characteristics of children according to two prescribed occlusion

	Prescribed occlusion (hours/day)	
	6 (n=40)	12 (n=40)
Mean (SD) baseline visual acuity	0.45 (0.30)	0.44 (0.30)
Type of amblyopia:		
Anisometropia	14	20
Strabismus	12	7
Mixed	14	13
Mean (SD) age (years)	5.4 (1.7)	5.6 (1.4)

BMJ

Antibiotic treatment for pyelonephritis in children: multicentre randomised controlled non-inferiority trial

Giovanni Montini, Antonella Toffolo, Pietro Zucchetta, Roberto Dall'Amico, Daniela Gobber, Alessandro Calderan, Francesca Maschio, Luigi Pavanello, Pier Paolo Molinari, Dante Scorrano, Sergio Zanchetta, Walburga Cassar, Paolo Brisotto, Andrea Corsini, Stefano Sartori, Liviana Da Dalt, Luisa Murer and Graziella Zacchello

BMJ 2007;335;386-; originally published online 4 Jul 2007; doi:10.1136/bmj.39244.692442.55

Table 1| Demographic and clinical characteristics of 502 children with acute pyelonephritis (APN) according to allocation to new treatment (oral co-amoxiclav) or standard treatment (intravenous ceftriaxone followed by oral co-amoxiclav)

	New treatment (n=244)	Standard treatment (n=258)
Age (months):		
Median (range)	8.1 (1-81)	7.9 (1-99)
Mean (SD)	12.7 (14.2)	11.9 (13.9)
No of children	244	258
No (%) girls	159/244 (65.2)	163/258 (63.2)
Max body temperature (°C):		
Median (range)	39.25 (36.5-41)	39.2 (36.8-41.5)
Mean (SD)	39.1 (0.77)	39.2 (0.78)
No of children	240	255
White cell count ($\times 10^9$/l):		
Median (range)	17.2 (5.5-45.9)	17.0 (3.8-37.1)
Mean (SD)	18.1 (6.4)	17.8 (6.0)
No of children	243	257

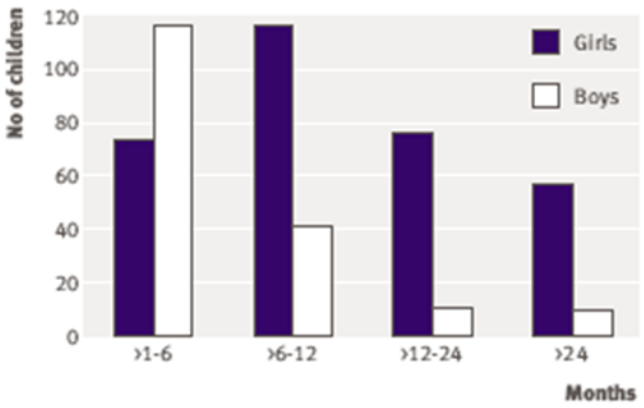


Fig 2| Distribution by age (months) and sex of 502 children

Testing hypotheses, motivating example I.

TABLE 2.1 Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Source: Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New Engl. J. Med.* 318: 262–264 (1988).

- This table is from a report on the relationship between aspirin use and heart attacks by the Physicians' Health Study Research Group at Harvard Medical School.
- The Physicians' Health Study was a 5-year randomized study of whether regular aspirin intake reduces mortality from cardiovascular disease.
- Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was *blind* those in the study did not know whether they were taking aspirin or a placebo.

Testing hypotheses, motivating example II.

TABLE 3.1 Swedish Study on Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Yes	No	Total
Placebo	28	656	684
Aspirin	18	658	676

Source: Based on results described in *Lancet* **338**: 1345–1349 (1991).

- The study randomly assigned 1360 patients who had already suffered a stroke to an aspirin treatment or to a placebo treatment.
- The table reports the number of deaths due to myocardial infarction during a follow-up period of about 3 years.

* Categorical Data Analysis , Alan Agresti (Wiley, 2002)

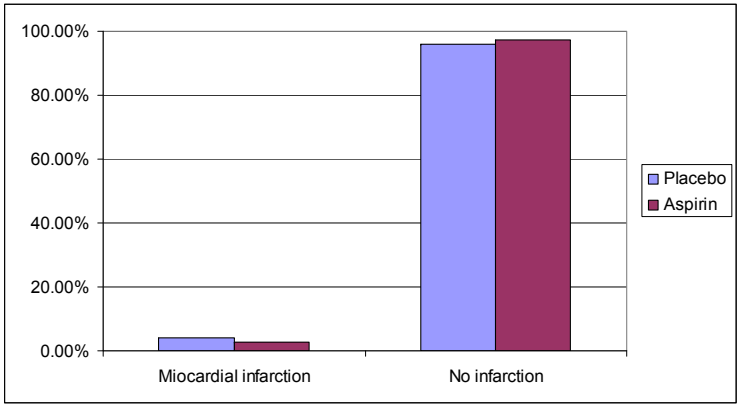
Questions

- Is the difference between the number of infarctions „meaningful”, i.e., statistically significant?
- Are these results caused only by chance or, can we claim that aspirin use decreases the ?
- If Aspirin has no effect, what is the probability that we get this difference?

	Miocardial infarction	No infarction
Placebo	28	656
Aspirin	18	658

	Miocardial infarction	No infarction
Placebo	4.09%	95.91%
Aspirin	2.66%	97.34%

- Answer: Prob=0.14. It is plausible that the true odds of death due to myocardial infarction are equal for aspirin and placebo.
- If there truly is a beneficial effect of aspirin but p-value is not too big, it may require a large sample size to show that benefit because of the relatively small number of myocardial infarction cases



Testing hypotheses, motivating example III.

Effect of Glucosamine on Pain-Related Disability in Patients With Chronic Low Back Pain and Degenerative Lumbar Osteoarthritis

A Randomized Controlled Trial

Philip Wilkens, MChiro
Inger B. Scheel, PhD
Oliver Grundnes, PhD
Christian Hellum, MD
Kjersti Storheim, PhD

Results At baseline, mean RMDQ scores were 9.2 (95% confidence interval [CI], 8.4-10.0) for glucosamine and 9.7 (95% CI, 8.9-10.5) for the placebo group ($P=.37$). At 6 months, the mean RMDQ score was the same for the glucosamine and placebo groups (5.0; 95% CI, 4.2-5.8). At 1 year, the mean RMDQ scores were 4.8 (95% CI, 3.9-5.6) for glucosamine and 5.5 (95% CI, 4.7-6.4) for the placebo group. No statistically significant difference in change between groups was found when assessed after the 6-month intervention period and at 1 year: RMDQ ($P=.72$), LBP at rest ($P=.91$), LBP during activity ($P=.97$), and quality-of-life EQ-5D ($P=.20$). Mild adverse events were reported in 40 patients in the glucosamine group and 46 in the placebo group ($P=.48$).

Conclusions Among patients with chronic LBP and degenerative lumbar OA, 6-month treatment with oral glucosamine compared with placebo did not result in reduced pain-related disability after the 6-month intervention and after 1-year follow-up.

Trial Registration clinicaltrials.gov Identifier: NCT00404079

JAMA. 2010;304(1):45-52 www.jama.com

Primary Outcome Measure

The primary outcome measure used was the Norwegian version of the RMDQ,[25] which is a widely used, back-specific, self-administered measure of pain-related disability. Greater levels of disability give higher numbers on a 24-point scale.

