ARISTOTLE UNIVERSITY
OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE
MSc Health Statistics and Data Analytics

Quantitative and qualitative data and their summary measures

NAME Kokkali Stamatia
Research Fellow- General Practitioner

THESSALONIKI 2021-22

Learning Objectives

Upon completion of this lecture you will be able to:

Define data and variables

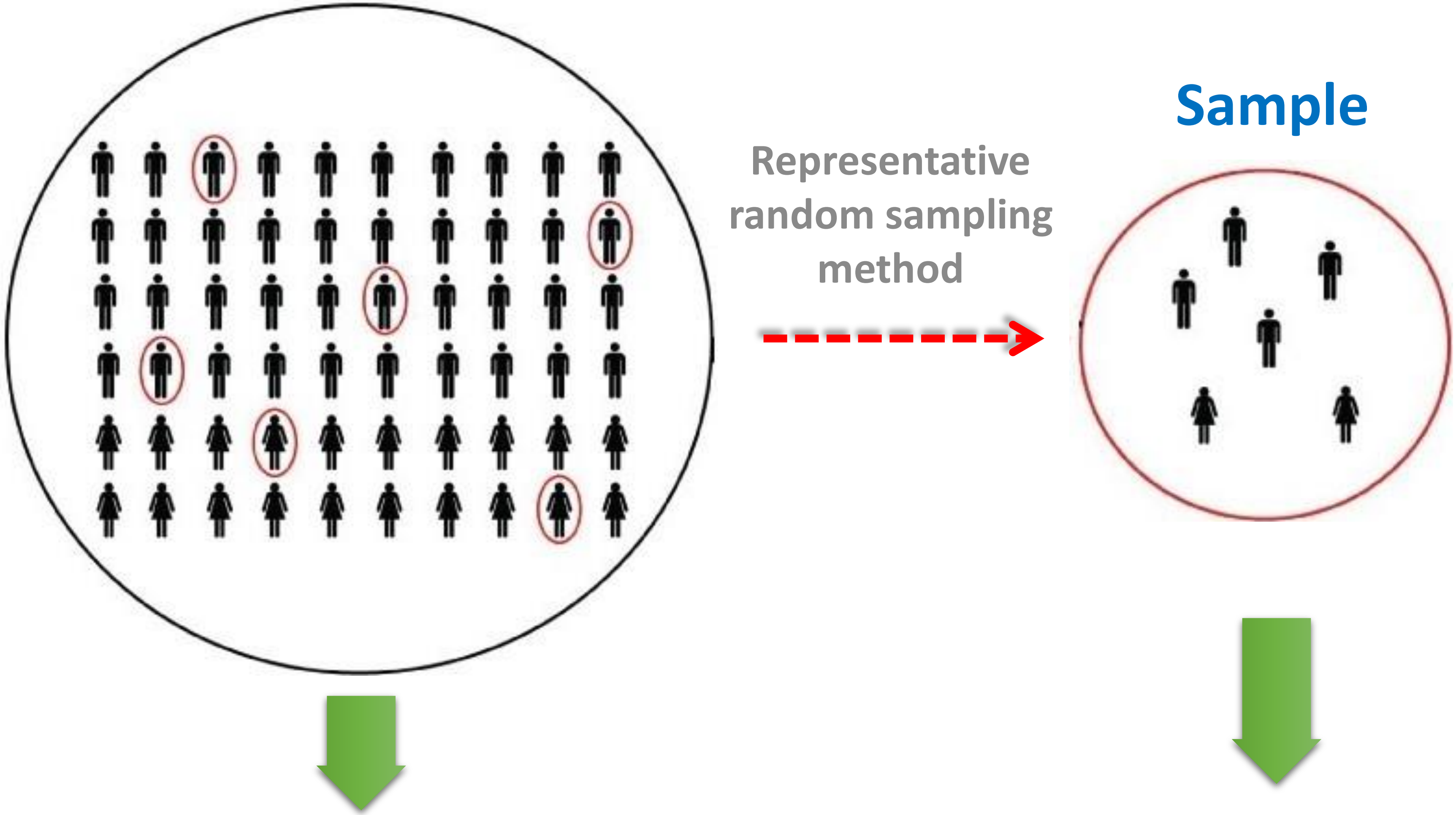
Define outcome and explanatory variables

