BIOSTATISTICS COURSE

DESCRIPTIVE STATISTICS

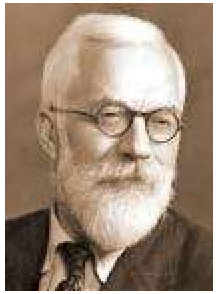
PAOLA REBORA paola.rebora@unimib.it

SCHOOL OF MEDICINE AND SURGERY

Introduction

"[…] Statistics may be regarded as the study of populations, the study of methods of the reduction of data, the study of variation."

Sir Ronald Aylmer Fisher (1950)



The key aspect is the possibility to isolate **common behaviors and associations** between phenomena while controlling for or taking into account the different sources of **variability**

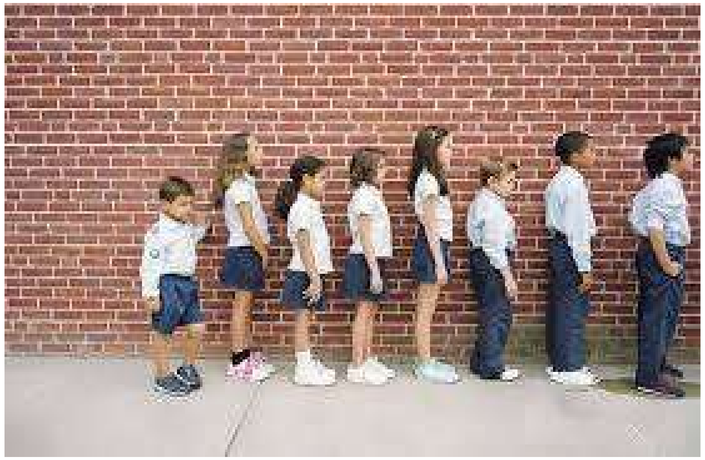
Statistical methodology is involved in all phases of scientific research:

- **Design** of the study/experiment
- Data collection and **elaboration**
- **Interpretation** of the results

Variability sources

Instrumental

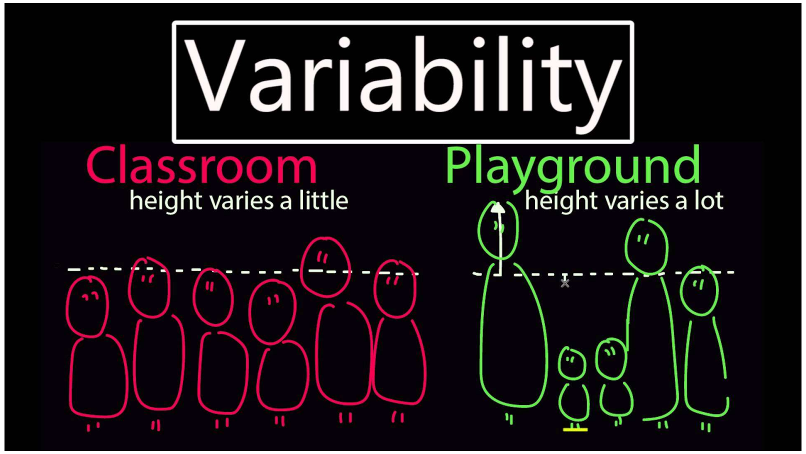
biological



Variability sources

Instrumental

biological



Instrumental variability can be completely controlled by acting on the measurement methods (optimizing the operating procedure, calibrating the instrument, training the staff).

Biological variability can only be partially limited, making the set of subjects analyzed more homogeneous.

Statistics

Statistical methodology is involved in all phases of scientific research:

1. **Design** of the study/experiment
2. Data collection and **elaboration**
3. **Interpretation** of the results

SCHOOL OF MEDICINE AND SURGERY

Planning/ Design of the study

- Primary question of interest
  - Quantifying information about a single group
  - Comparing multiple groups
- sample size
  - How many subject needed in total?
  - How many in each group to be compared?
- Selecting study participant
  - Randomly chosen
  - Selected from a pool of (interested) persons?
  - Take whoever shows up

  if group comparison: how to assign to groups

SCHOOL OF MEDICINE AND SURGERY

Data collection /data analysis

- How to best summarise information from raw data
- Dealing with variability (natural and instrumental/sampling related)
  - Important patterns in data obscured by variability
  - Distinguish real patterns from random variation
- Inference: using information from the single study coupled with info about variability to make statement about the larger population/process of interest

SCHOOL OF MEDICINE AND SURGERY

Presentation and Interpretation

- What summary measures will best convey the «main messagges» in the data on the question of interest
- How to convey uncertainty in estimates based on data
- What do the results mean?

Terminology: population and sample

Population (also called universe space):

Any collection of individuals in which we might be interested, where these individuals might be anything, and the number of individuals may be finite or infinite.

Sample:

a subset of the population, used to draw conclusions on the target population.

Terminology: population and sample

Population (also called universe space):
Any collection of individuals on which we might be interested, where these individuals might be anything, and the number of individuals may be finite or infinite.

Example:
-all people in Italy
-all people with diabetes
-all possible measures of blood pressure on a patient

Sample:
a subset of the population, used to draw conclusions on the target population.

Terminology: population and sample

Target Population:

the population to whom we wish to generalise our findings

Study population:

the population from which we sampled, also called the "study base"

Sample:

a subset of the population, used to draw conclusions on the target population.

SCHOOL OF MEDICINE AND SURGERY

Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial

Yeming Wang*, Dingyu Zhang*, Guanhua Du*, Ronghui Du*, Jianping Zhao*, Yang Jin*, Shouzhi Fu*, Ling Gao*, Zhenshun Cheng*, Qiaofa Lu*, Yi Hu*, Guangwei Luo*, Ke Wang, Yang Lu, Huadong Li, Shuzhen Wang, Shunan Ruan, Chengqing Yang, Chunlin Mei, Yi Wang, Dan Ding, Feng Wu, Xin Tang, Xianzhi Ye, Yingchun Ye, Bing Liu, Jie Yang, Wen Yin, Aili Wang, Guohui Fan, Fei Zhou, Zhibo Liu, Xiaoying Gu, Jiuyang Xu, Lianhan Shang, Yi Zhang, Lianjun Cao, Tingting Guo, Yan Wan, Hong Qin, Yushen Jiang, Thomas Jaki, Frederick G Hayden, Peter W Horby, Bin Cao, Chen Wang

Summary

Background No specific antiviral drug has been proven effective for treatment of patients with severe coronavirus disease 2019 (COVID-19). Remdesivir (GS-5734), a nucleoside analogue prodrug, has inhibitory effects on pathogenic animal and human coronaviruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in vitro, and inhibits Middle East respiratory syndrome coronavirus, SARS-CoV-1, and SARS-CoV-2 replication in animal models.

Methods We did a randomised, double-blind, placebo-controlled, multicentre trial at ten hospitals in Hubei, China. Eligible patients were adults (aged ≥18 years) admitted to hospital with laboratory-confirmed SARS-CoV-2 infection, with an interval from symptom onset to enrolment of 12 days or less, oxygen saturation of 94% or less on room air or a ratio of arterial oxygen partial pressure to fractional inspired oxygen of 300 mm Hg or less, and radiologically confirmed pneumonia. Patients were randomly assigned in a 2:1 ratio to intravenous remdesivir (200 mg on day 1 followed by 100 mg on days 2–10 in single daily infusions) or the same volume of placebo infusions for 10 days. Patients were permitted concomitant use of lopinavir–ritonavir, interferons, and corticosteroids. The primary endpoint was time to clinical improvement up to day 28, defined as the time (in days) from randomisation to the point of a decline of two levels on a six-point ordinal scale of clinical status (from 1=discharged to 6=death) or discharged alive from hospital, whichever came first. Primary analysis was done in the intention-to-treat (ITT) population and safety analysis was done in all patients who started their assigned treatment. This trial is registered with ClinicalTrials.gov, NCT04257656.

Findings Between Feb 6, 2020, and March 12, 2020, 237 patients were enrolled and randomly assigned to a treatment group (158 to remdesivir and 79 to placebo); one patient in the placebo group who withdrew after randomisation was not included in the ITT population. Remdesivir use was not associated with a difference in time to clinical improvement (hazard ratio 1·23 [95% CI 0·87–1·75]). Although not statistically significant, patients receiving remdesivir had a numerically faster time to clinical improvement than those receiving placebo among patients with symptom duration of 10 days or less (hazard ratio 1·52 [0·95–2·43]). Adverse events were reported in 102 (66%) of 155 remdesivir recipients versus 50 (64%) of 78 placebo recipients. Remdesivir was stopped early because of adverse events in 18 (12%) patients versus four (5%) patients who stopped placebo early.

Interpretation In this study of adult patients admitted to hospital for severe COVID-19, remdesivir was not associated with statistically significant clinical benefits. However, the numerical reduction in time to clinical improvement in those treated earlier requires confirmation in larger studies.

Lancet 2020; 395: 1569–78

Published Online April 29, 2020 https://doi.org/10.1016/S0140-6736(20)31022-9

This online publication has been corrected. The corrected version first appeared at thelancet.com on May 28, 2020

See Comment page 1525

*Contributed equally

Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, National Clinical Research Center for Respiratory Diseases (Ye Wang MD, F Zhou MD, Z Liu MD, L Shang MD, Y Zhang MD, Prof B Cao MD, Prof C Wang MD) and Institute of Clinical Medical Sciences (G Fan MS, X Gu PhD), China-Japan Friendship Hospital, Beijing, China; Department of Respiratory Medicine, Capital Medical University, Beijing, China (Ye Wang, Prof B Cao); Jin Yin-tan Hospital, Wuhan, Hubei, China (D Zhang MD, H Li MD, S Wang MS, S Ruan MS); Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China (Prof G Du PhD, Prof K Wang PhD, Prof Y Lu PhD); Wuhan Lung Hospital, Wuhan, China (Prof R Du MD, C Yang MD,

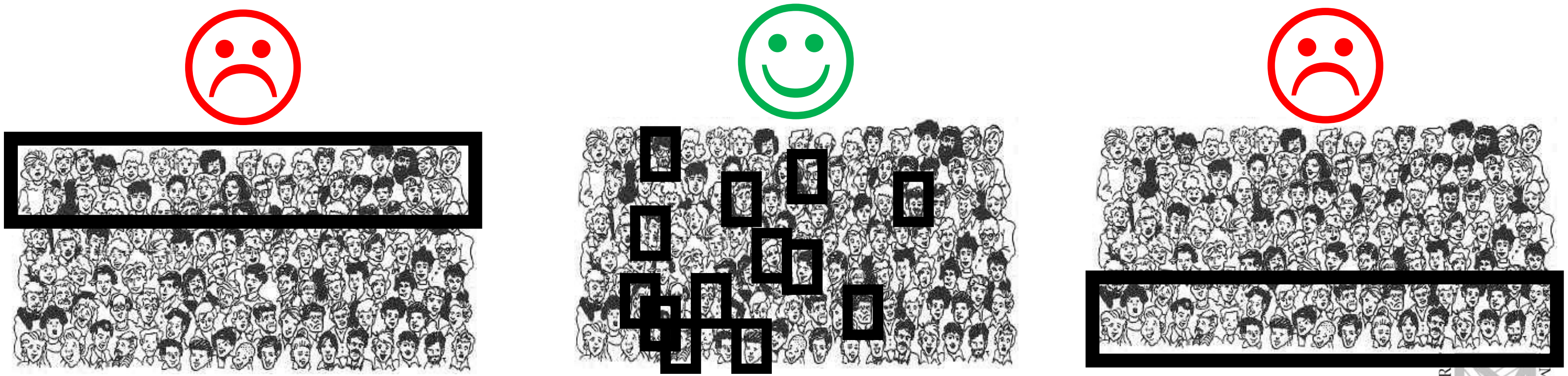
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Sample

A subset of the study population, used to draw conclusions on the target population.

It has to be a probabilistic sample, possibly a random sample of the population

A subset of the population.



last row | **random sample** | **1st row**

Sample

SCHOOL OF MEDICINE AND SURGERY

A **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

Represent the population

↓

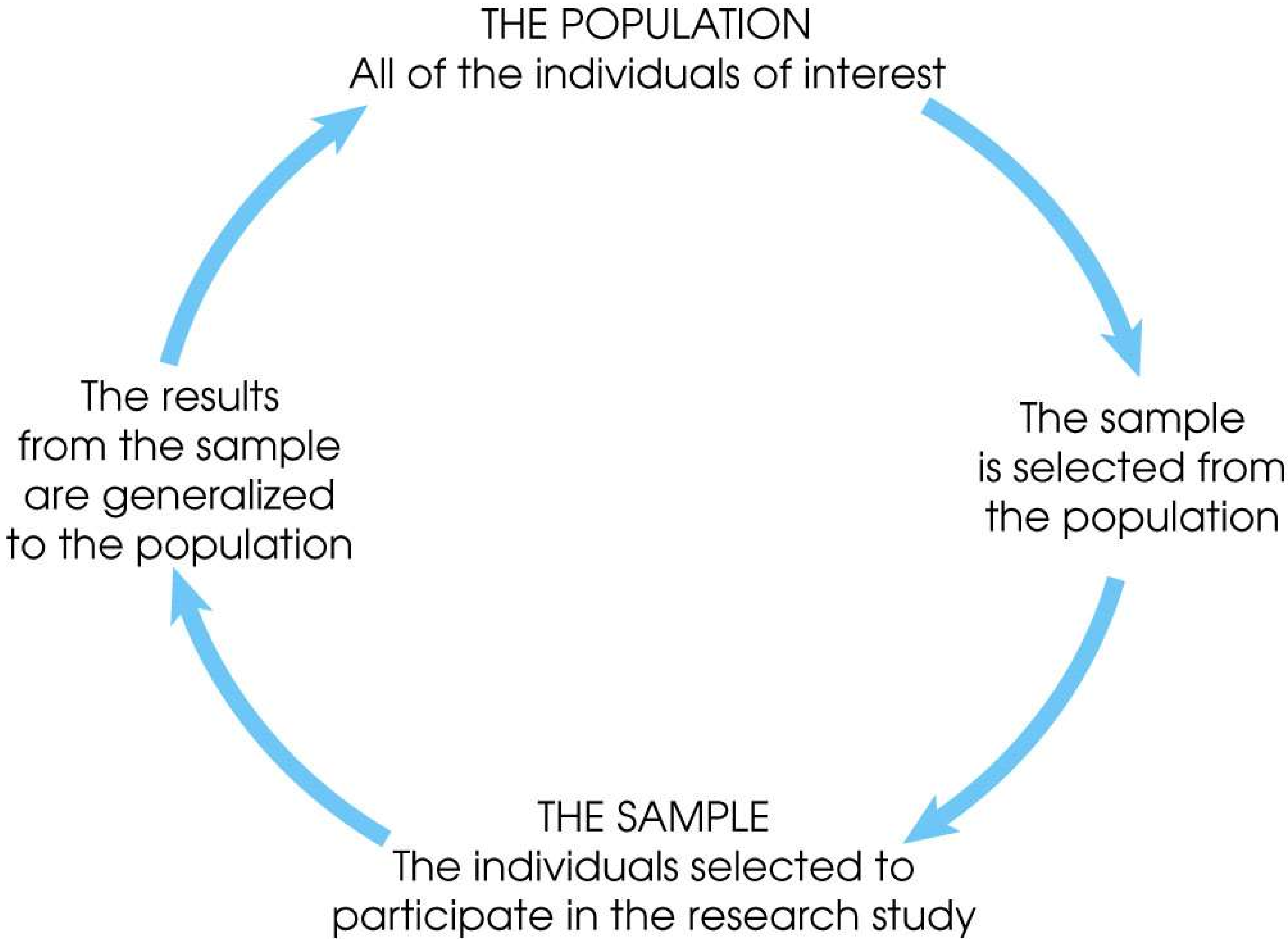
Not verifiable condition to generalise results of the study to target population!