Results

Table 2. Primary and Secondary Outcomes

Assessment (Range) and Time of Evaluation	Mean SD (95% CI)[a] Glucosamine (n = 125)	Placebo (n = 125)	Treatment Effect[b]	*P* Value[c]
RMDQ (0 to 24)				
Baseline	9.2 (8.4 to 10.0)	9.7 (8.9 to 10.5)	NA	NA
6 wk	7.0 (6.1 to 7.8)	7.1 (6.3 to 7.9)	–0.1 (–1.3 to 1.0)	.82
3 mo	5.8 (5.0 to 6.6)	6.5 (5.7 to 7.3)	–0.7 (–1.8 to 0.5)	.24
6 mo	5.0 (4.2 to 5.8)	5.0 (4.2 to 5.8)	0.0 (–1.1 to 1.2)	.72
1 y	4.8 (3.9 to 5.6)	5.5 (4.7 to 6.4)	–0.8 (–2.0 to 0.4)	.50

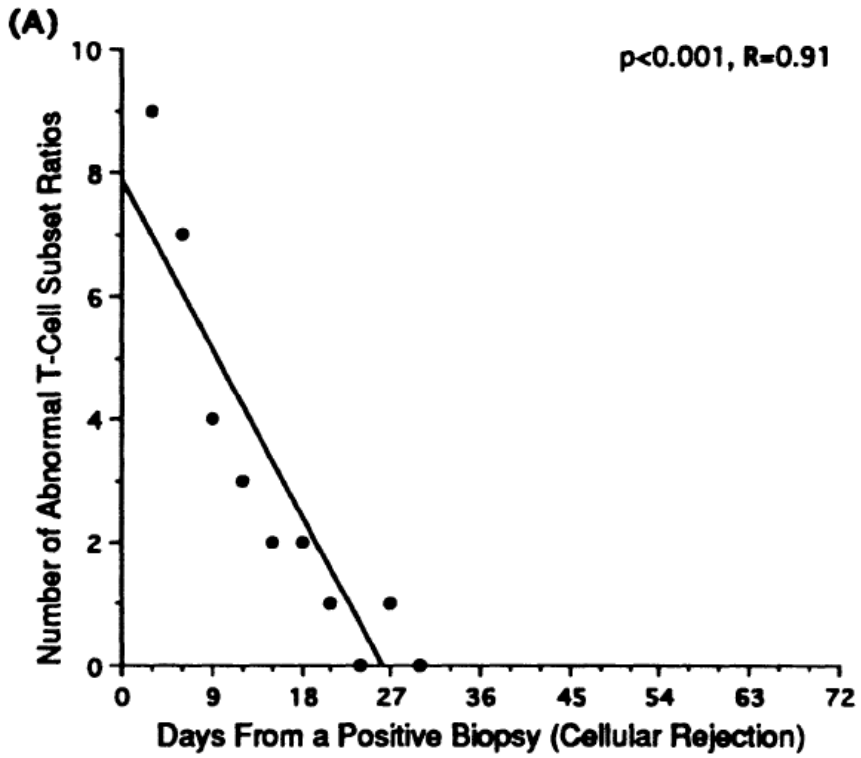
CONCLUSION

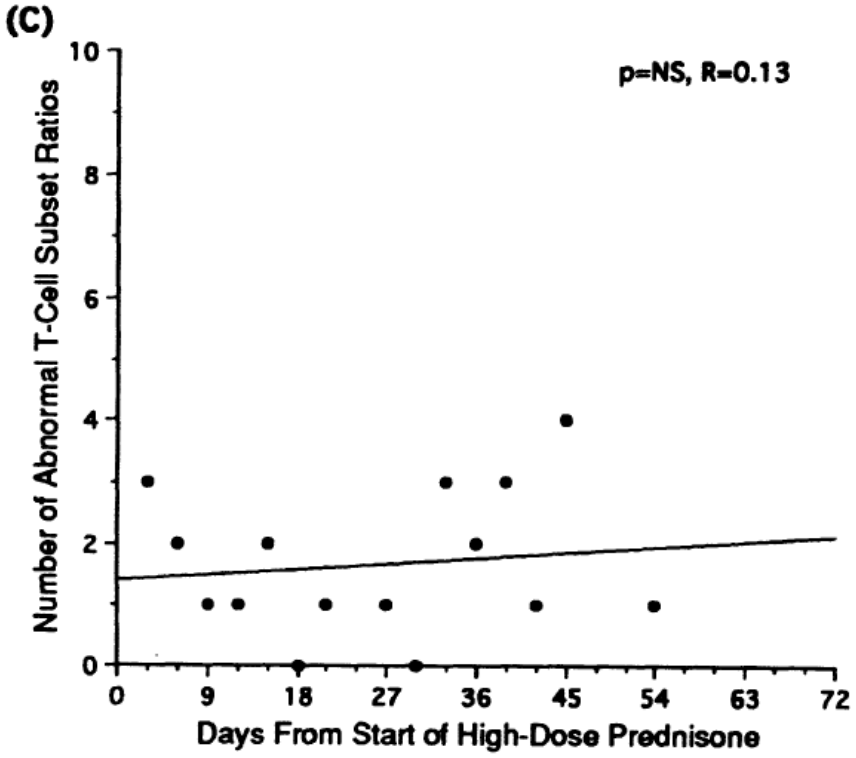
No significant differences were found between glucosamine and placebo during the intervention period or at 1-year follow-up. Both interventions improved functional status by the end of treatment by a similar amount. No serious adverse events were associated with either of the study agents. Based on our results, it seems unwise to recommend glucosamine to all patients with chronic LBP and degenerative lumbar OA. Further research is needed to clarify whether glucosamine is advantageous in an alternative LBP population.

Motivating example IV.

Linear relationship between two measurements – correlation, regression analysis

690 **Circulation** *Vol 90, No 2 August 1994*





Good relationship

week relationship

Descriptive statistics

The data set

- A data set contains information on a number of individuals.
- **Individuals** are objects described by a set of data, they may be people, animals or things. For each individual, the data give values for one or more variables.
- A **variable** describes some characteristic of an individual, such as person's age, height, gender or salary.

The data-table

- Data of one experimental unit ("individual") must be in one record (row)
- Data of the answers to the same question (variables) must be in the same field of the record (column)

Number	SEX	AGE	....
1	1	20	....
2	2	17	....
.	.	.	...

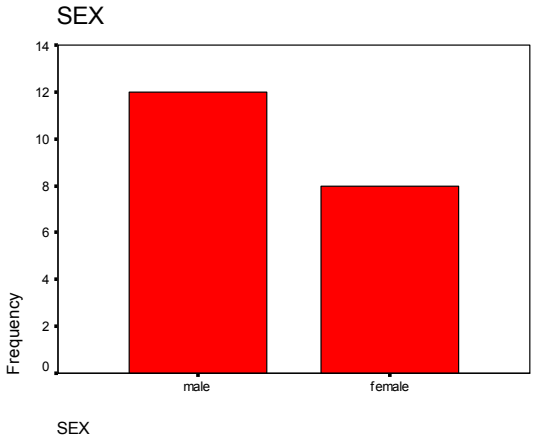
Type of variables

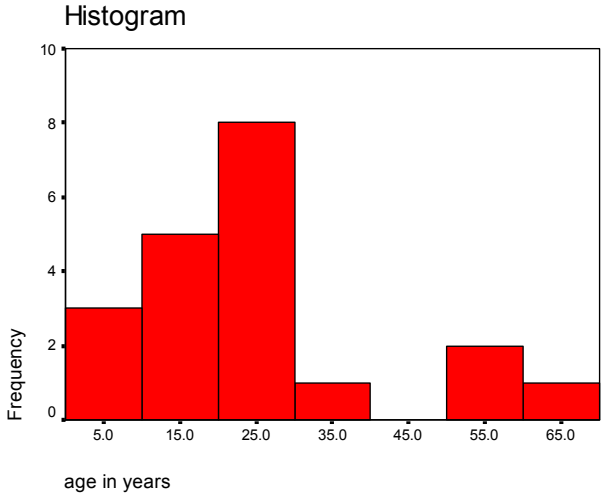
- **Categorical (discrete)**
  A discrete random variable **X** has finite number of possible values
  - Gender
  - Blood group
  - Number of children
  - …
- **Continuous**
  A continuous random variable **X** has takes all values in an interval of numbers.
  - Concentration
  - Temperature
  - …

Distribution of variables

Discrete: the distribution of a categorical variable describes what values it takes and how often it takes these values.