Distinguish between qualitative and quantitative variables

Compute the measures of central location and variation

Present and summarize appropriately your data

Interpret the summary measures



Population

Representative random sampling method

Sample

Population parameters

μ, σ^2, σ

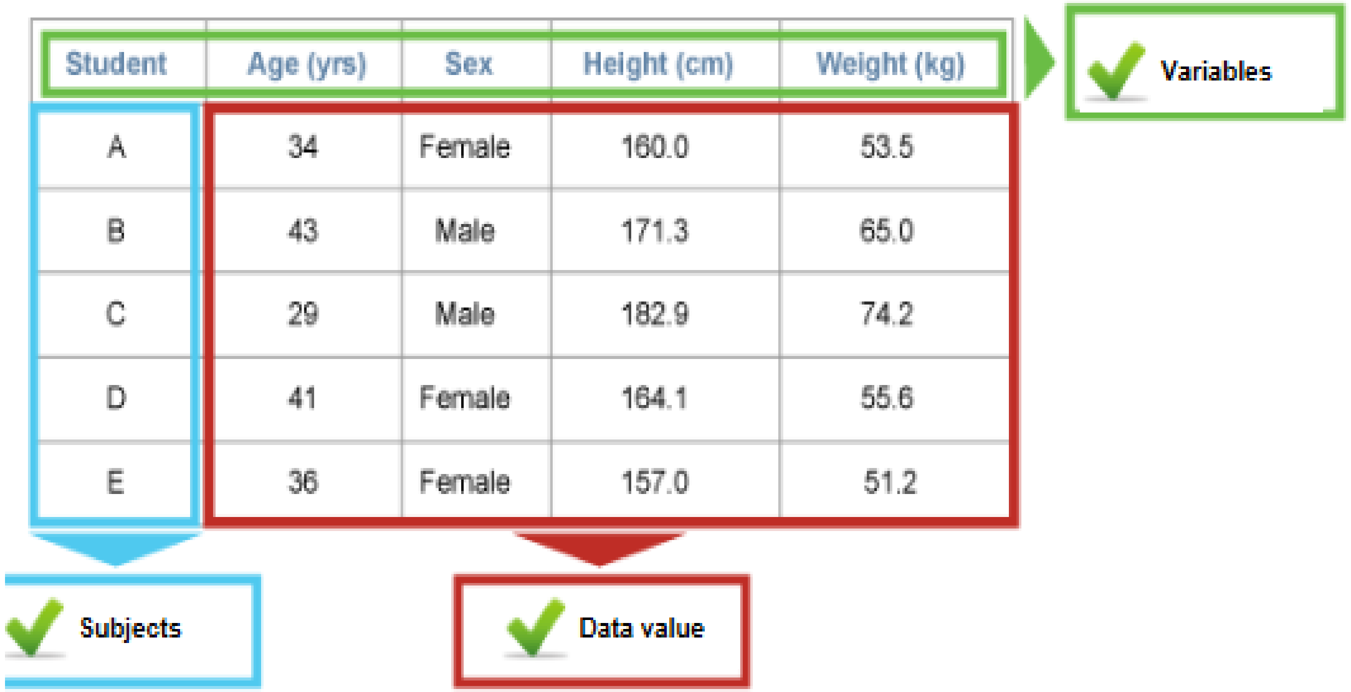
Sample statistics

$\bar{x}, s^2, s$

Data and variables

- Data are collected on the specific characteristics of each subject, and groups are formed and compared on the basis of these characteristics.
- In statistical terms these characteristics are called variables since they vary from subject to subject.
- Suppose we wanted to study a group of medical students. We might ask about their:
  - Age
  - Sex
  - Place of residence
  - Weight
  - Height
- Each of these characteristics varies from student to student. They are what we call variables, and the values we collect from the students are called data.