Example. A small number of people accurately reflect the members of an entire population. In a classroom of 30 students, in which half the students are male and half are female, a representative sample might include six students: three males and three females.

DIPARTIMENTO DI MEDICINA E CHIRURGIA

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

The **relationship** between a population and a sample:



Terminology: Random Variable & data

SCHOOL OF MEDICINE AND SURGERY

Random Variable: Characteristic observable in a target population. It may vary from subject to subject.
Data: Numbers or modalities expressed by a random variable (r.v.)

Example:
RANDOM VARIABLE (denoted with UPPER CASE):

X : SEX

Data (denoted with lower case):

x_1=Female x_2=Male x_3=Female



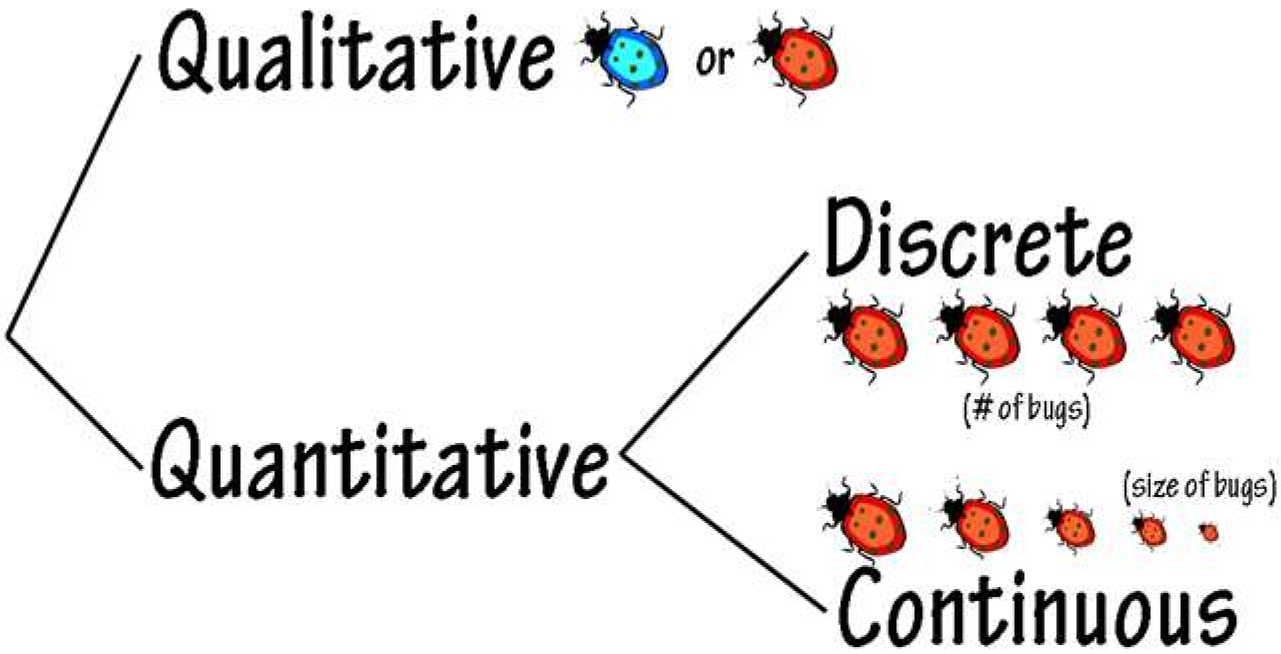
RANDOM VARIABLE: **Y : AGE**

Data: y_1=1 y_2=4 y_3=12

DIPARTIMENTO DI MEDICINA E CHIRURGIA

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

<u>Terminology:</u> Random Variables classification



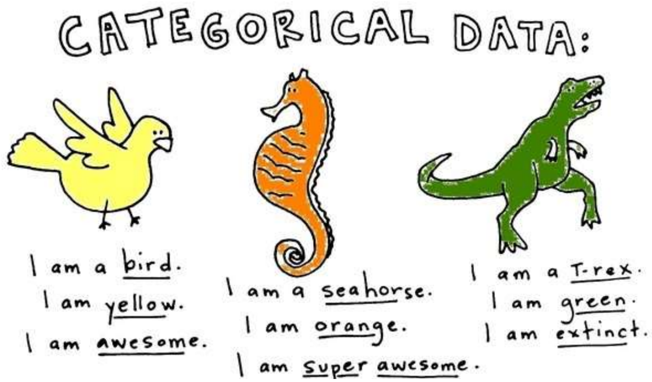
Terminology: Random Variables classification

Qualitative Variable – Non numerical characteristic, attribute

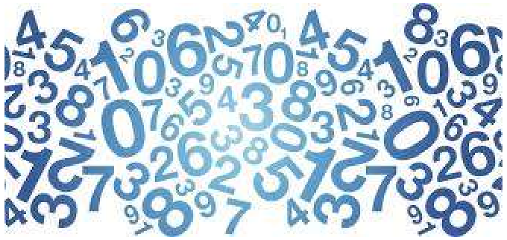
Nominal – Words without a logical natural ordering (e.g. marital status, ethnicity)
Ordinal – Words with logical natural ordering (e.g. perceived pain, education)

Some nominal variables could become ordinal (e.g. patient diseases can be ranked according to a severity of the disease), thus the boundary between the two definitions becomes flexible.

Qualitative variables are also called categorical variables.



<u>Terminology:</u> Random Variables classification



Quantitative Variable – numerical characteristic

Discrete – natural numbers obtained by counts (e.g. number of sons or daughters)

Continuous – continuous numbers obtained by measurements (e.g. weight, income, mechanical strength)

Any continuous number is rounded to a unit measurement (e.g. weight in kilograms), thus the boundary between the two definitions becomes flexible.

Discrete variables usually assumes a limited number of values, whereas continuous variables assumes a relatively big number of values even under this rounding operation.

Terminology: Variables classification

	Remdesivir group (n=158)	Placebo group (n=78)
Age, years	66·0 (57·0–73·0)	64·0 (53·0–70·0)
Sex		
Men	89 (56%)	51 (65%)
Women	69 (44%)	27 (35%)
Any comorbidities	112 (71%)	55 (71%)
Hypertension	72 (46%)	30 (38%)
Diabetes	40 (25%)	16 (21%)
Coronary heart disease	15 (9%)	2 (3%)
Body temperature, °C	36·8 (36·5–37·2)	36·8 (36·5–37·2)
Fever	56 (35%)	31 (40%)
Respiratory rate >24 breaths per min	36 (23%)	11 (14%)
White blood cell count, $\times 10^9$ per L		
Median	6·2 (4·4–8·3)	6·4 (4·5–8·3)
4–10	108/155 (70%)	58 (74%)
<4	27/155 (17%)	12 (15%)
>10	20/155 (13%)	8 (10%)
Lymphocyte count, $\times 10^9$ per L	0·8 (0·6–1·1)	0·7 (0·6–1·2)
≥1·0	49/155 (32%)	23 (29%)
<1·0	106/155 (68%)	55 (71%)
Platelet count, $\times 10^9$ per L	183·0 (144·0–235·0)	194·5 (141·0–266·0)
≥100	148/155 (95%)	75 (96%)
<100	7/155 (5%)	3 (4%)

SCHOOL OF MEDICINE AND SURGERY

Which variable in table 1 is Quantitative-continuous?

1. **Sex (Males or Female)**
2. **Hypertension (yes or no)**
3. **Blood pressure (mmHg)**
4. **COPD (yes or no)**
5. **Respiratory support (Oxygen Mask, Noninvasive ventilation, Invasive mechanical ventialtion)**
6. **PEEP**
7. **No. of patients with PEEP measurement**

Descriptive Methods

Descriptive statistics are methods for organizing and summarizing data

For example, tables or graphs are used to organize data, and descriptive indicators such as mean, median are used to summarize data.

TEAM-BASED LEARNING 06/10 on descriptive statistics :

1. Read chapter “4. Summarising data” from the book “An Introduction to Medical Statistics, Martin Bland 2015.” (pdf in the course webpage)

Contents of the chapter:

4.1 Types of data
4.2 Frequency distributions
4.3 Histograms and other frequency graphs
4.4 Shapes of frequency distribution
4.5 Medians and quantiles
4.6 The mean
4.7 Variance, range, and interquartile range
4.8 Standard deviation

Frequency Tables

List of possible values assumed by the random variable with corresponding frequencies (absolute, relative, relative %)

Nominal categorical/qualitative variable – Values are listed according to a chosen ordering. This ordering will be used also for graphical representation.

Ordinal categorical/qualitative variable – Values are listed according to their natural ordering

Discrete quantitative variable – Values are listed in increasing order

Continuous quantitative variable – Values are aggregated in small intervals mutually exclusive. Values are listed in increasing ordering

Frequency Table of a categorical variable

SCHOOL OF MEDICINE AND SURGERY

$$p\% = \frac{frequency\ (f)}{sample\ size\ (n)} * 100$$

School degree	f	p%
Elementary	42	13.2
High school	105	32.9
College/University	172	53.9
	n=319	**100**

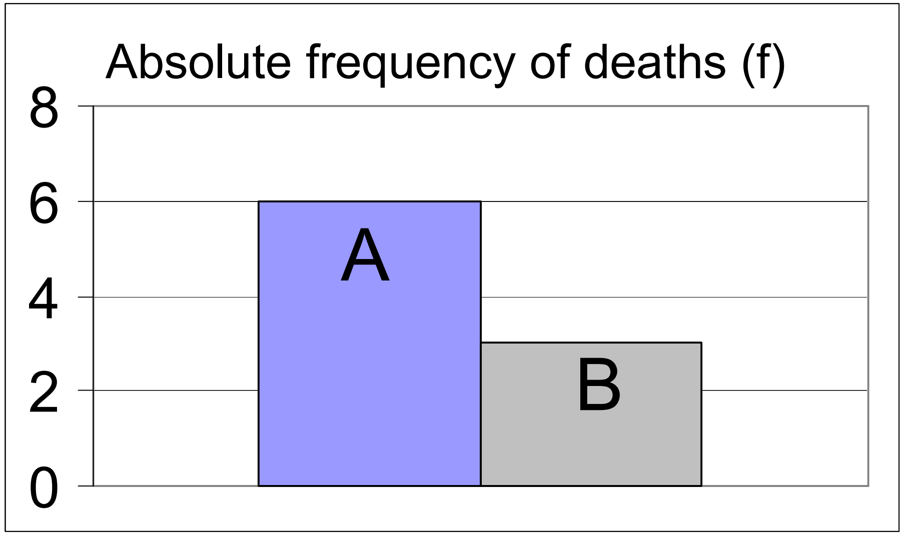
(absolute) frequencies — relative frequencies

What's the advantage of computing relative frequencies?

Example

With the aim of evaluating the efficacy of a new drug (B) on mortality after miocardial infarction (1 month), 150 patients were recruited. 100 patients were randomised to receive the standard drug A and 50 the new drug B.

	Drug A	Drug B
Dead	6	3
Alive	94	47
Total	100	50



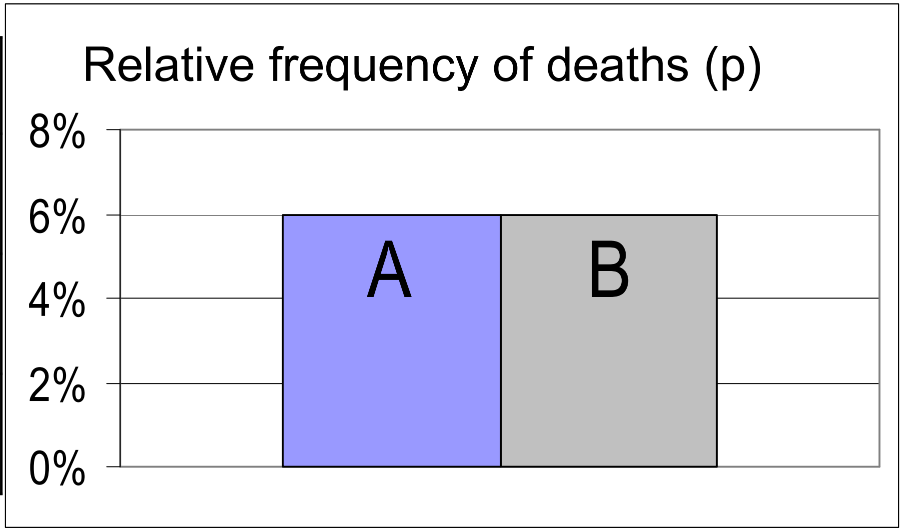
Would you recommend the new drug B to future patients?

Example

SCHOOL OF MEDICINE AND SURGERY

With the aim of evaluating the efficacy of a new drug (B) on mortality after miocardial infarction (1 month), 150 patients were recruited. 100 patients were randomised to receive the standard drug A and 50 the new drug B.

	Drug A	Drug B
Dead	6(6%)	3(6%)
Alive	94	47
Total	100	50



They provide the same information, but relative frequencies are useful to compare groups with different sample size and they facilitate the perception of modality's weight

But they can also be misleading:

SCHOOL OF MEDICINE AND SURGERY

“The antibiotic phosphomycin is advertised as being 100% effective in chronic urinary tract infections.”

This information is based on a trial recruiting 8 patients, after excluding patients whose urine contained phosphomycin-resistant bacteria.

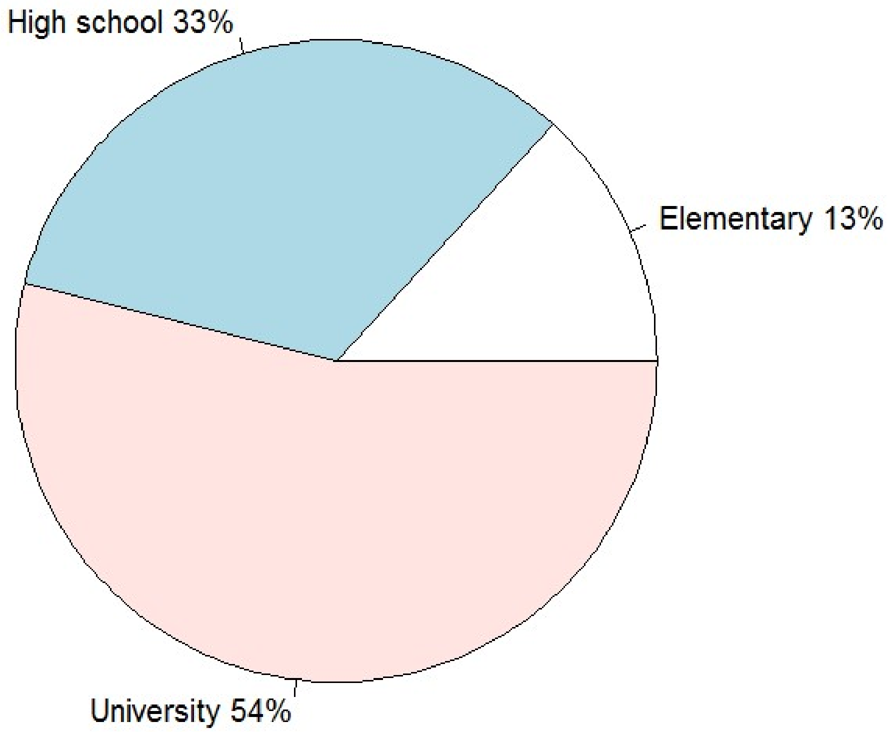
Relative frequencies should always be accompanied by the number on which they were calculated!