Continuous: the distribution of a continuous variable describes what values it takes and how often these values fall into an interval.

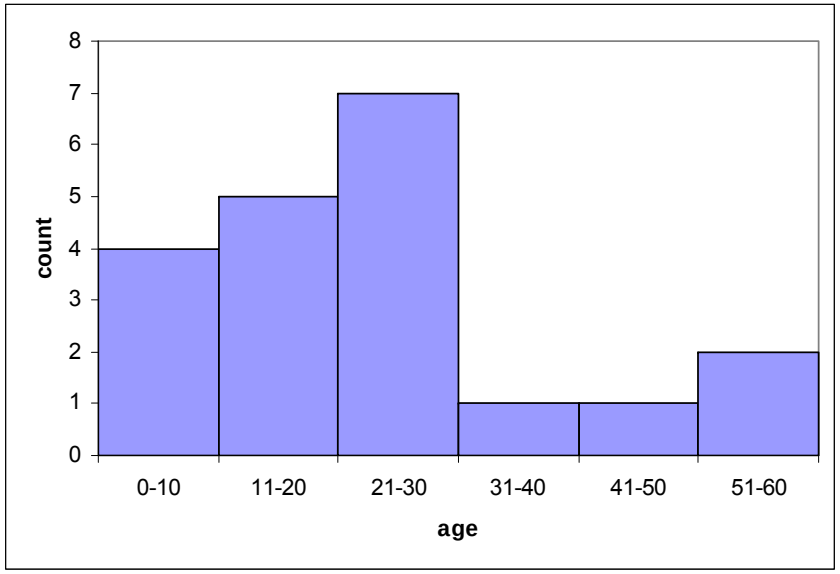


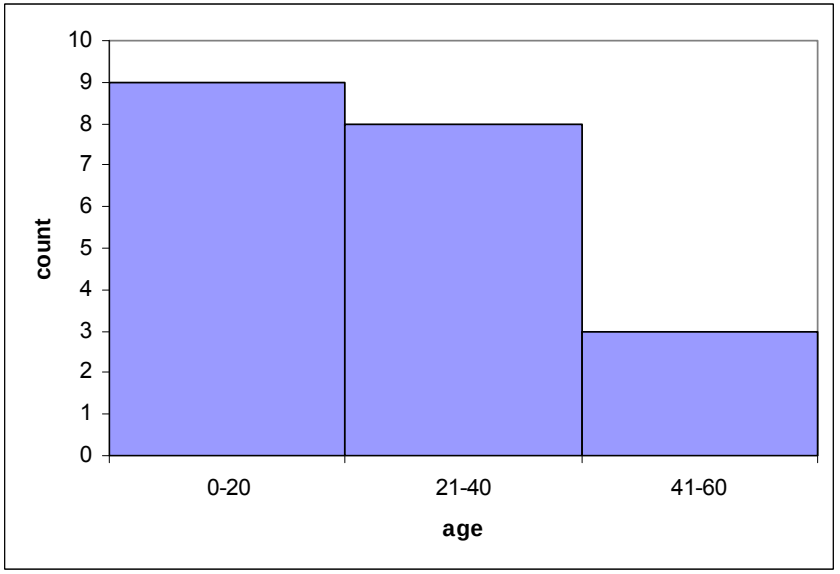


The distribution of a continuous variable, example

Values:	Categories:	Frequencies
20.00	0-10	4
17.00	11-20	5
22.00	21-30	7
28.00	31-40	1
9.00	41-50	1
5.00	51-60	2
26.00		
60.00		
35.00		
51.00		
17.00		
50.00		
9.00		
10.00		
19.00		
22.00		
25.00		
29.00		
27.00		
19.00		