Example of a dataset



Student	Age (yrs)	Sex	Height (cm)	Weight (kg)
A	34	Female	160.0	53.5
B	43	Male	171.3	65.0
C	29	Male	182.9	74.2
D	41	Female	164.1	55.6
E	36	Female	157.0	51.2

✓ Variables

✓ Subjects

✓ Data value

Outcome variable

An **outcome** variable is a characteristic which we believe to be affected by the values taken by other variables. It is also called a **response** or **dependent** variable

Examples of Outcome Variables:

In a study of the effect of cigarette smoking on the incidence of <u>lung cancer</u>

In a study of the link between an outbreak of <u>salmonella poisoning</u> and a particular supplier of frozen chicken

Explanatory variable

- An **explanatory** variable may influence the outcome. Such a variable partly explains the variability of the outcome

- **Independent** or **predictor** variables

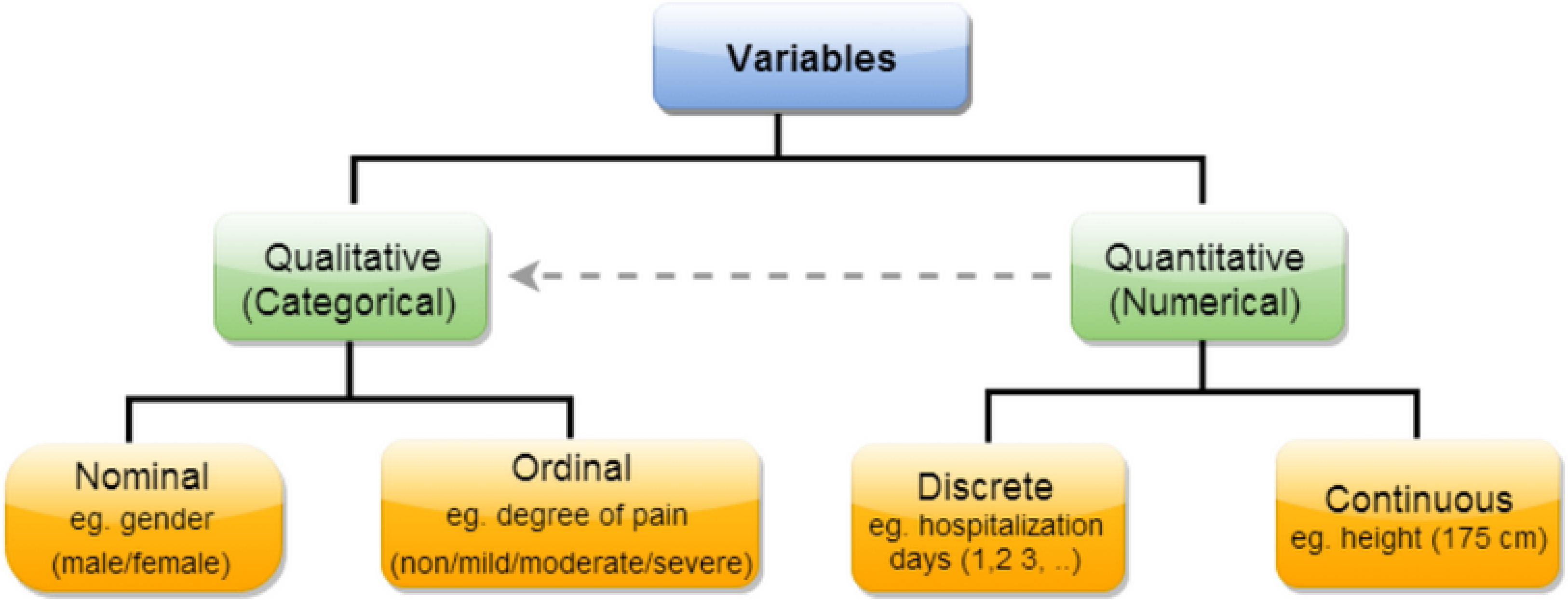
- <u>For example</u>:

  - In a randomised controlled trial for a new drug for the treatment of hypertension:

    - the outcome is blood pressure or a change in blood pressure

    - the explanatory variable is the treatment, since the main factor to influence blood pressure is whether a subject is assigned to the new drug or control drug.