Graphical Representations

	Nominal	Ordinal	Discrete	Continuous
Pieplot	■	■	■	
Barplot	■	■	■	
Histogram				■
Distribution function (or ogive)				■
Boxplot				■

Pieplot

The angle of the slice is calculated by the product 360°*relative frequency



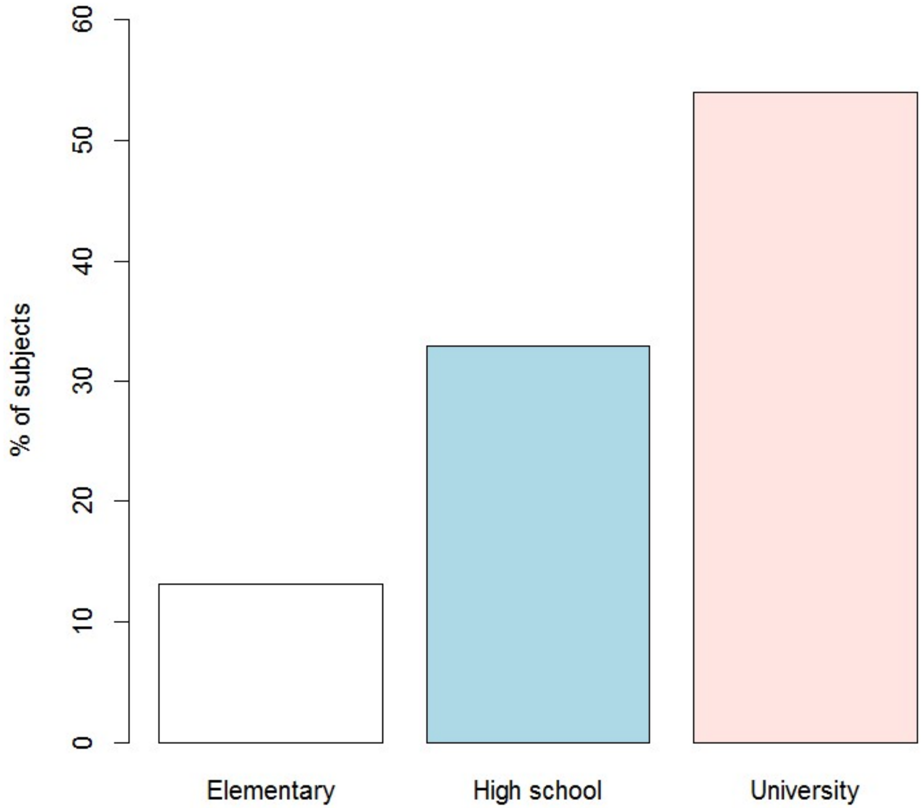
n=319

Note – Usable only for variables with a limited number of modalities

Barplot

The y height of the bars is equal to the frequency (absolute, relative, relative %).

The x axis has no numerical relevance.



n=319

Note – Usable even for variables with a big number of modalities

Frequency Table of a continuous variable?

SCHOOL OF MEDICINE AND SURGERY

In a survey conducted by a group of neonatologists, the values of supine length (cm) were found in a sample of 60 newborns. The measurements, taken with the Harpenden infantometer, are shown below.

51.0	46.5	48.7	54.5	46.0	51.2	55.0	50.2	44.5	56.3
49.4	47.8	50.0	48.2	52.2	51.1	50.2	53.4	49.2	46.5
49.0	49.7	52.9	48.9	47.0	54.7	50.3	47.4	50.5	51.5
52.5	44.4	50.8	51.2	50.8	52.3	47.7	50.5	49.5	50.9
51.5	49.8	46.2	49.5	50.0	48.2	48.5	51.7	52.9	51.6
51.8	53.0	48.9	54.0	52.5	50.8	53.8	49.5	50.5	52.7

?

Frequency Table of a continuous variable

SCHOOL OF MEDICINE AND SURGERY

aggregate values in small classes:

$$p\% = \frac{\text{frequency (f)}}{\text{sample size (n)}}$$

Lower Interval Limits

Upper Interval Limits

Interval of supine lenght (cm)		f		p%	
(44.25,45.75]		2		3.3	
(45.75,47.25]		5		8.3	
(47.25,48.75]		7		11.7	
(48.75,50.25]		14		23.3	
(50.25,51.75]		16		26.7	
(51.75,53.25]		9		15.0	
(53.25,54.75]		5		8.3	
(54.75,56.25]		1		1.7	
(56.25,57.75]		1		1.7	

(absolute) frequencies   relative frequencies

Intervals for a continuous variable

[44.25-45.75) o 44.25 |- 45.75

Interval closed on the left and open on the right

left extreme included

(44.25-45.75] o 44.25 - |45.75

Interval closed on the right and open on the left

right extreme included

[44.25-45.75] o 44.25 - 45.75

Interval closed both on the right and on the left

right and left extremes included

(44.25-45.75) o 44.25 - 45.75

Interval open both on the right and on the left

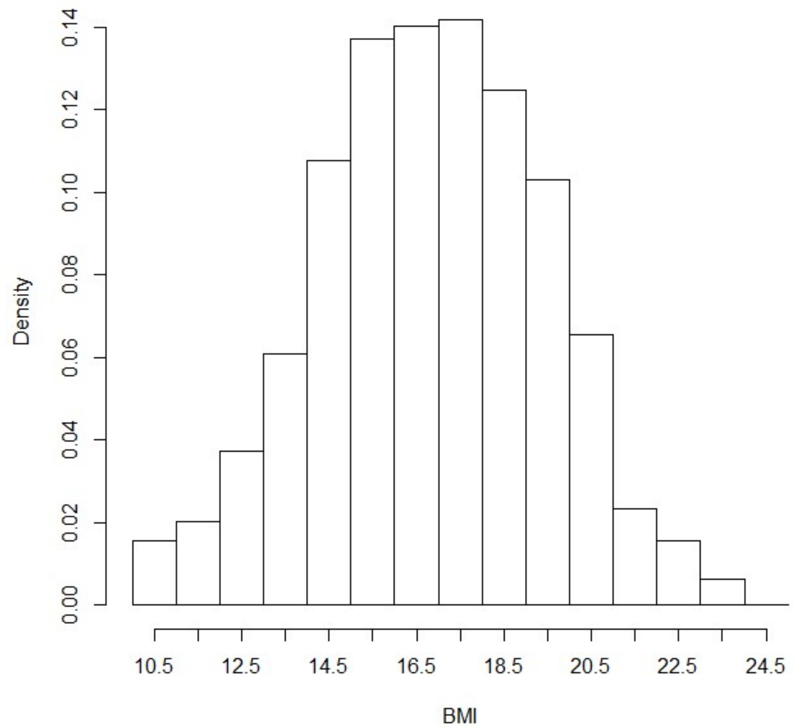
right and left extremes excluded

Histogram

The y height of the bars is equal to relative frequency/width of the class.

$$\text{density} = \frac{\text{relative frequency}}{\text{width of the class}}$$

The x width of the bars is that of the class. The x axis has numerical relevance.



Note1 – The area of each rectangle is equal to the relative frequency

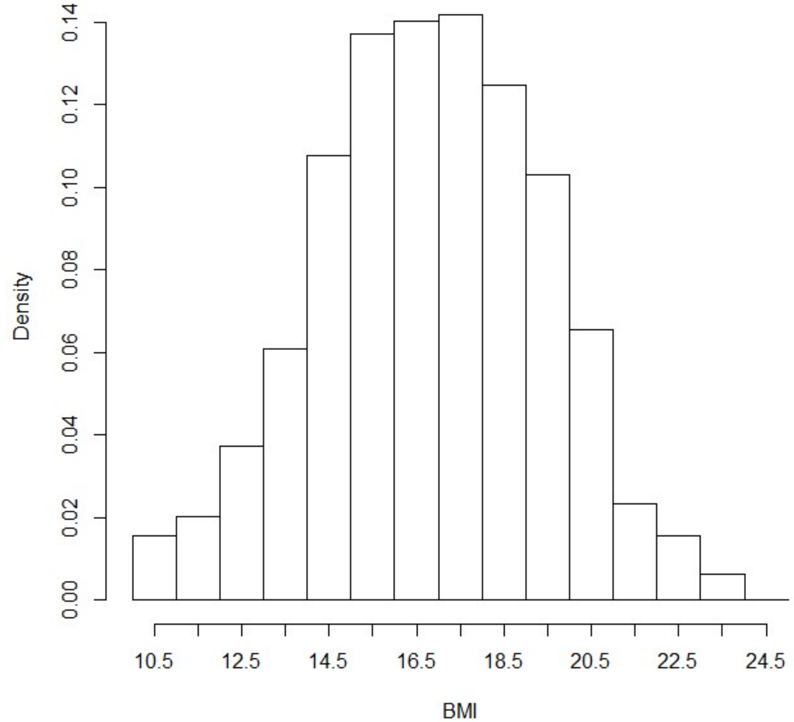
Note2 – The whole area of the rectangles is equal to one

Note3 – A bell shaped histogram can be approximated by a Gaussian distribution

Histogram

BMI kg/m^2	f	p
(10,11]	10	0.0156
(11,12]	13	0.0203
(12,13]	24	0.0374
(13,14]	39	0.0608
(14,15]	69	0.1076
(15,16]	88	0.1373
(16,17]	90	0.1404
(17,18]	91	0.1420
(18,19]	80	0.1248
(19,20]	66	0.1030
(20,21]	42	0.0655
(21,22]	15	0.0234
(22,23]	10	0.0156
(23,24]	4	0.0062
(24,25]	0	0.0000
	641	

$$\text{density}=\frac{\text{relative frequency(p)}}{\text{width of the class}}$$



Note1 – The area of each rectangle is equal to the relative frequency

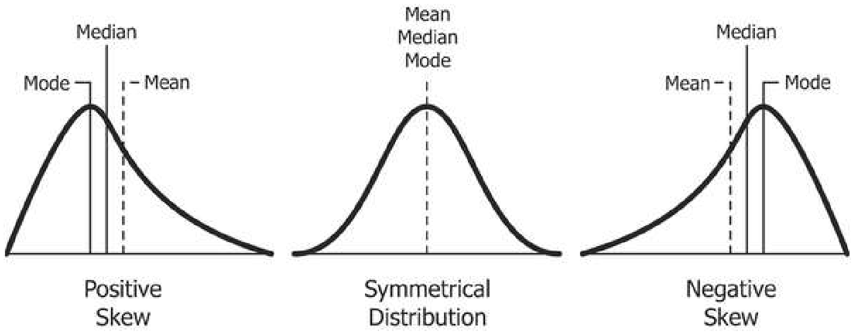
Note2 – The whole area of the rectangles is equal to one

Note3 – A bell shaped histogram can be approximated by a Gaussian distribution

Shape of data is measured by

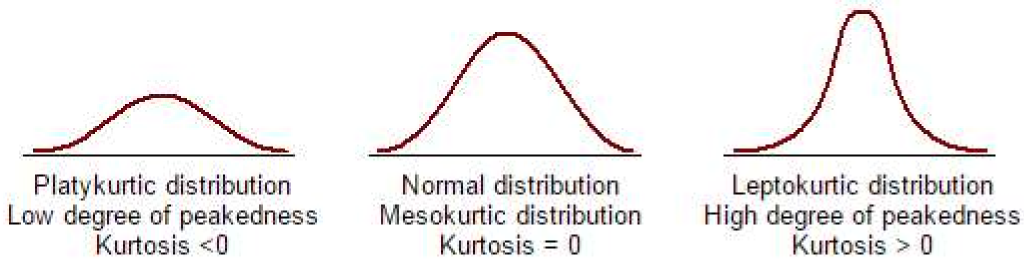
- ✓ Skewness

  Positive or right skewed: Longer right tail
  Negative or left skewed: Longer left tail



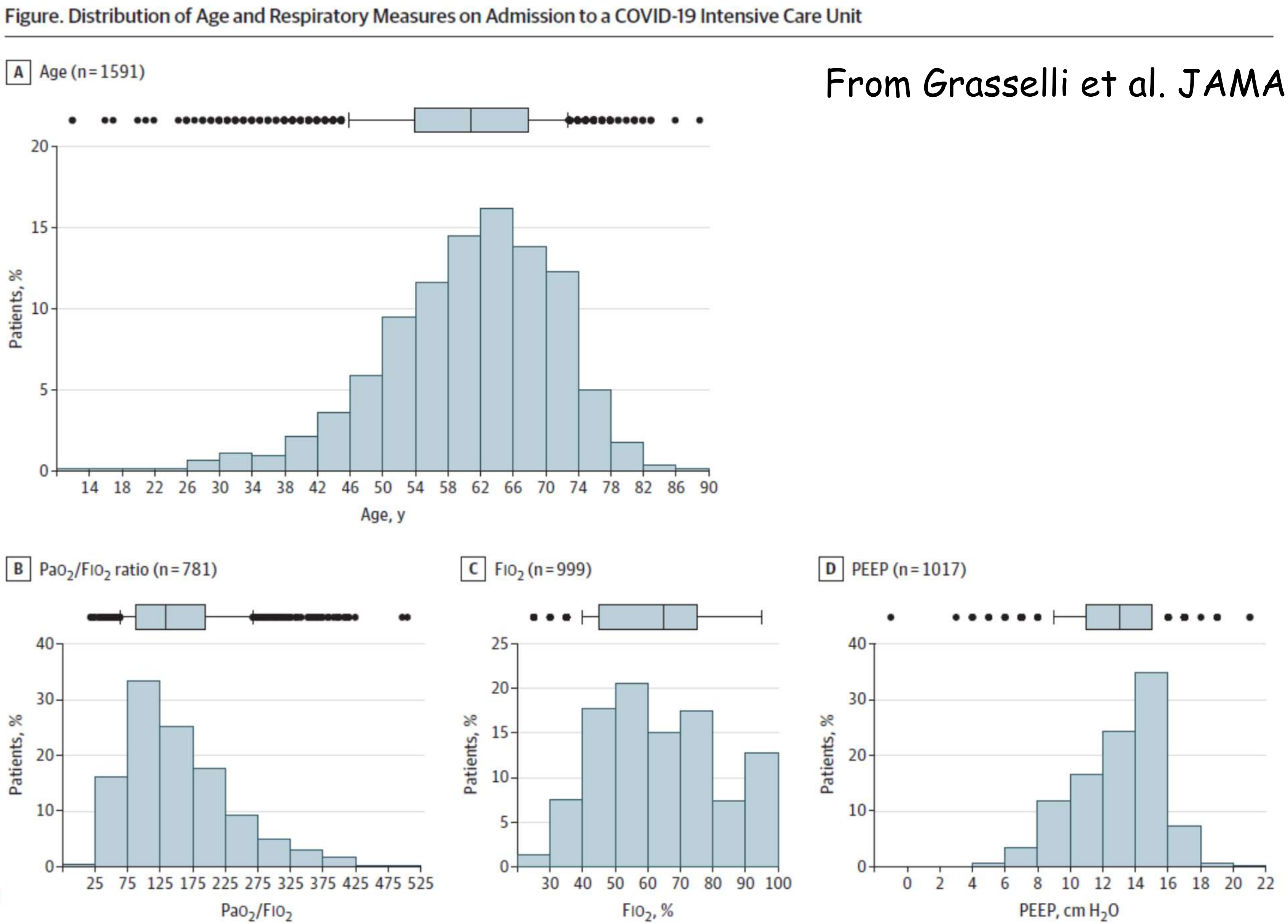
- ✓ Kurtosis

  Measures peakedness of the distribution of data.
  The kurtosis of normal distribution is 0