The length of the intervals (or the number of intervals) affect a histogram

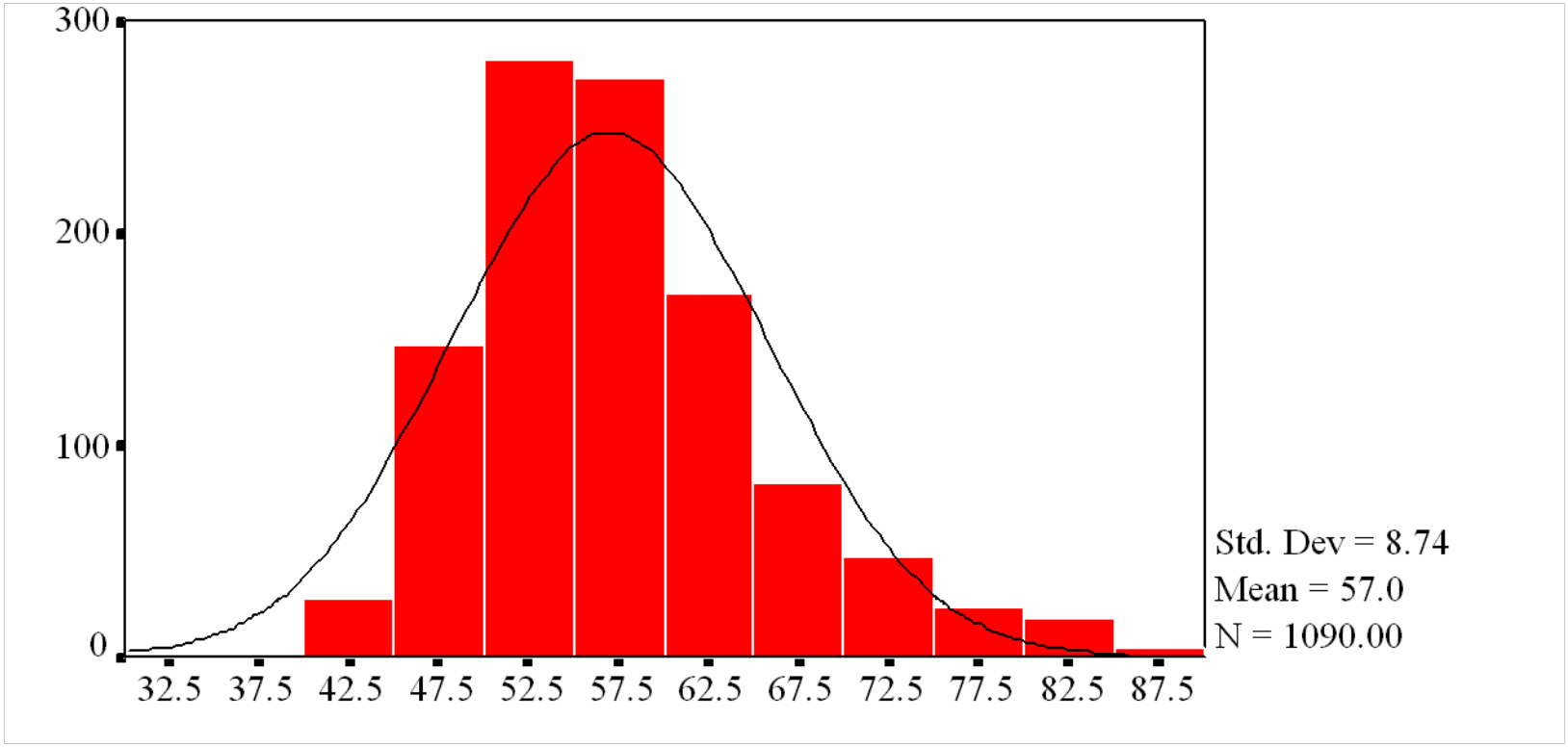




The overall pattern of a distribution

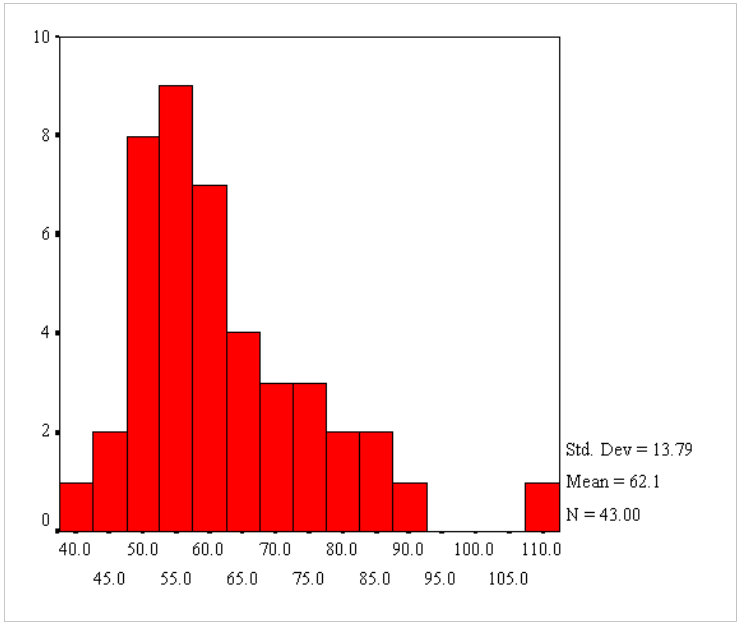
- The **center**, **spread** and **shape** describe the overall pattern of a distribution.
- Some distributions have simple shape, such as **symmetric** and **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.
- A distribution is skewed to the right if the right side of the histogram extends much farther out then the left side.

Histogram of body mass (kg)



Outliers

- Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them (real data, typing mistake or other).



Describing distributions with numbers

- **Measures of central tendency**: the mean, the mode and the median are three commonly used measures of the center.
- **Measures of variability** : the range, the quartiles, the variance, the standard deviation are the most commonly used measures of variability .
- **Measures of an individual**: rank, z score

Measures of the center

- **Mean:** $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **Mode: is the most frequent number**
- **Median: is the value that half the members of the sample fall below and half above. In other words, it is the middle number when the sample elements are written in numerical order**

- **Example: 1,2,4,1**
- **Mean**
- **Mode**
- **Median**

Measures of the center

- **Mean:**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Mode: is the most frequent number**
- **Median: is the value that half the members of the sample fall below and half above. In other words, it is the middle number when the sample elements are written in numerical order**

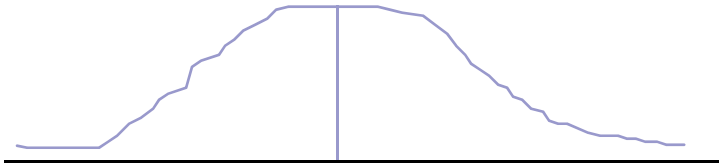
- **Example: 1,2,4,1**
- **Mean=8/4=2**
- **Mode=1**
- **Median**
  - **First sort data 1 1 2 4**
  - **Then find the element(s) in the middle**
    - **If the sample size is odd, the unique middle element is the median**
    - **If the sample size is even, the median is the average of the two central elements**
    - **1 1 2 4**
    - **Median=1.5**

Example

- The grades of a test written by 11 students were the following:
- 100 100 100 63 62 60 12 12 6 2 0.
- A student indicated that the class average was 47, which he felt was rather low. The professor stated that nevertheless there were more 100s than any other grade. The department head said that the middle grade was 60, which was not unusual.
  The mean is 517/11=47, the mode is 100, the median is 60.

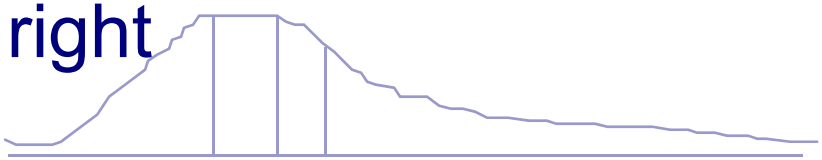
Relationships among the mean(m), the median(M) and the mode(Mo)

- A symmetric curve



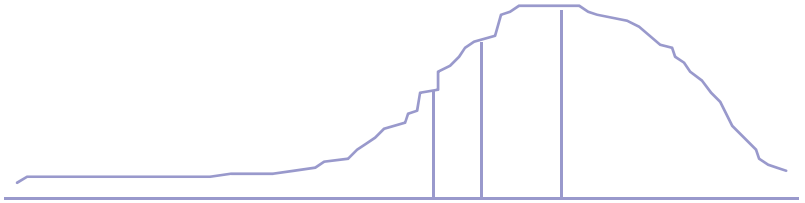
m=M=Mo

- A curve skewed to the right



Mo<M< m

- A curve skewed to the left



M < M < Mo

Measures of variability (dispersion)

- The **range** is the difference between the largest number (maximum) and the smallest number (minimum).
- **Percentiles** (5%-95%): 5% percentile is the value below which 5% of the cases fall.
- **Quartiles**: 25%, 50%, 75% percentiles

- **The variance=**

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

- **The standard deviation:**

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\text{var}\,iance}$$

Example

- **Data: 1 2 4 1, in ascending order: 1 1 2 4**
- **Range: max-min=4-1=3**
- **Quartiles:**
- **Standard deviation:**

Percentiles

	Percentiles		
	25	50	75
Weighted Average(Definition 1)	1.0000	1.5000	3.5000
Tukey's Hinges	1.0000	1.5000	3.0000

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1-2=-1	1
1	1-2=-1	1
2	2-2=0	0
4	4-2=2	4
Total	0	6

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{6}{3}} = \sqrt{2} = 1.414$$

The meaning of the standard deviation

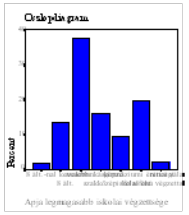
- A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations.
- For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

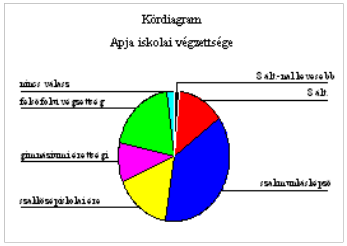
The use of sample characteristics in summary tables

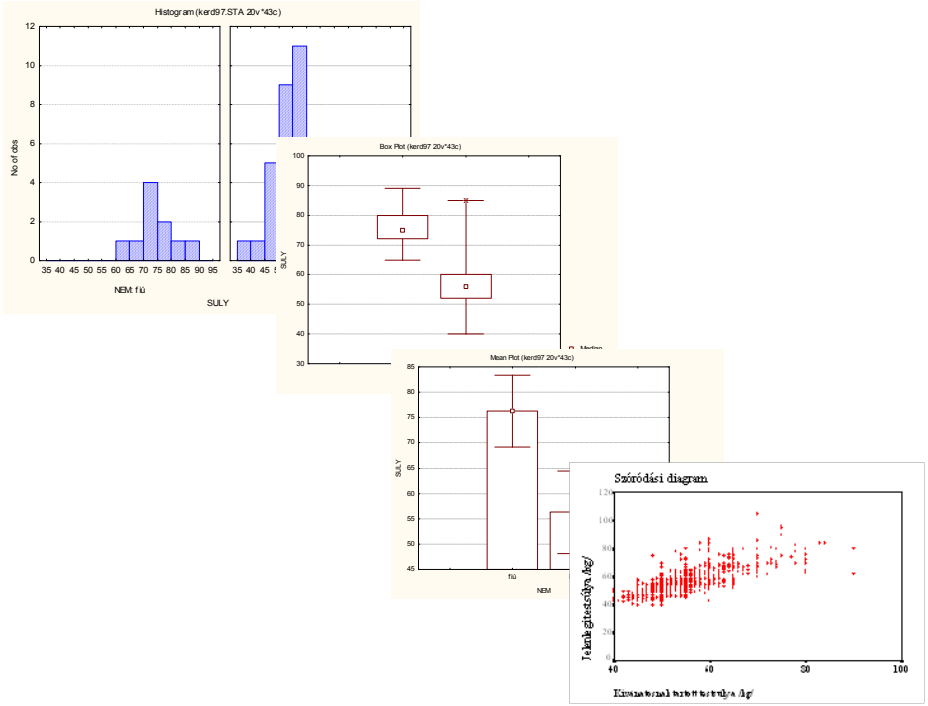
Center	Dispersion	Publish
Mean	Standard deviation, Standard error	Mean (SD) Mean ± SD Mean ± SE Mean ± SEM
Median	Min, max 5%, 95%s percentile 25 % , 75% (quartiles)	Med (min, max) Med(25%, 75%)