Types of variables



Variables

Qualitative (Categorical)

Quantitative (Numerical)

Nominal eg. gender (male/female)

Ordinal eg. degree of pain (non/mild/moderate/severe)

Discrete eg. hospitalization days (1,2 3, ..)

Continuous eg. height (175 cm)

Representation and summarization of qualitative variables

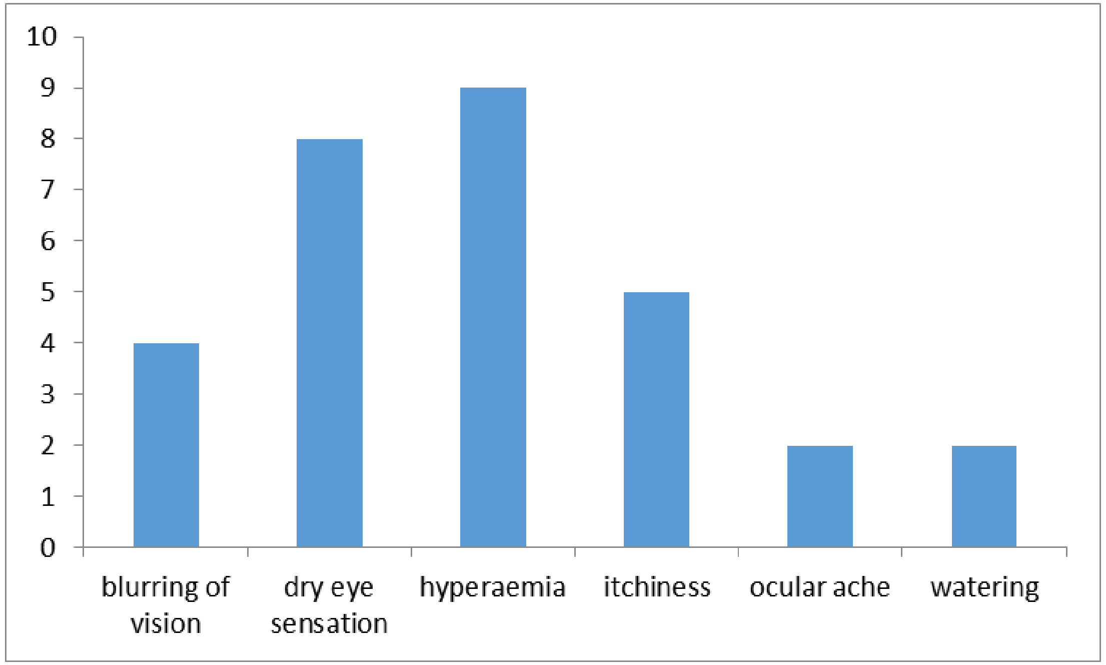
One-way table

$$\text{Relative frequency (\%)} = \frac{\text{frequency in category x 100\%}}{\text{total frequency}}$$

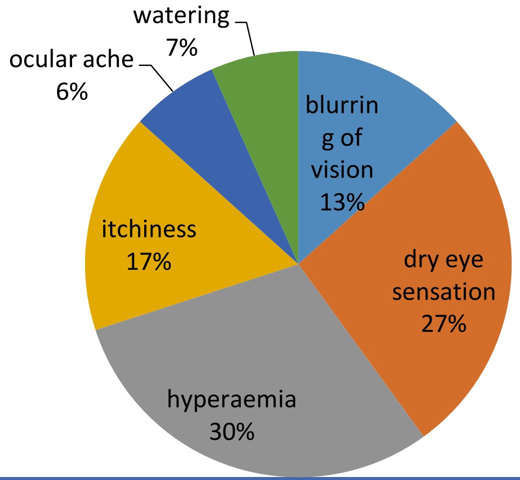
Adverse event	Number of Cases	Relative Frequency (%)
blurring of vision	4	13.3
dry eye sensation	8	26.7
hyperaemia	9	30.0
itchiness	5	16.7
ocular ache	2	6.7
watering	2	6.7
Total	**30**	**100.0**

"Thirty percent of the patients experienced hyperaemia, while 26.7% experienced dry eye sensation."