4.3 Histograms and other frequency graphs

From Grasselli et al. JAMA 2020



4.4 Shapes of frequency distribution

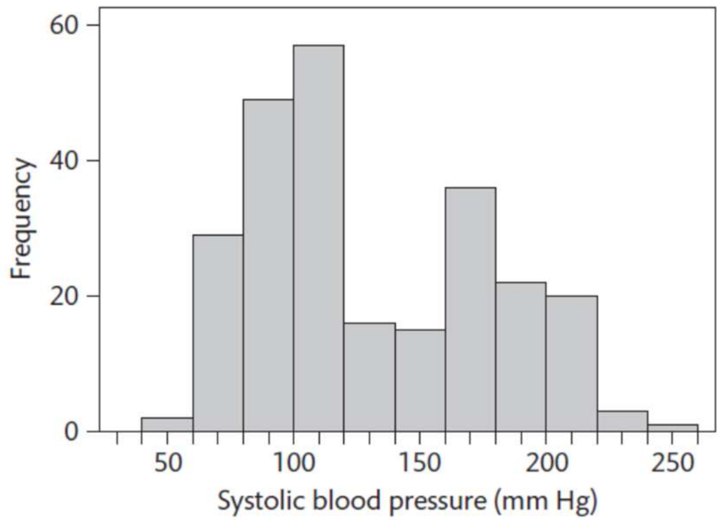


Figure 4.13 Systolic blood pressure in a sample of patients in an intensive therapy unit (data from Friedland *et al.* 1996).

(Cumulative) Distribution function – Ogiva

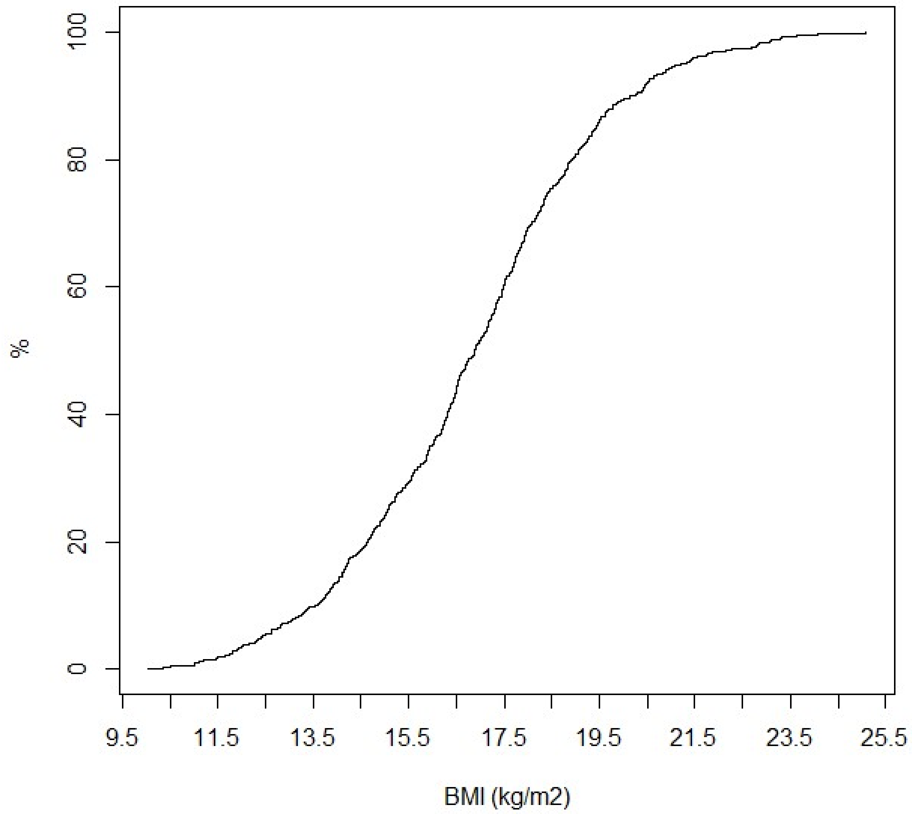
The x values are the ordered variable values.

The x axis has numerical relevance.

The y value corresponding to x is [sum(data <=x)/sample size] *100

Cumulative relative frequency % without collapsing data into classes.

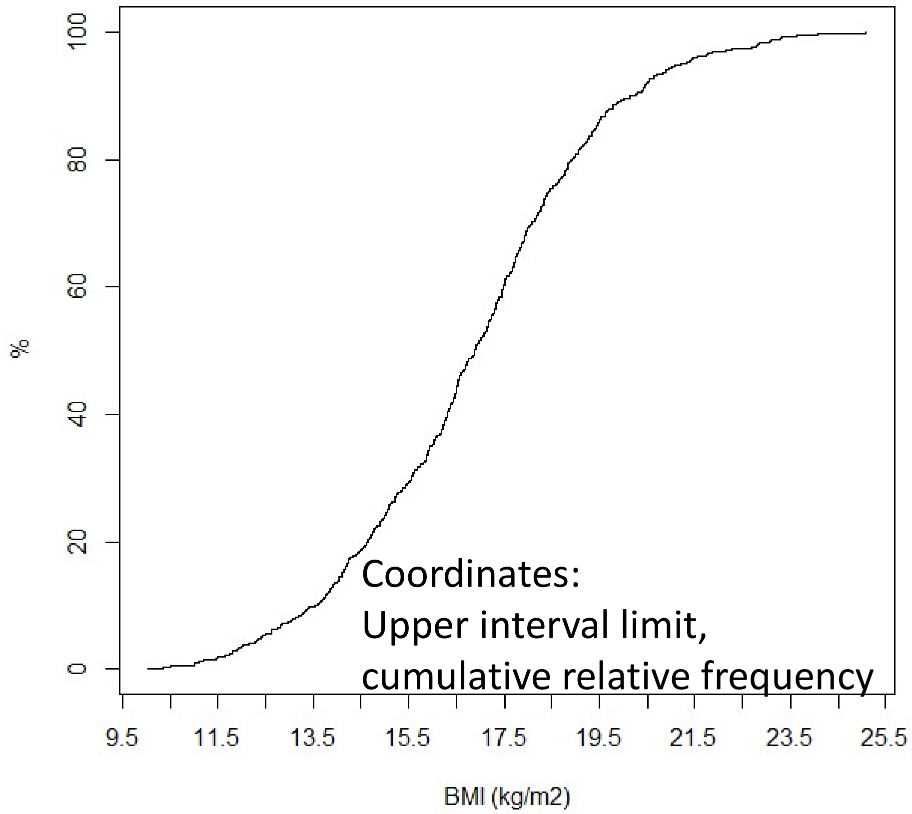
The y value corresponding to x is an estimate of P(X<=x)=F(x)



(Cumulative) Distribution function

$$P(x) = \sum_{i=mi}^{x} p_i$$

BMI kg/m^2	f	p	P
(10,11]	10	0.0156	0.0156
(11,12]	13	0.0203	0.0359
(12,13]	24	0.0374	0.0733
(13,14]	39	0.0608	0.1342
(14,15]	69	0.1076	0.2418
(15,16]	88	0.1373	0.3791
(16,17]	90	0.1404	0.5195
(17,18]	91	0.1420	0.6615
(18,19]	80	0.1248	0.7863
(19,20]	66	0.1030	0.8892
(20,21]	42	0.0655	0.9548
(21,22]	15	0.0234	0.9782
(22,23]	10	0.0156	0.9938
(23,24]	4	0.0062	1.0000
(24,25]	0	0.0000	1.0000
	641		

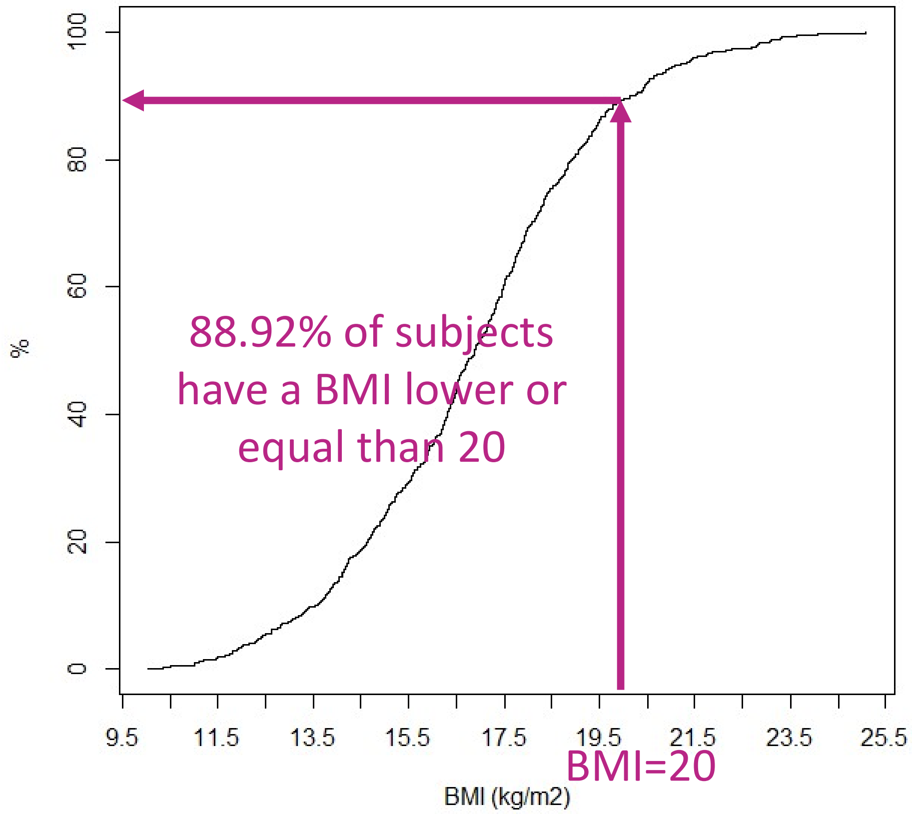


The y value corresponding to x is an estimate of P(X<=x)=F(x)

(Cumulative) Distribution function – in practice

Note1 – It has always a monotonic shape (boring…)

Note2 – It enables to calculate quickly the % of the sample with value less or equal (or greater) than a given threshold (e. g. BMI=20).



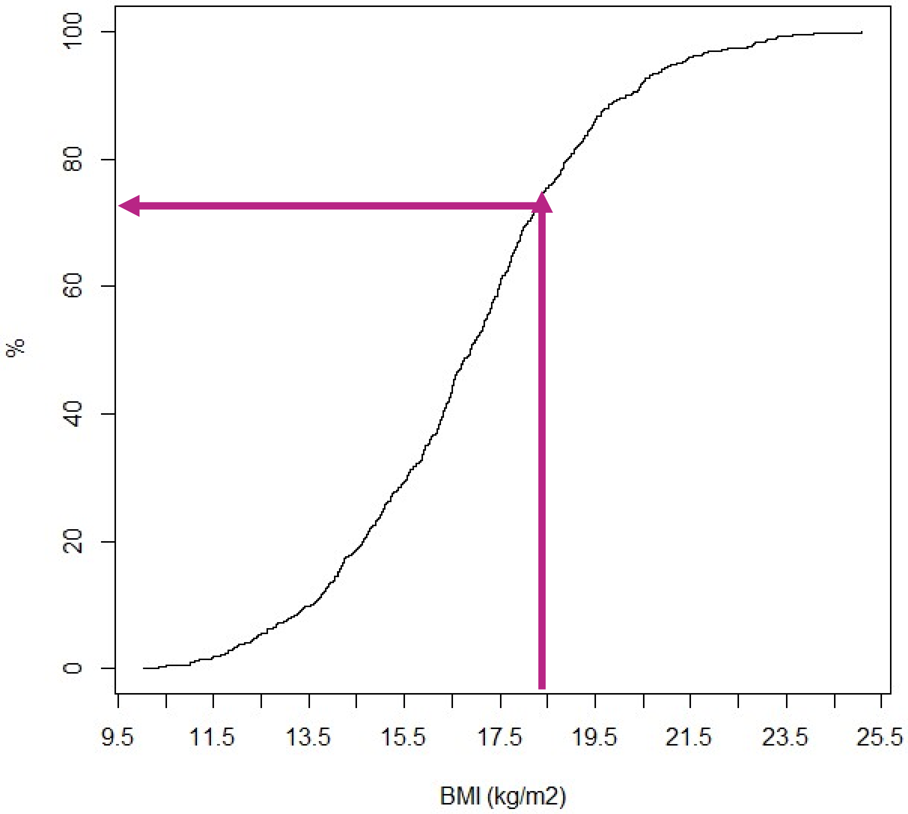
(Cumulative) Distribution function - in practice

SCHOOL OF MEDICINE AND SURGERY

Note2 – It enables to calculate quickly the % of the sample with value greater (or less or equal) than a given (clinical/epidemiological) threshold.