Displaying data

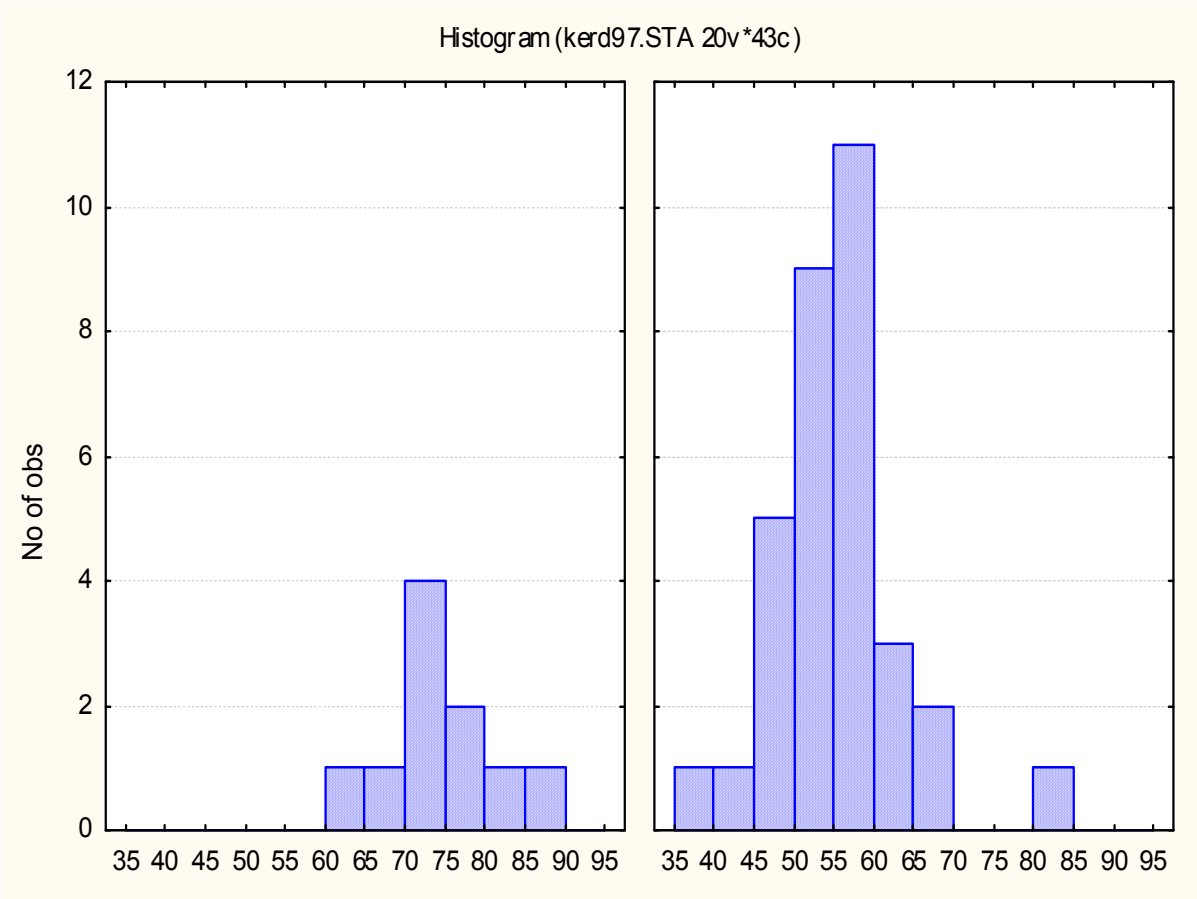
- Categorical data
  - bar chart
  - pie chart
- Continuous data
  - histogram
  - box-whisker plot
  - mean-standard deviation plot
  - scatter plot





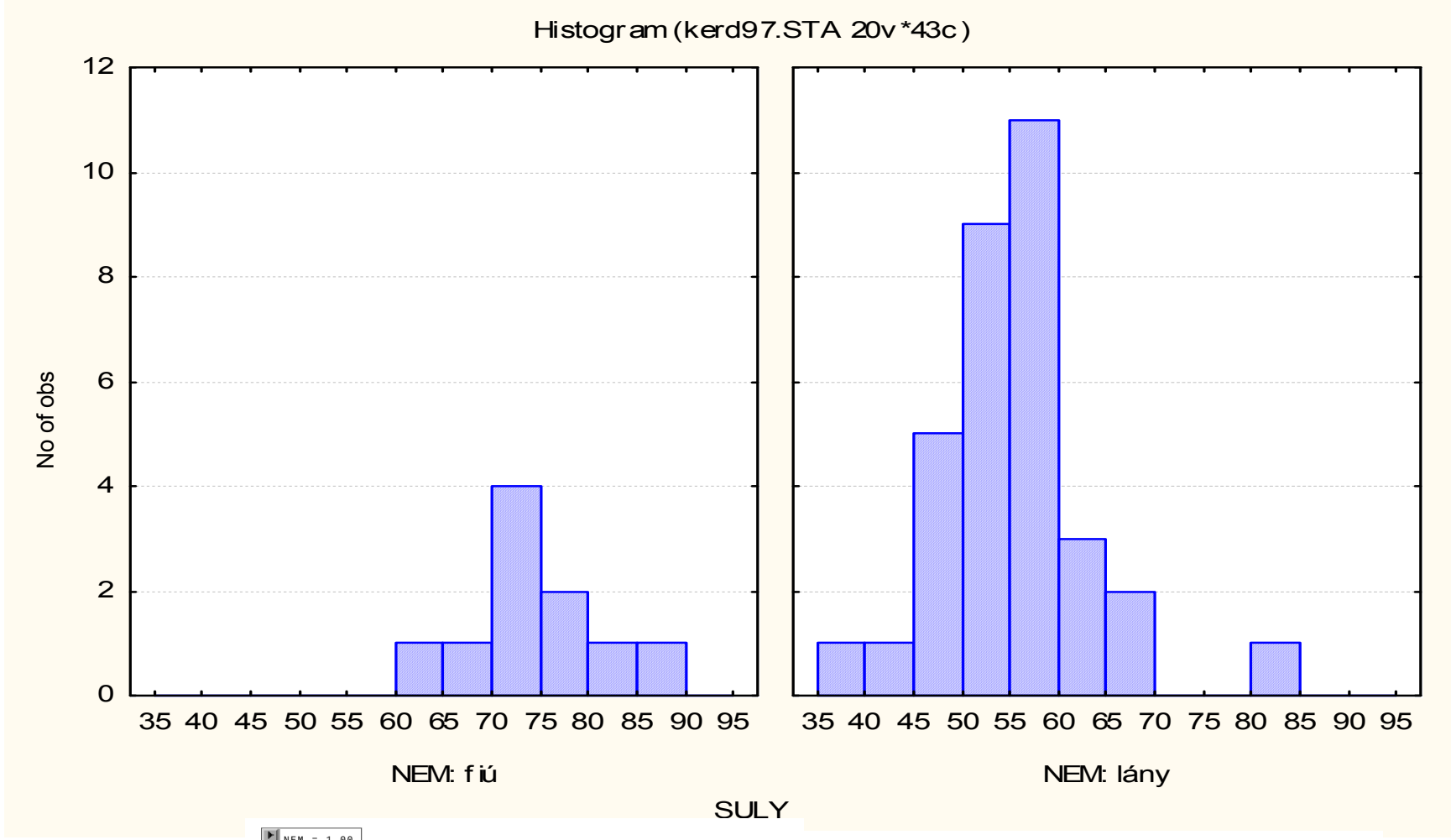


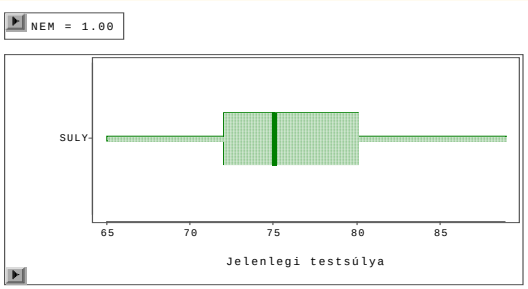
Distribution of body weights
The distribution is skewed in case of girls

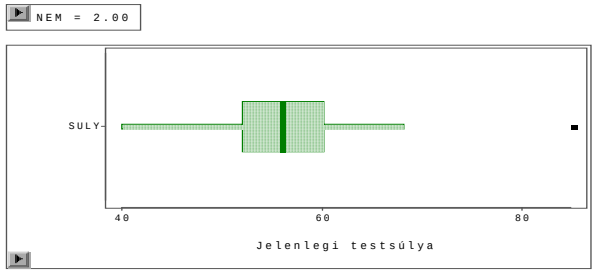


boys

girls

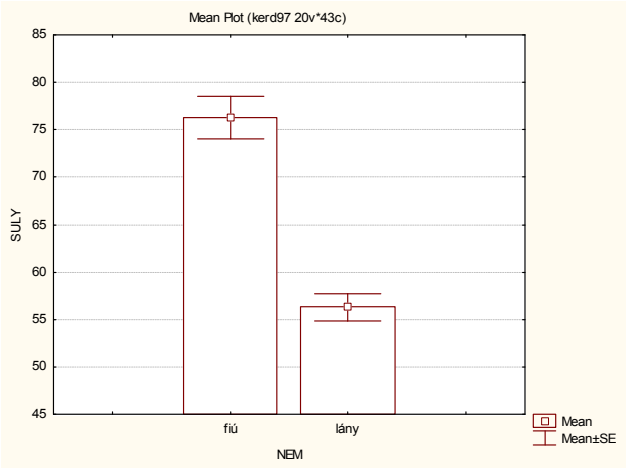

Histogram (kerd97.STA 20v*43c)
No of obs
0
2
4
6
8
10
12
35 40 45 50 55 60 65 70 75 80 85 90 95
NEM: fiú
NEM: lány
SULY


NEM = 1.00
SULY
65 70 75 80 85
Jelenlegi testsúlya

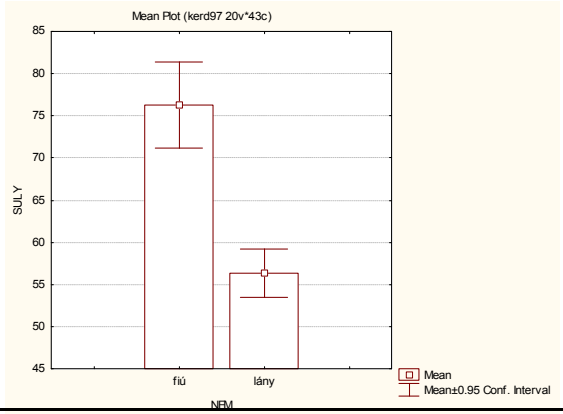

NEM = 2.00
SULY
40 60 80
Jelenlegi testsúlya

Mean-dispersion diagrams

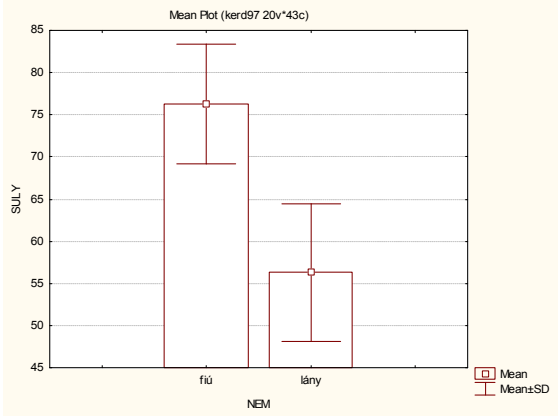
- **Mean + SD**
- **Mean + SE**
- **Mean + 95% CI**