Useful to categorize continuous variables according to standard cut-points (e.g. 18.5, 25 thresholds for BMI kg/m^2) and to obtain directly the frequencies

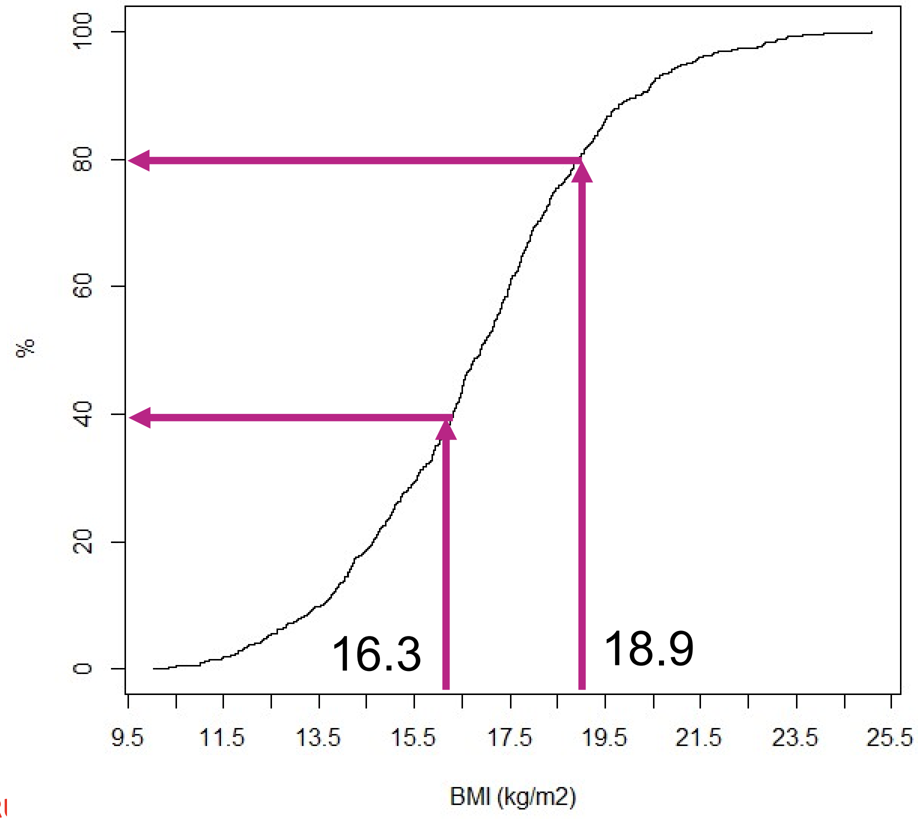
BMI kg/m2	f	p%
<=18.5	484	75.62
>18.5	156	24.38
	641	



(Cumulative) Distribution function - in practice

SCHOOL OF MEDICINE AND SURGERY

Useful to measure the placement of a measure with respect to the overall distribution – percentile (the pth percentile is the value which is greater than P% of the data).



UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

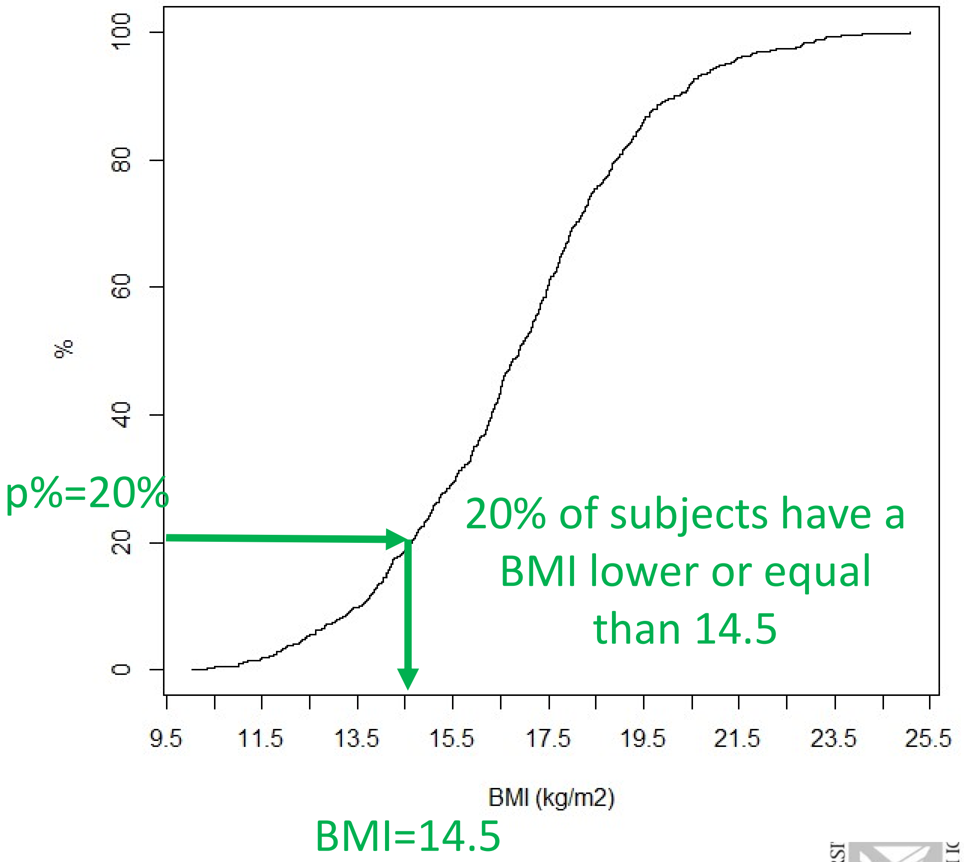
(Cumulative) Distribution function – in practice

SCHOOL OF MEDICINE AND SURGERY

Note1 – It has always a monotonic shape (boring…)

Note2 – It enables to calculate quickly the % of the sample with value less or equal (or greater) than a given threshold (e. g. BMI=20).

Note3 – It enables to calculate quickly the x value such that a given % of the sample has value less or equal (or greater) than x (e.g. BMI=14.5) - quantiles



UNIVERSITÀ DI MILANO BICOCCA

Summary indicators (position and variability)

	Nominal	Ordinal	Discrete	Continuous
Modal value	■	■	■	■
Mean			■	■
Median		■	■	■
Standard Deviation			■	■
Interquartile range		■	■	■
Range		■	■	■

Measures of Central Tendency (Location): the median

If the sample data are arranged in increasing order, the **median** is
- the <u>middle</u> value if *n* is an odd number, or
- <u>midway</u> between the two middle values if *n* is an even number

Exercise: find the median of height for the following data

Dispersion index: the interquartile range

The median divides a distribution into two halves.

The first and third quartiles (denoted Q1 and Q3) are defined as follows:

- ✓ 25% of the data lie below Q1 (and 75% is above Q1),
- ✓ 25% of the data lie above Q3 (and 75% is below Q3)

The inter-quartile range (IQR) is the difference between the first and third quartiles, i.e.
IQR = Q3- Q1

Example

The ordered blood pressure data is:

113 124 124 132 146 151 170

Q_1 (124), Q_2 (132), Q_3 (151)

Inter Quartile Range (IQR) is 151-124 = 27

Percentiles definition

The percentile x_p ($0<p<1$) of the distribution of a continuous variable is that value of the variable that satisfies these conditions:

1. p% of the observations assume values ≤ of x_p,
2. the (1-p)% of the observations take values> of x_p

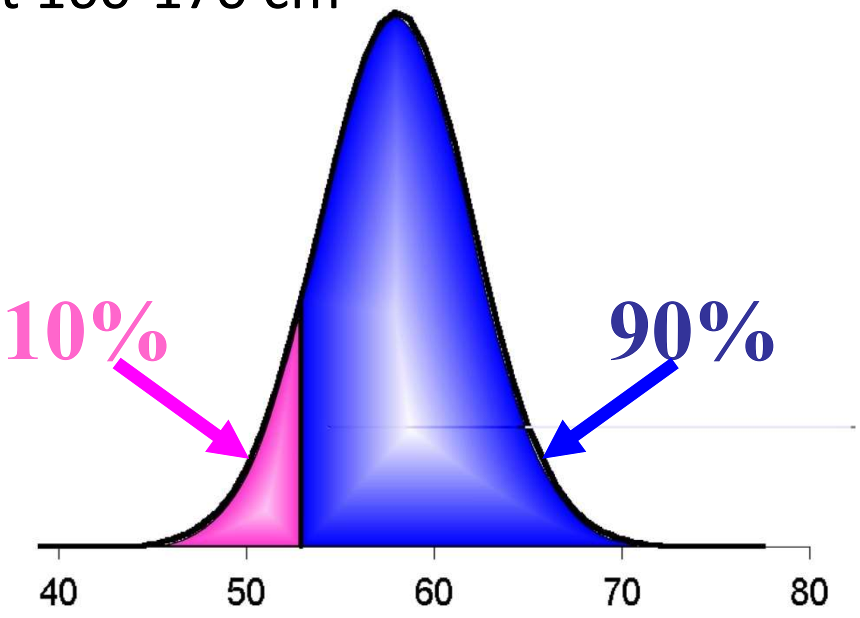
Percentiles are useful for:

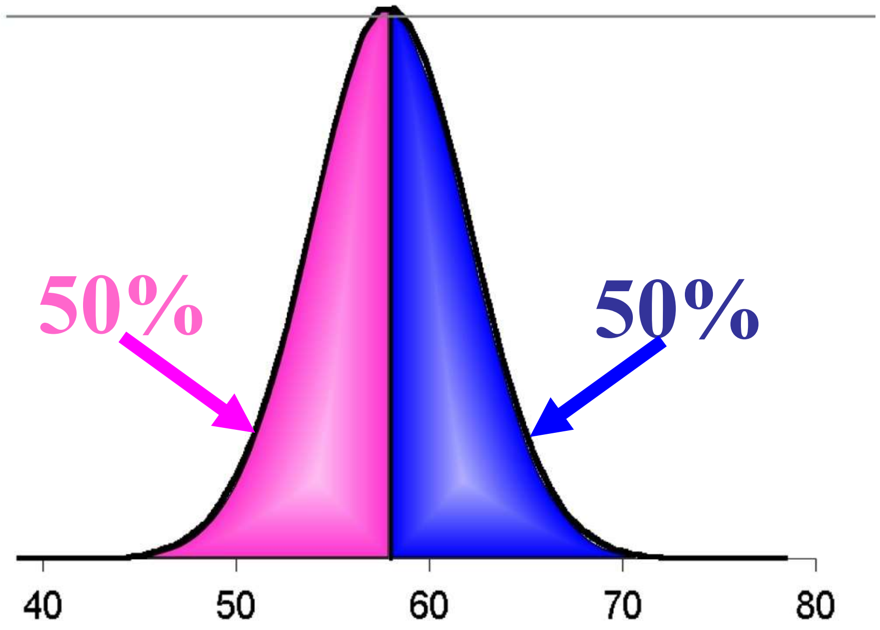
- Describe a distribution
- Identify normal range
- Classify the value of a subject with respect to the distribution of the phenomenon

Percentiles definition

Example: Weight of women of height 160-170 cm

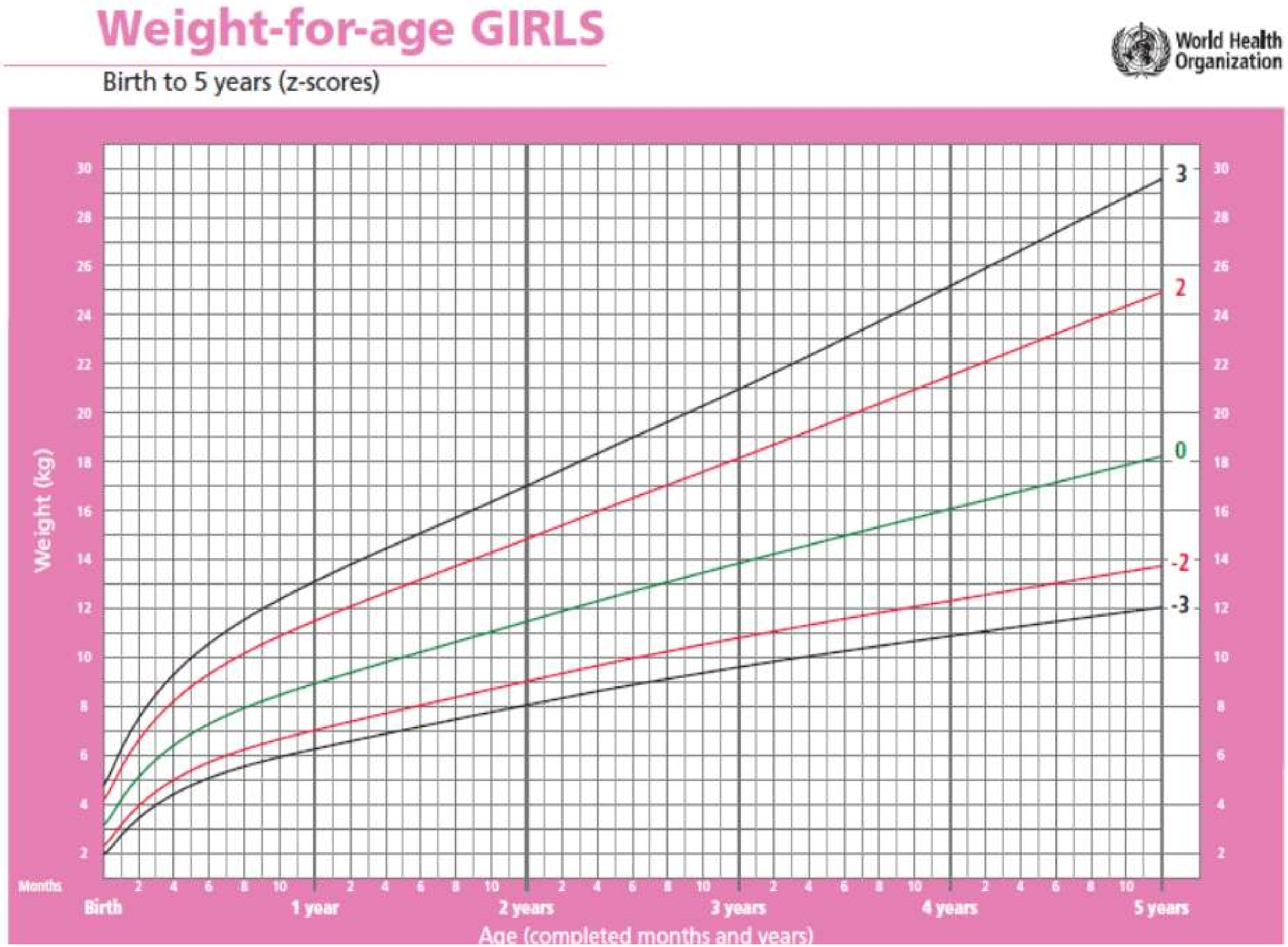
$p = 0.10 \quad x_{0.10} = 53$





$p = 0.50 \quad x_{0.50} = 58$

(Cumulative) Distribution function – percentiles



3=99.5p

2=97.5p

0=50p

-2=2.5p

-3=0.5p

Note – Only approximated percentiles can be calculated

Boxplot

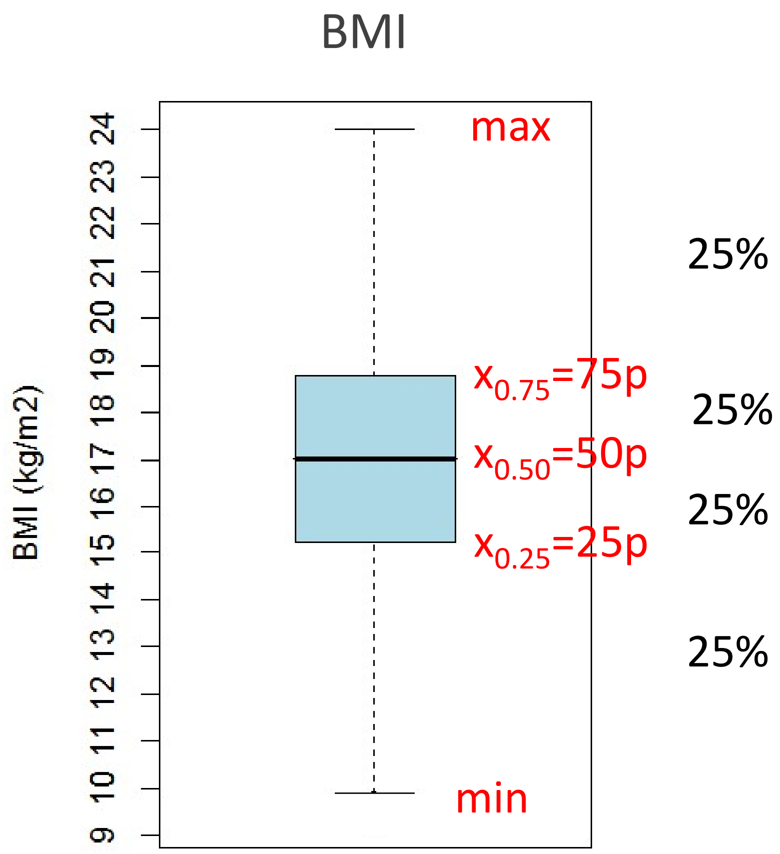
The y bottom base is the 25p percentile– the value such as 25% of the sample has a value <= 25p

The y top base is the 75p– the value such as 75% of the sample has a value <= 75p

The black line is the 50p (median) – the value such as 50% of the sample has a value <= 50p

The bottom and top whiskers are minimum and maximum

The x axis has no numerical relevance



Note – For a sample with limited sample size, raw data can be shown as points.