Mean ± SE

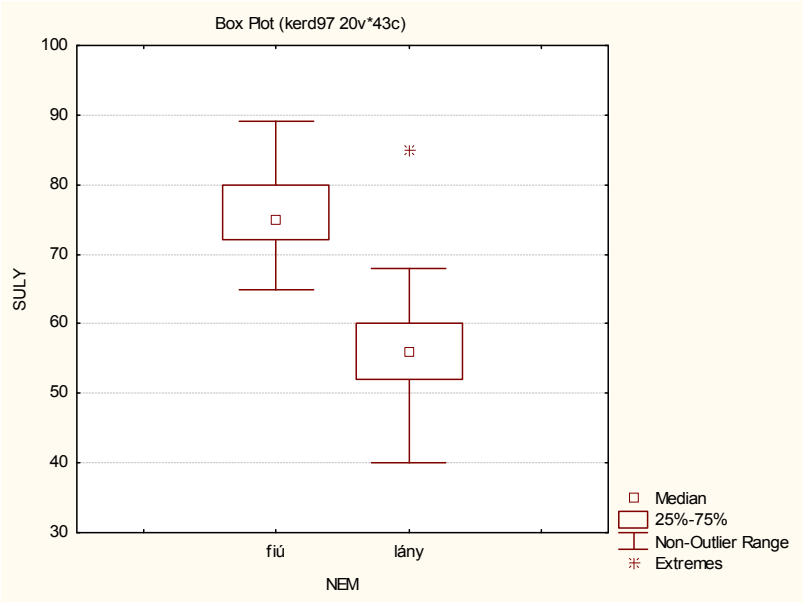


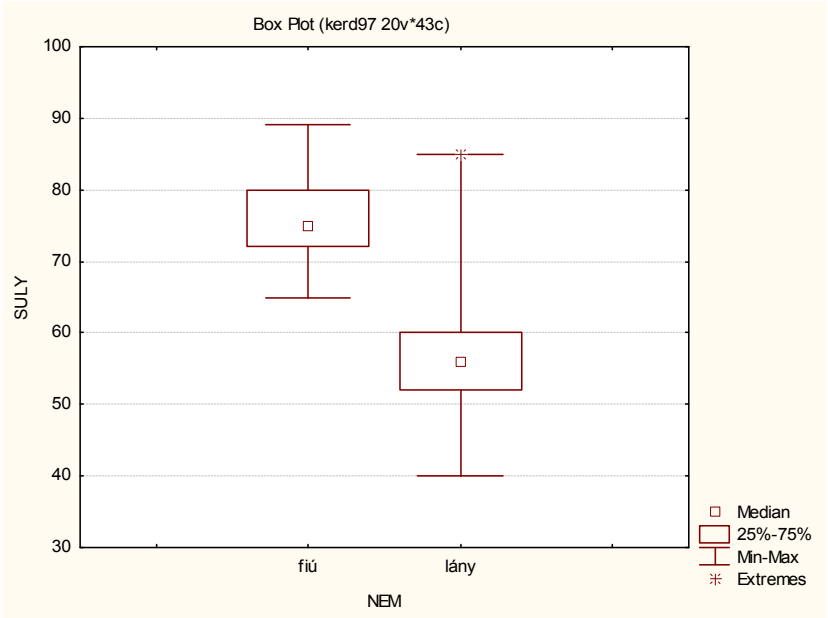
Mean ± 95% CI



Mean ± SD

Box diagram



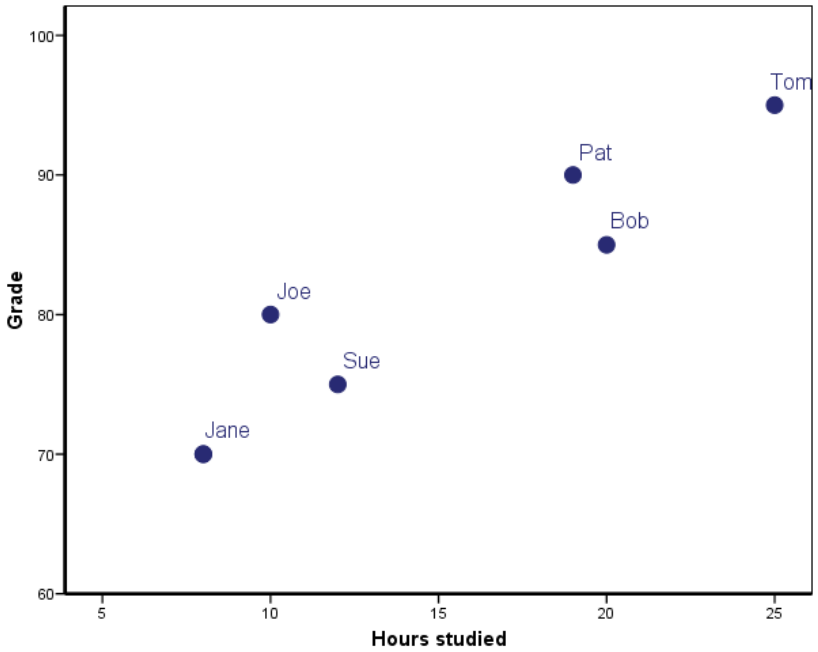


A box plot, sometimes called a box-and-whisker plot displays the median, quartiles, and minimum and maximum observations .

Scatterplot

Relationship between two continouous variables

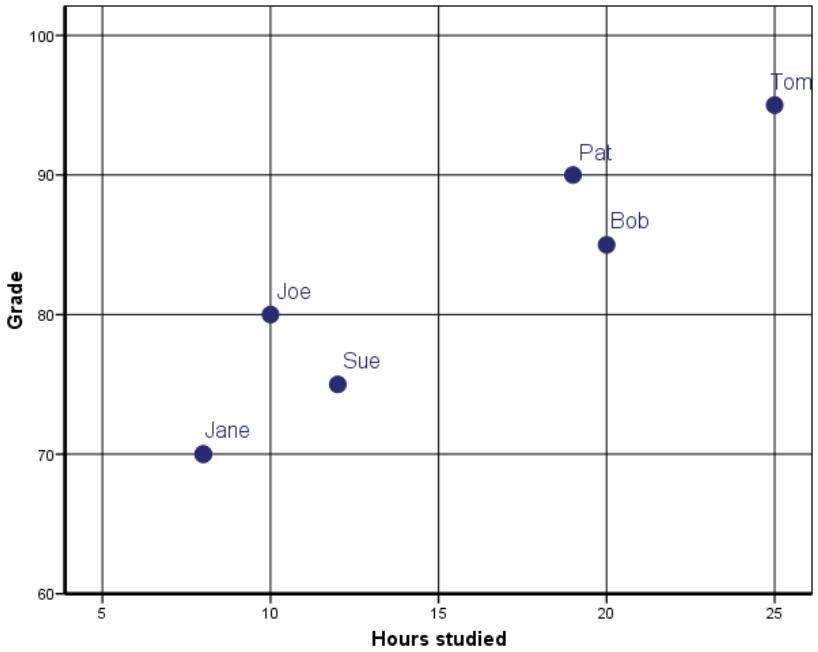
Student	Hours studied	Grade
Jane	8	70
Joe	10	80
Sue	12	75
Pat	19	90
Bob	20	85
Tom	25	95



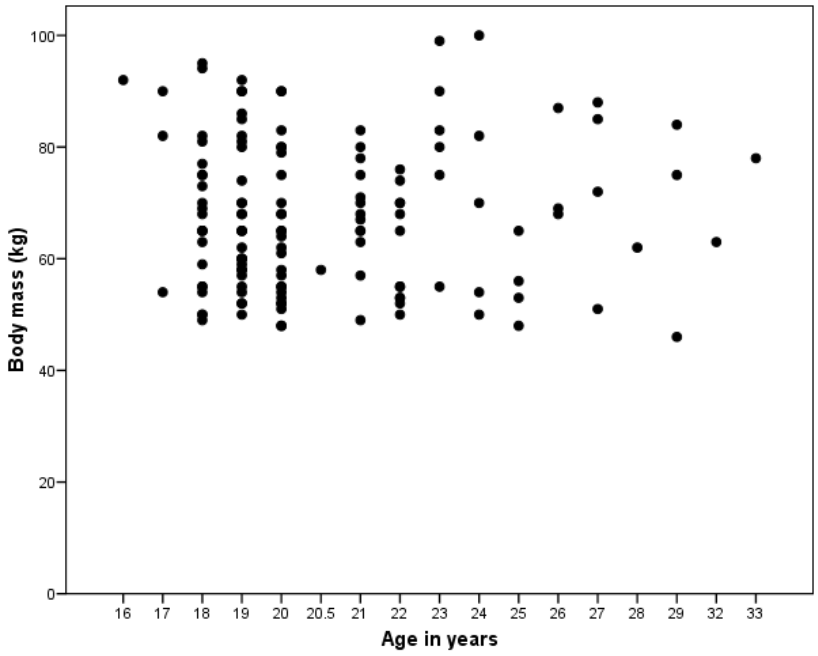
Scatterplot

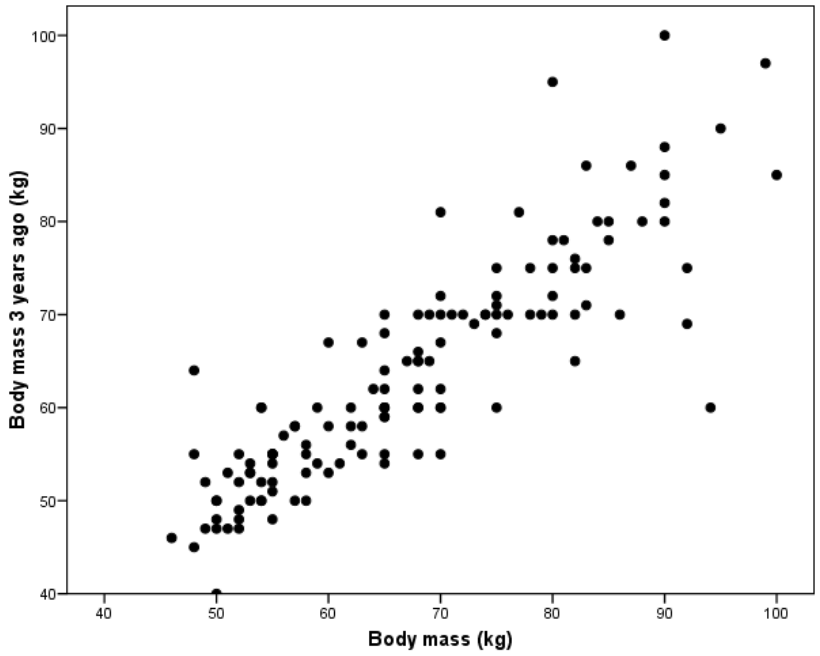
Relationship between two continouous variables

Student	Hours studied	Grade
Jane	8	70
Joe	10	80
Sue	12	75
Pat	19	90
Bob	20	85
Tom	25	95



Scatterplot
Other examples





Transformations of data values Addition, subtraction

- Adding (or subtracting) the same number to each data value in a variable shifts each measures of center by the amount added (subtracted).
- Adding (or subtracting) the same number to each data value in a variable does not change measures of dispersion.

Transformations of data values Multiplication, division

- Measures of center and spread change in predictable ways when we multiply or divide each data value by the same number.
- Multiplying (or dividing) each data value by the same number multiplies (or divides) all measures of center or spread by that value.

Proof.
The effect of linear transformations

- Let the transformation be x ->ax+b
- Mean: $\frac{\sum_{i=1}^{n} ax_i+b}{n} = \frac{ax_1+b+ax_2+b+\ldots+ax_n+b}{n} = \frac{a(x_1+x_2+\ldots+x_n)+nb}{n} = a\bar{x}+b$
- Standard deviation:

$$\sqrt{\frac{\sum_{i=1}^{n}((ax_i+b)-(a\bar{x}+b))^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}((ax_i+b-a\bar{x}-b))^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(ax_i-a\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}a^2(x_i-\bar{x})^2}{n-1}} = |a|\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = |a|SD$$

Example: the effect of transformations

Sample data (x_i)	Addition (x_i +10)	Subtraction (x_i -10)	Multiplication (x_i *10)	Division (x_i /10)
1	11	-9	10	0.1
2	12	-8	20	0.2
4	14	-6	40	0.4
1	11	-9	10	0.1
Mean=2	12	-8	20	0.2
Median=1.5	11.5	-8.5	15	0.15
Range=3	3	3	30	0.3
St.dev.≈1.414	≈1 .414	≈ 1.414	≈ 14.14	≈ 0.1414

Special transformation: standardisation

- The z score measures how many standard deviations a sample element is from the mean. A formula for finding the z score corresponding to a particular sample element x_i is
- $$z_i = \frac{x_i - \bar{x}}{s}$$, i=1,2,...,n.
- We standardize by subtracting the mean and dividing by the standard deviation.
- The resulting variables (z-scores) will have
  - Zero mean
  - Unit standard deviation
  - No unit

Example: standardisation

	Sample data (x_i)	Standardised data (z_i)
	1	-1
	2	0
	4	2
	1	1
Mean	2	0
St. deviation	≈1 .414	1

Population, sample

- **Population**: the entire group of individuals that we want information about.
- **Sample**: a part of the population that we actually examine in order to get information
- A simple random sample of size *n* consists of *n* individuals chosen from the population in such a way that every set of n individuals has an equal chance to be in the sample actually selected.

Examples

- Sample data set
  - Questionnaire filled in by a group of pharmacy students
  - Blood pressure of 20 healthy women
  - …
- Population
  - Pharmacy students
  - Students
  - Blood pressure of women (whoever)
  - …

Sample

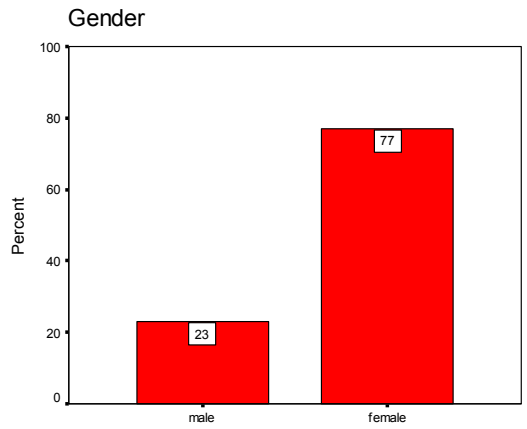
Population

(approximates)

- **Bar chart of relative frequencies of a categorical variable**
- **Distribution of that variable in the population**

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	20	23.0	23.0	23.0
	female	67	77.0	77.0	100.0
	Total	87	100.0	100.0	



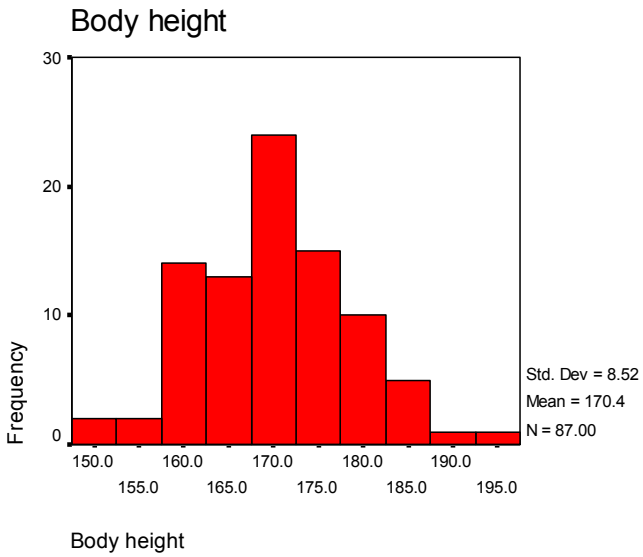
Sample

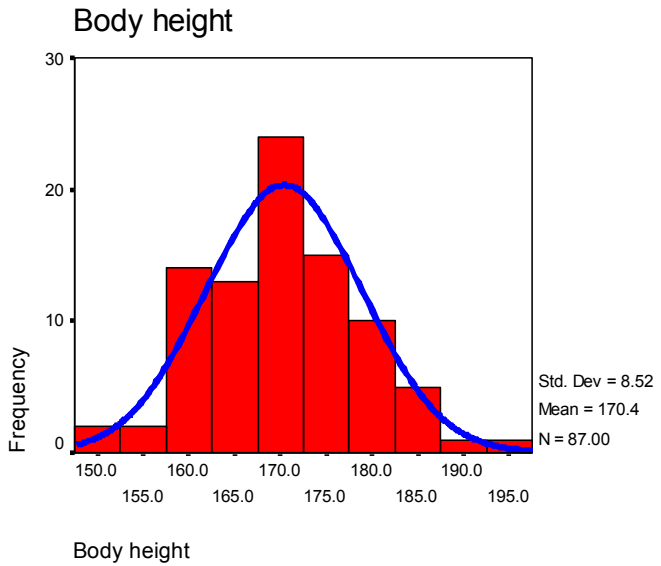
Population

(approximates)

- **Histogram of relative frequencies of a continuous variable**

- **Distribution of that variable in the population**





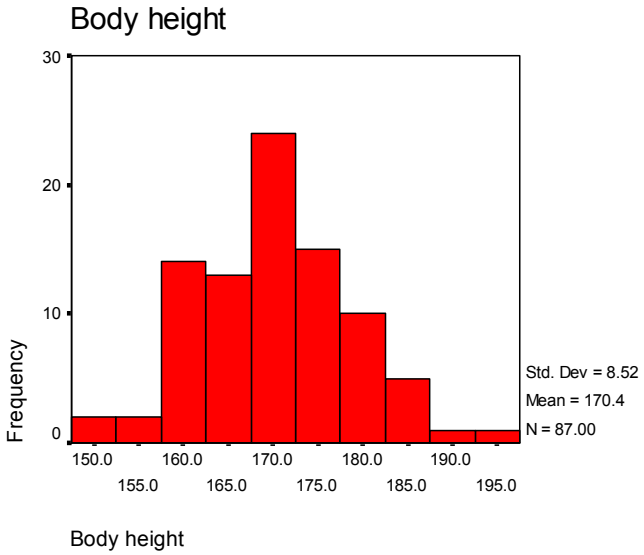
Sample

- Mean ($\bar{x}$)
- Standard deviation (SD)
- Median

(approximates)

Population

- Mean μ (unknown)
- Standard deviation σ (unknown)
- Median (unknown)



Useful WEB pages

- http://onlinestatbook.com/rvls.html
- http://www-stat.stanford.edu/~naras/jsm
- http://my.execpc.com/~helberg/statistics.html
- http://www.math.csusb.edu/faculty/stanton/m262/index.html