Boxplot : example

SCHOOL OF MEDICINE AND SURGERY

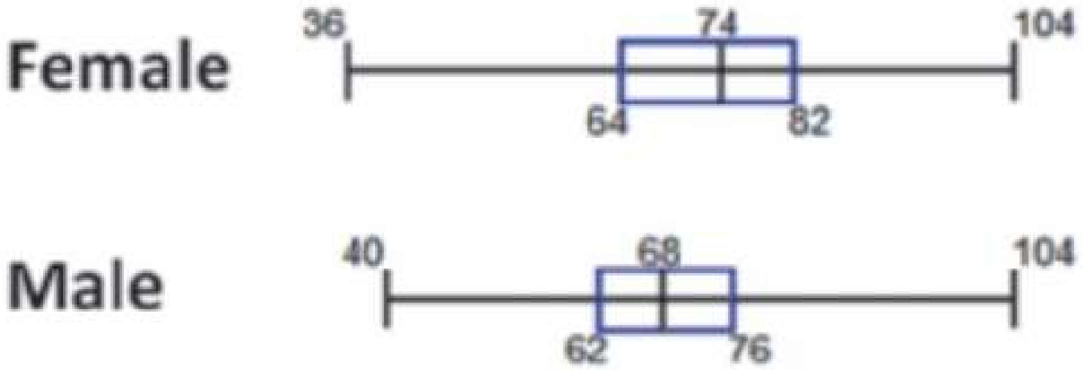


FIGURE 3-7 **Boxplots of Pulse Rates of Men and Women**

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Outliers

An **outlier** is an observation which does not appear to belong with the other data as it stands very far away from the most of other observations.

Can arise because of a measurement or recording error or because of equipment failure during an experiment, etc.

An outlier might be indicative of a sub-population, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.

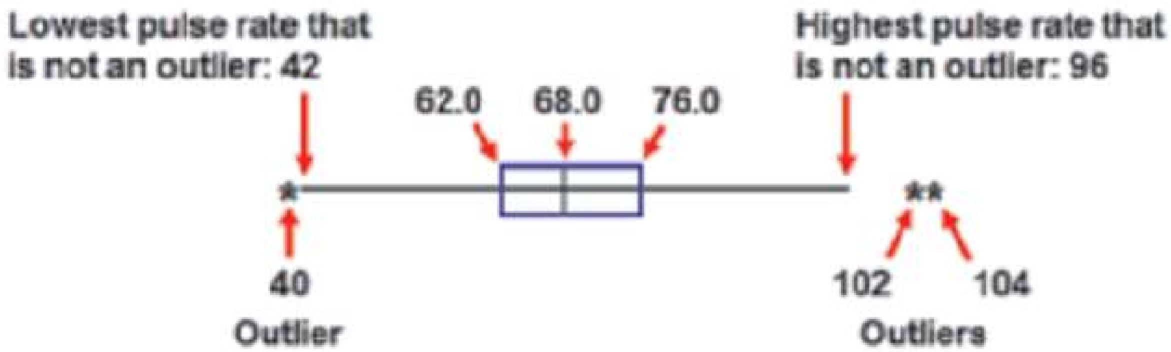


FIGURE 3-8 Modified Boxplot of Male Pulse Rates (BPM)

NOTE: What "very far away" means?

e.g. observations above 75p and below 25p by an amount greater than 1.5*(75p-25p) (Tukey)

Outliers

What “very far away” means?
e.g. observations above 75p and below 25p by an amount greater than 1.5*(75p-25p) (Tukey) :

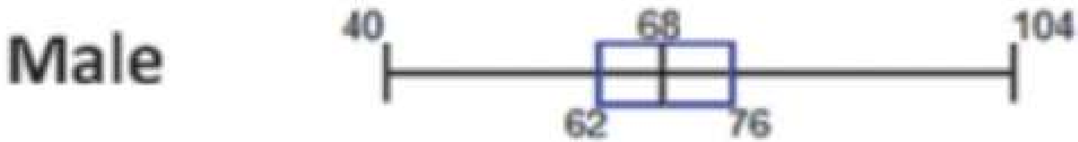


FIGURE 3-7 Boxplots of Pulse Rates of Men and Women

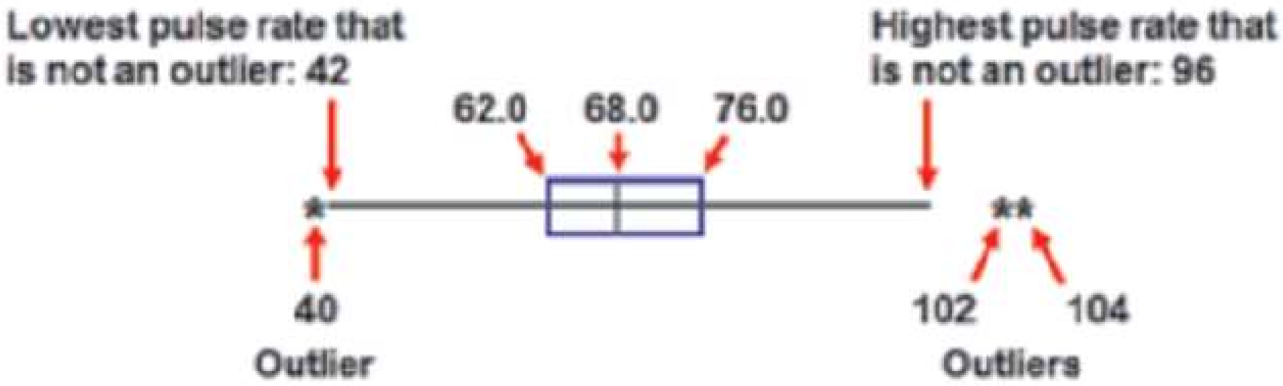


FIGURE 3-8 Modified Boxplot of Male Pulse Rates (BPM)

Measures of Central Tendency (Location): The arithmetic mean

The **mean** Let $x_1, x_2, x_3, \ldots, x_n$ be the realised values of a ***quantitative*** random variable **X**, from a sample of size **n.** The **sample arithmetic mean** is defined as:

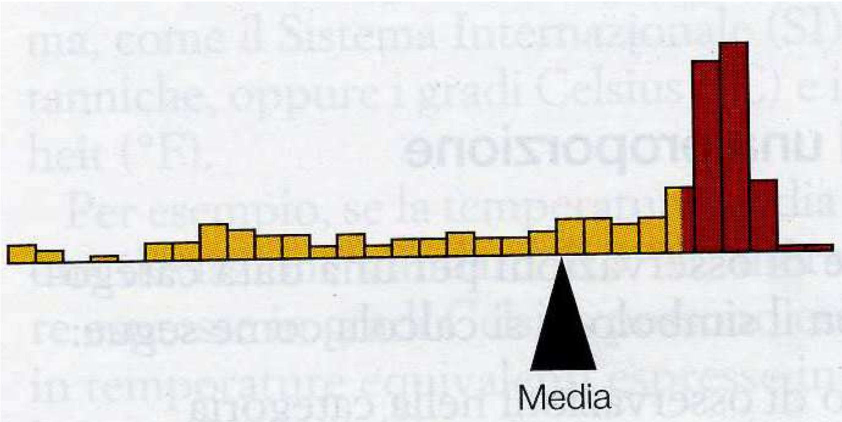
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n)$$

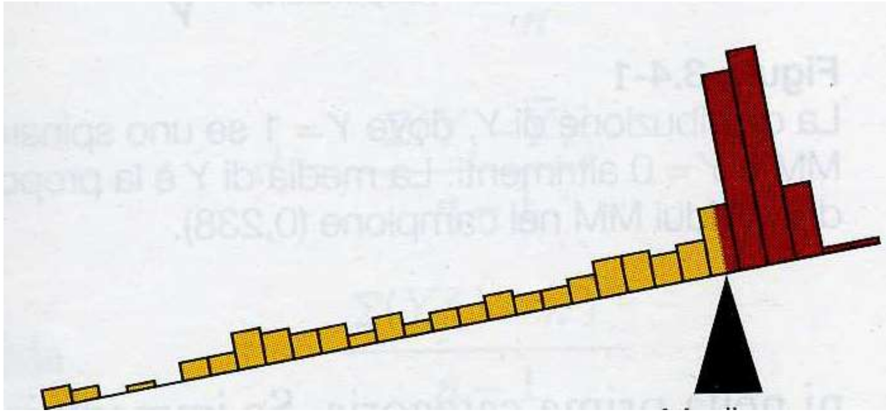
Properties:

1. It is the center of gravity of the distribution
2. It is always between the smallest and largest of the observed values
3. The average of a set of observations organized in k groups is equal to the weighted average of the partial averages with weights equal to the number of subgroups
4. The sum of the differences in the observations from the average is null

Properties o the arithmetic mean:

1. center of gravity of the distribution

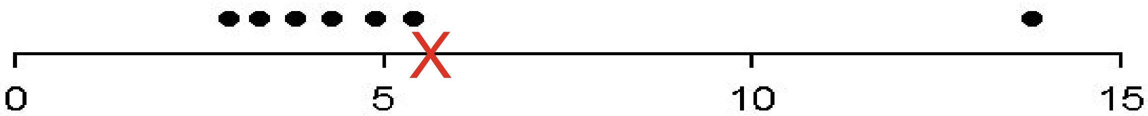




⇒ **strong dependence on extreme values**

Example: {2.9, 3.3, 3.8, 4.3, 4.9, 5.4, 13.8}

$$\bar{x} = (2.9+3.3+3.8+4.3+4.9+5.4+13.8)/7=38.4/7=5.5$$



Properties o the arithmetic mean:

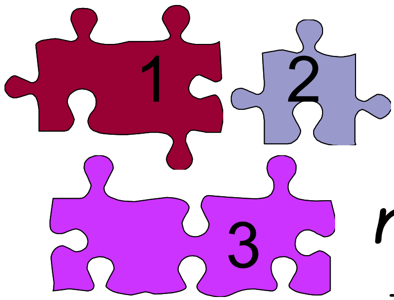
2. It is between the smallest and largest of the observed values:

Properties o the arithmetic mean:

3. The average of a set of observations organized in k groups is equal to the weighted average of the partial averages with weights equal to the size of subgroups.

High (mt) of 85 boys in 3 different classroms

$$n_1 = 20 \qquad \bar{x}_1 = 1.68$$



$$n_2 = 15 \qquad \bar{x}_2 = 1.60$$

$$n_3 = 50 \qquad \bar{x}_3 = 1.90$$

$$\bar{x} = \frac{1.68 \cdot 20 + 1.60 \cdot 15 + 1.90 \cdot 50}{85} = \frac{152.6}{85} = 1.80$$

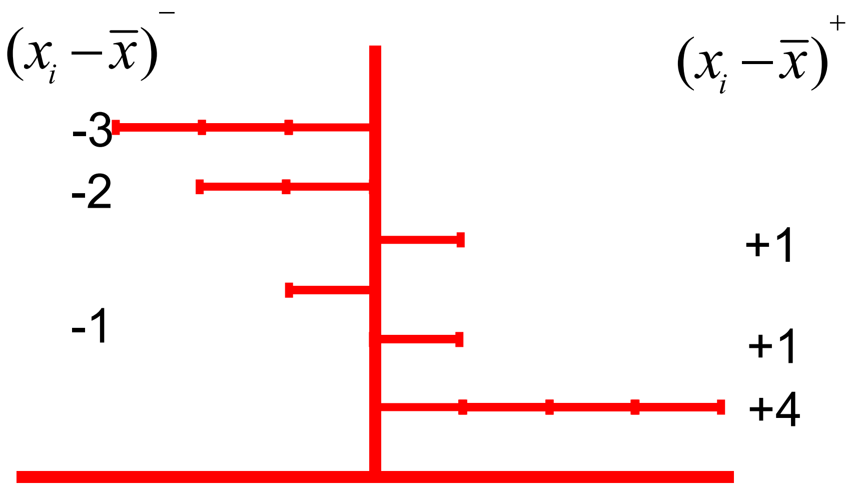
Properties o the arithmetic mean:

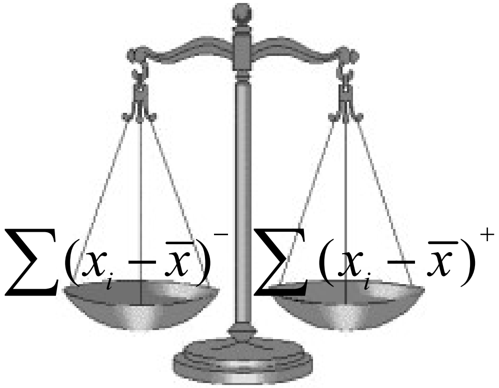
4. The sum of the differences in the observations from the average is null

Exam grades: $\{23, 24, 27, 25, 27, 30\}$ $\bar{x} = 26$

(23-***26***)+(24-***26***)+…

$\bar{x} = 26$





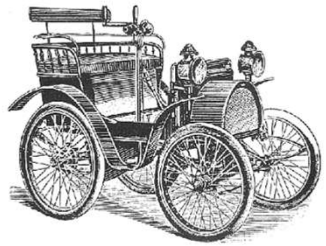
$\sum (x_i - \bar{x})^- = -6$ $\sum (x_i - \bar{x})^+ = +6$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Exercise:

Number of cars per family

Calculate the mean



x_i	f_i
0	2
1	20
2	25
3	3
Tot.	50

Using f:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \frac{0 \cdot 2 + ...3 \cdot 3}{50} = \frac{79}{50} = 1.6$$

Using p:

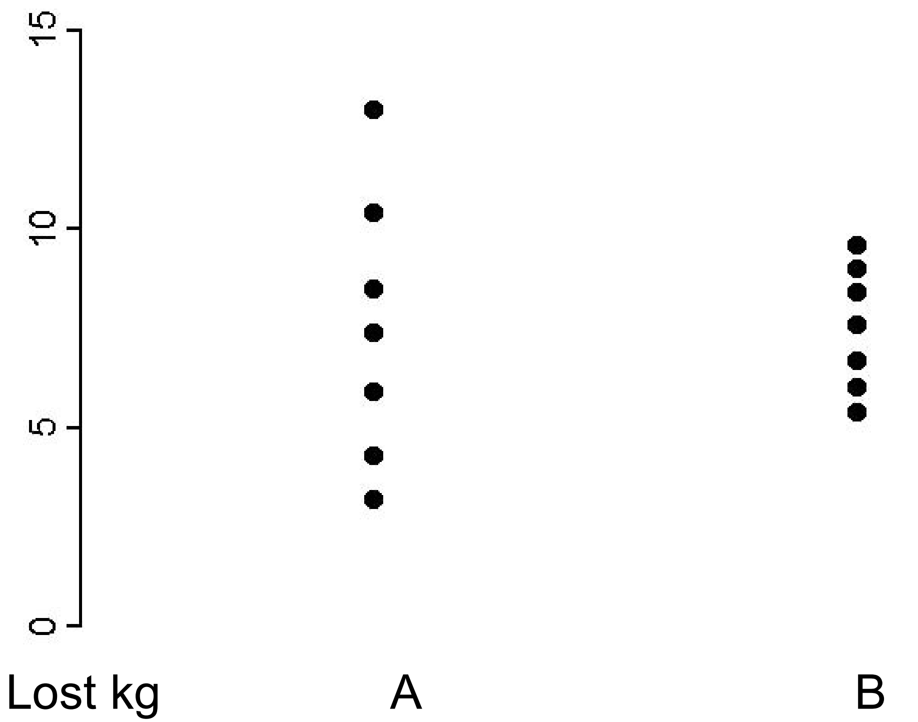
$$\bar{x} = \sum_{i=1}^{k} x_i \cdot p_i = 0 \cdot 0.04 + ...3 \cdot 0.06 = 1.6$$

Is it the mean sufficient to describe a sample?

Is it more effective a diet (A) or a pharmacological treatment (B) to decrease weight?

To assess the extent of weight loss (kg) that occurs after treatment (A or B), 14 comparable subjects were considered.

Intervention	
A	**B**
13.0	8.4
3.2	5.4
7.4	7.6
4.3	6.0
8.5	9.6
5.9	6.7
10.0	9.0



Which is the conclusion?

Dispersion index:

Measures of dispersion characterise how spread out the distribution is, i.e., how variable the data are.

We need a summary index for variability that:

1. uses all values in the sample
2. measures the dispersion about a «certain» value, i.e. the mean

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

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

but we can square them to get a positive result!

Dispersion index: the variance

SCHOOL OF MEDICINE AND SURGERY

The **VARIANCE** is the average of the squares of the deviation of the single observations from the sample mean:

Sample variance(s^2)=sum of(single observation-sample mean)^2/(sample size-1)

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

- it can take strictly positive values;
- it is null in the absence of variability; (eg 3.5, 3.5, 3.5, 3.5)
- it is higher as data are dispersed in a wide range of values;
- It is strongly influenced by the presence of extreme data due to the fact that the squares of distances are used;
- It has the square of the scale of the phenomenon per unit of measuremen.

Dispersion index: the standard deviation

To get an indicator with the same unit of measurement of the mean one takes the square root of the variance

Using single values:

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

Using frequency tables:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1}}$$

diet A	
$_{A}x_i$	$({}_{A}x_i - \overline{x}_A)^2$
13.0	30.57
3.2	18.25
7.4	0.01
4.3	10.06
8.5	1.06
5.9	2.47
10.0	6.39

$$\text{Sum} = 68.79$$

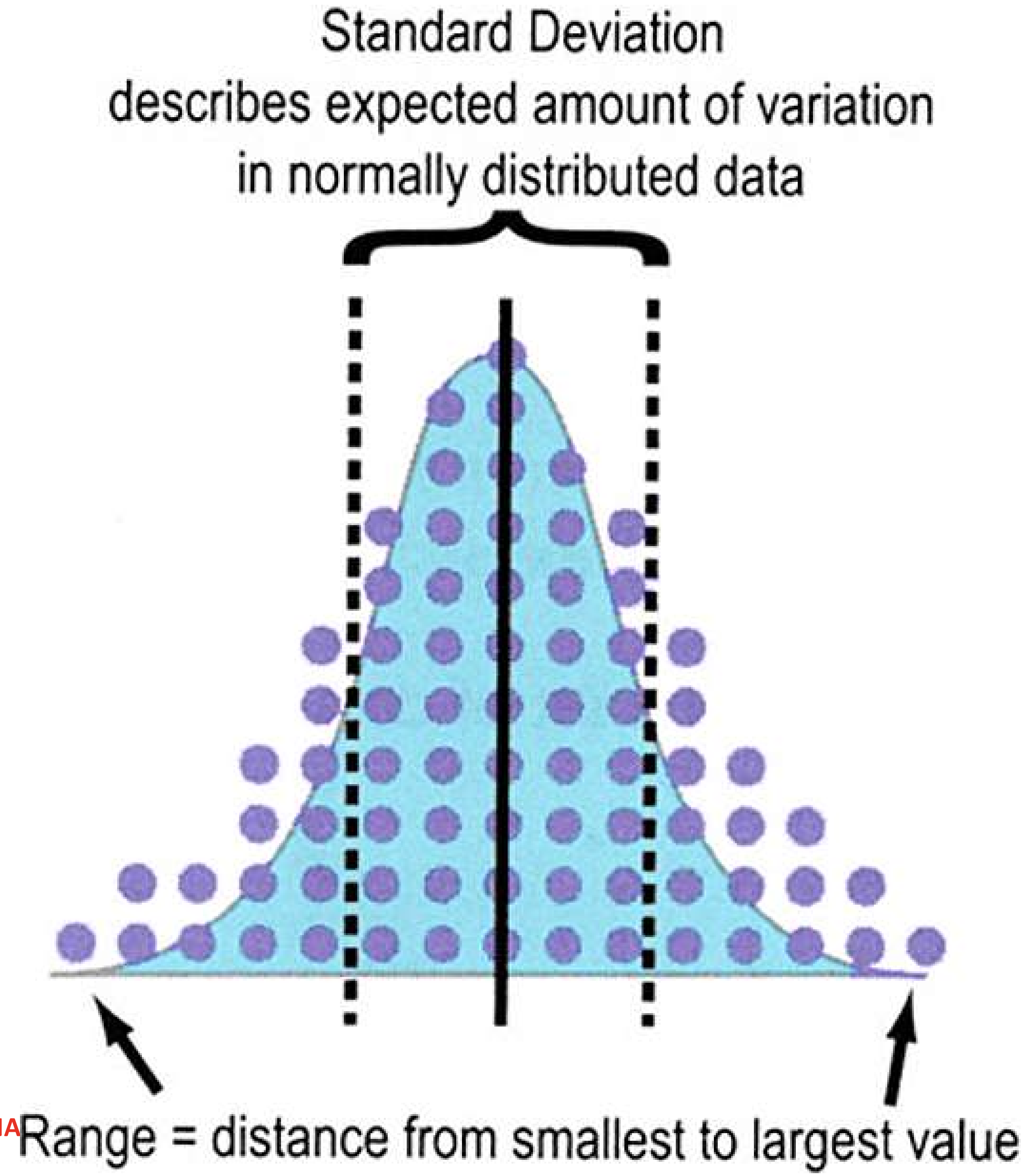
drug B	
$_{B}x_i$	$({}_{B}x_i - \overline{x}_B)^2$
8.4	0.76
5.4	4.54
7.6	0.01
6.0	2.34
9.6	4.28
6.7	0.69
9.0	2.16

$$\text{Sum} = 14.78$$

$$s_A = \sqrt{11.47} = 3.39$$

$$s_B = \sqrt{2.46} = 1.57$$

Dispersion index: the standard deviation



Exercise: compute mean and standard deviation

Use the weights of freshmen males in September to construct a frequency distribution. Begin with a lower class limit of 50 kg and use a class width of 10 kg.

interval	f		
[50;60)	2		
[60;70)	11		
[70;80)	13		
[80;90)	2		
[90;100)	4		
total			

1. Calculate the sample mean
2. Calculate the standard deviation

Variance

The sample **variance**, **s²**, is the arithmetic mean of the squared deviations from the sample mean:

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

Standard deviation

The sample **standard deviation**, s**,** is the square-root of the variance:

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

***s** has the advantage of being in the same units as the original variable **x**

Note1 - The standard deviation gives a rough estimate of the typical distance of a data values from the mean

Note2 - The larger the standard deviation, the more variability there is in the data and the more spread out the data are

Mean and standard deviation are summary indexes of location and variability

The 68% of observation lie within 1 standard deviation (s) $[\bar{x} - s \quad , \quad \bar{x} + s]$

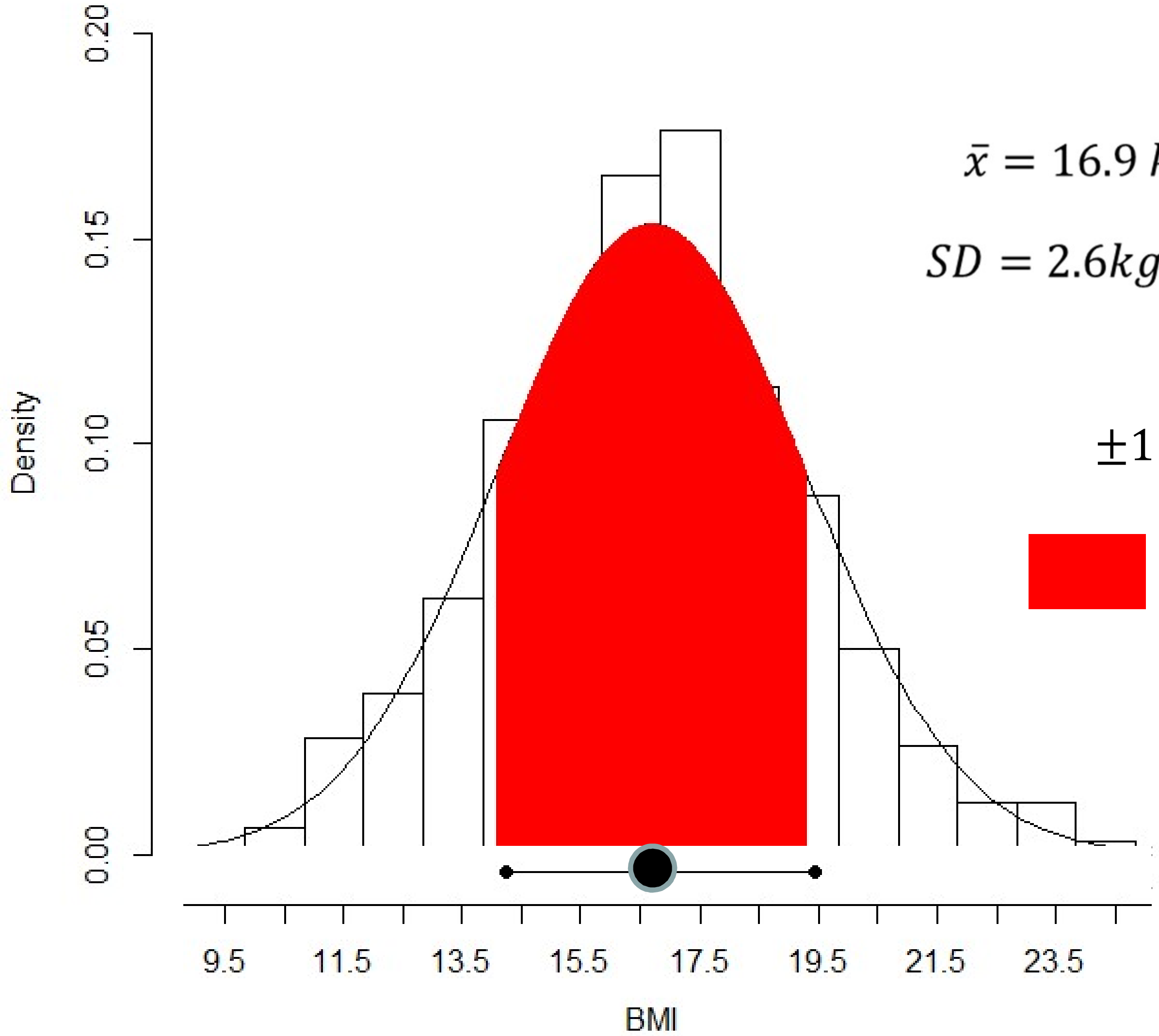
The 95% of observation lie within 2 ***s*** $[\bar{x} - 2 \cdot s \quad , \quad \bar{x} + 2 \cdot s]$

The 99% of observation lie within 1 ***s*** $[\bar{x} - 3 \cdot s \quad , \quad \bar{x} + 3 \cdot s]$

They are suitable only for representing symmetric distributions (with approximately normal shape)

They allow comparison between phenomena in the same unit of measurement and with the same order of magnitude

Histogram with Gaussian approximation – intervals around μ



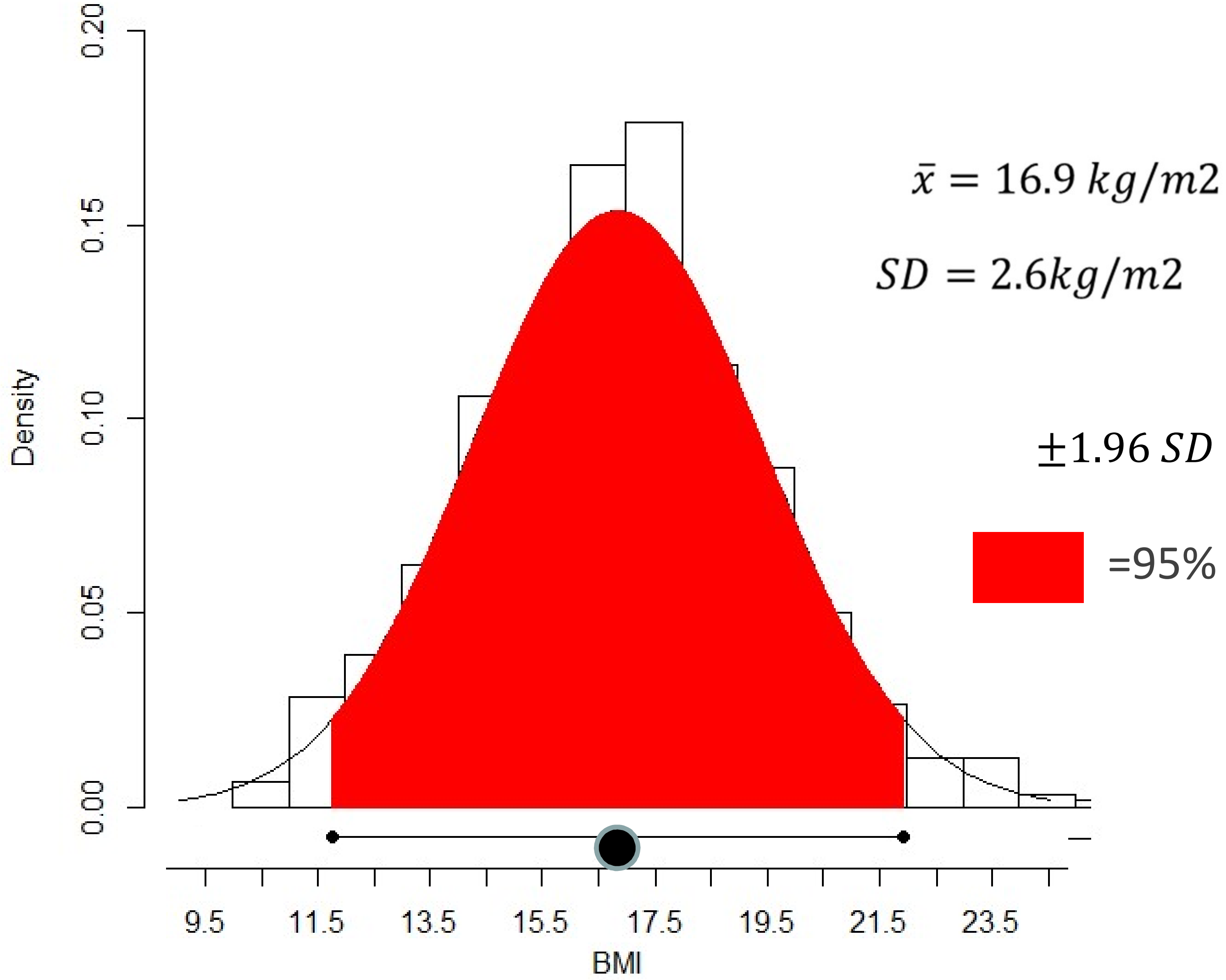
$\bar{x} = 16.9\ kg/m2$

$SD = 2.6 kg/m2$

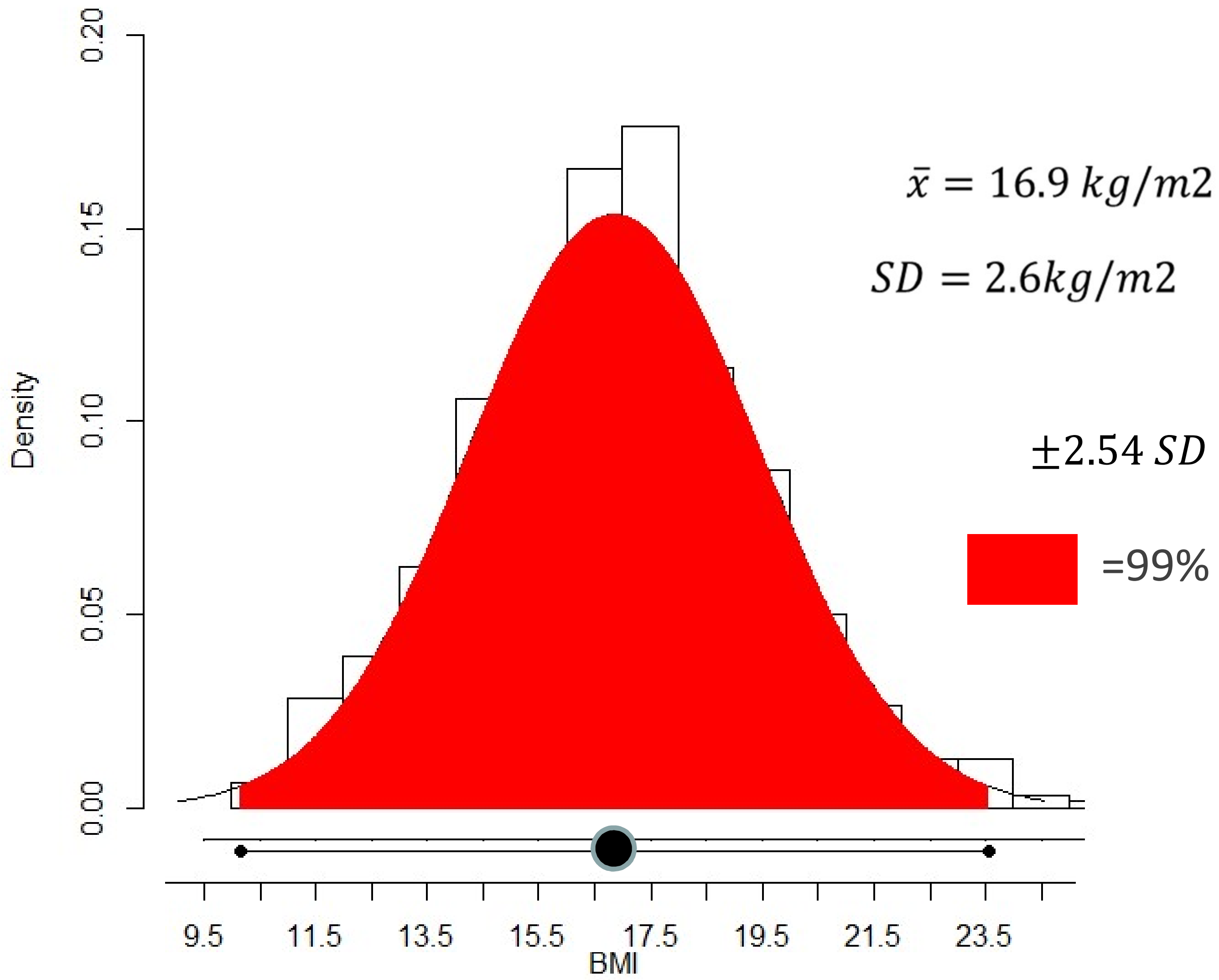
$\pm 1\ SD$

=68%

Histogram with Gaussian approximation – intervals around μ



Histogram with Gaussian approximation – intervals around μ



Omega-3 trial – what do we learn from table 1?

SCHOOL OF MEDICINE AND SURGERY

	n-3 PUFA (n=3494)	Placebo (n=3481)
Patients' characteristics		
Age (years)	67 (11)	67 (11)
Age >70 years	1465 (41·9%)	1482 (42·6%)
Women	777 (22·2%)	739 (21·2%)
Heart disease risk factors		
BMI (kg/m²)	27 (5)	27 (5)
SBP (mm Hg)	126 (18)	126 (18)
DBP (mm Hg)	77 (10)	77 (10)
Heart rate (beats per min)	72 (13)	73 (14)
Current smoking	502 (14·4%)	485 (13·9%)
History of hypertension	1886 (54·0%)	1923 (55·2%)
NYHA class		
II	2226 (63·7%)	2199 (63·2%)
III	1178 (33·7%)	1187 (34·1%)
IV	90 (2·6%)	95 (2·7%)
LVEF (%)	33·0% (8·5)	33·2% (8·5)
LVEF >40%	333 (9·5%)	320 (9·2%)
Medical history		
Admission for HF in previous year	1746 (50·0%)	1638 (47·1%)
Previous AMI	1461 (41·8%)	1448 (41·6%)

Which is the range of BMI in this sample?

4.7 Variance, range, and interquartile range
4.8 Standard deviation

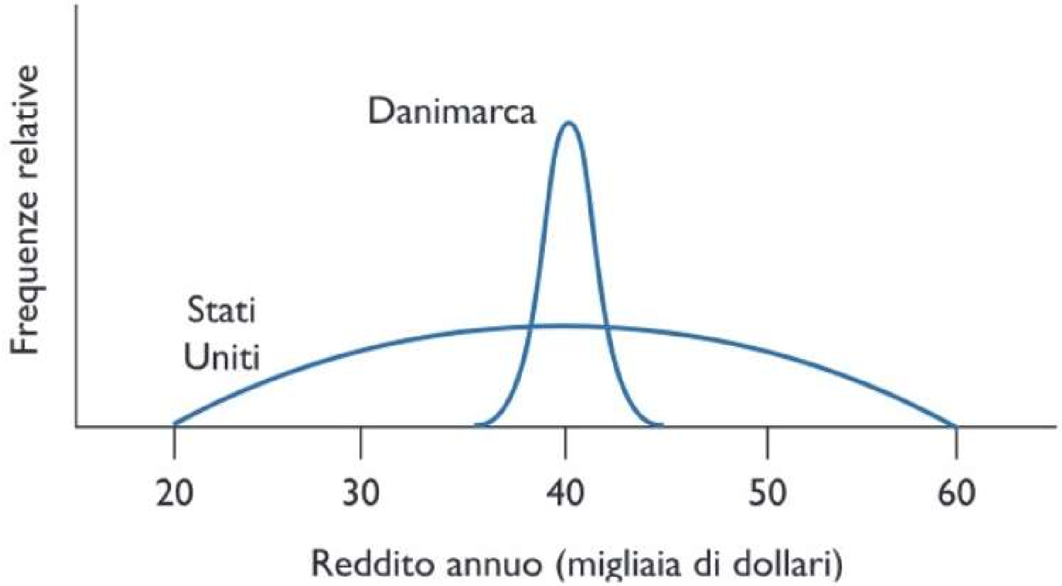


Figura 2.10 Distribuzioni dei redditi per gli insegnanti di musica in Danimarca e negli Stati Uniti. Le distribuzioni hanno la stessa media. La distribuzione relativa agli Stati Uniti è però molto più variabile intorno alla media.

From Agresti, Franklin - Statistica: l'arte e la scienza di imparare dai dati

Coefficient of variation:

Can we use the standard deviation to compare the variability of SBP and DBP?

They are expressed in the same unit of measurement but have different orders of magnitude!

We can use the coefficient of variation (CV) that is the standard deviation divided by the mean:

$$CV = \left(\frac{s}{\bar{x}} \right) \times 100\%$$

It is a pure number that can take positive or negative values depending on the sign of the average.

- the measurement unit is deleted
- the variability is standardized for the order of magnitude of the phenomenon

The CV is not affected by multiplicative changes in scale

It is a useful way of comparing the dispersion of variables measured on different scales

Exercise:

Compare variability between BMI and blood pressure (SBP and DBP) in the GISSI-prevention trial in the n-3PUFA arm

	n-3 PUFA (n=3494)	Placebo (n=3481)
Patients' characteristics		
Age (years)	67 (11)	67 (11)
Age >70 years	1465 (41·9%)	1482 (42·6%)
Women	777 (22·2%)	739 (21·2%)
Heart disease risk factors		
BMI (kg/m²)	27 (5)	27 (5)
SBP (mm Hg)	126 (18)	126 (18)
DBP (mm Hg)	77 (10)	77 (10)
Heart rate (beats per min)	72 (13)	73 (14)
Current smoking	502 (14·4%)	485 (13·9%)
History of hypertension	1886 (54·0%)	1923 (55·2%)
NYHA class		
II	2226 (63·7%)	2199 (63·2%)
III	1178 (33·7%)	1187 (34·1%)
IV	90 (2·6%)	95 (2·7%)
LVEF (%)	33·0% (8·5)	33·2% (8·5)
LVEF >40%	333 (9·5%)	320 (9·2%)
Medical history		

$$CV(BMI) = \frac{5}{27} = 0.1852$$

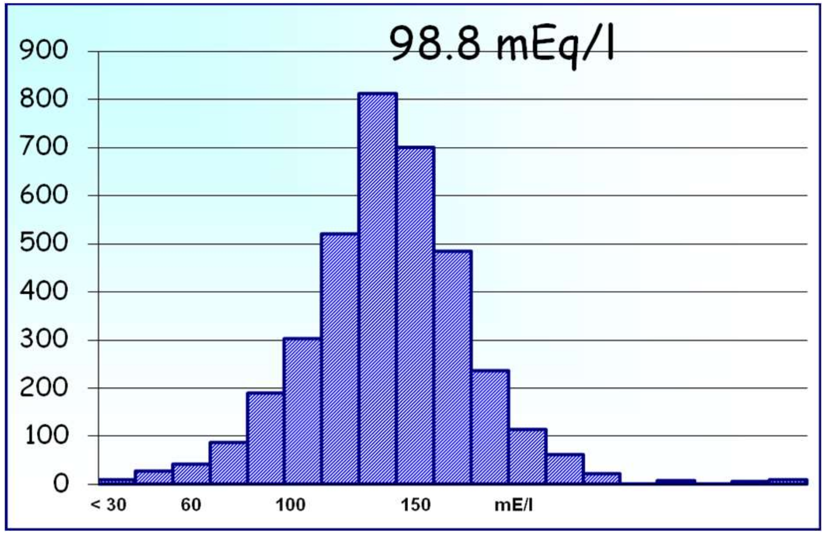
$$CV(SBP) = \frac{18}{126} = 0.1429$$

$$CV(DBP) = \frac{10}{77} = 0.1299$$

Examples:

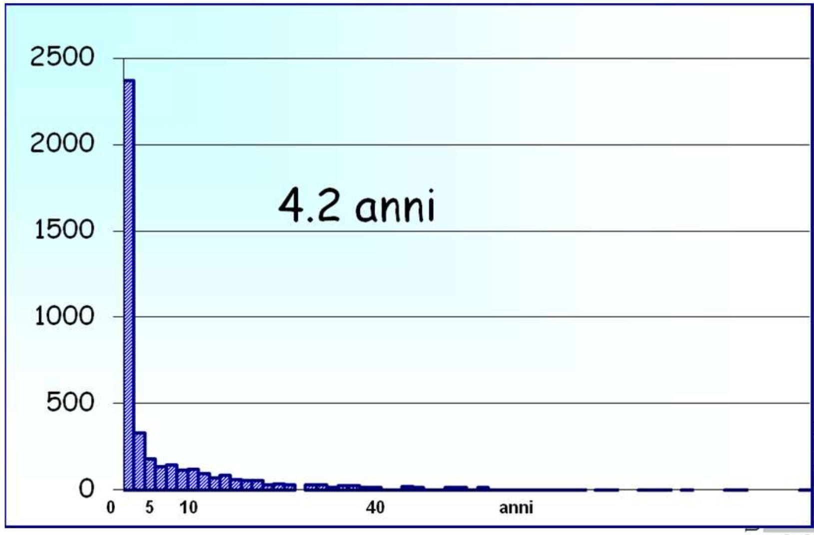
Chlorine concentration in sweat (symmetrical)

mean=98.8 mEq/l



Age at disgnosis in cystic fibrosis (positive asimmetry)

mean=4.2 years



Summary indicators: mode

Modal value - modality(ies) with maximum frequency

Some indexes of **variability** for categorical variables have been proposed (e.g. in Agresti 1990) but none has become a "standard" in practice

Note – For a binary (yes/no) variable the % of yes (or no) summarizes the **variability** (maximum when the proportion is near 50%, minimum when the proportion is near 0% or 100%)

Agresti, Alan (1990). Categorical Data Analysis. John Wiley and Sons, Inc. 24-25