Bar chart



Pie chart

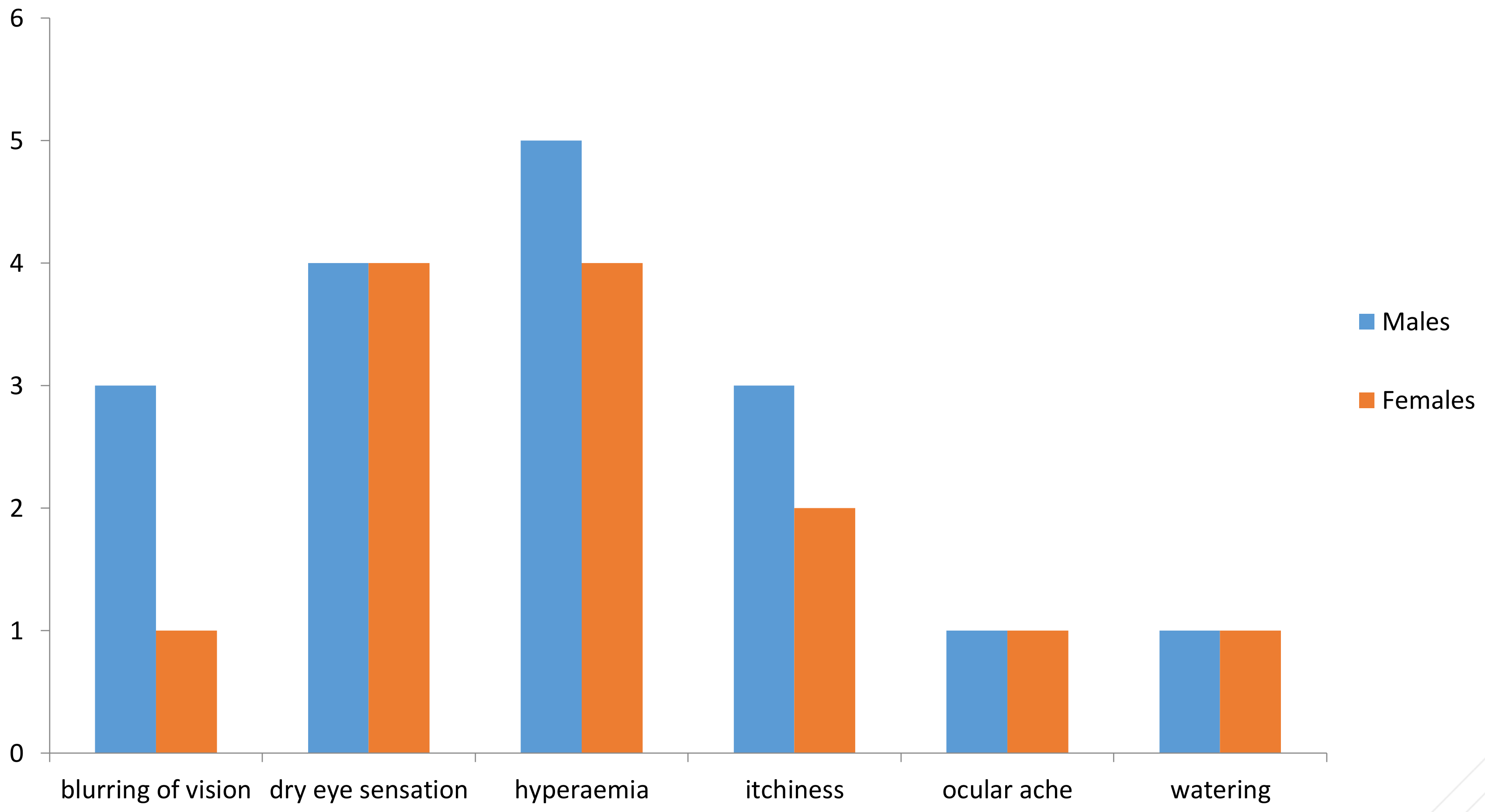


Representation and summarization of qualitative variables

- Two-way table

	Males		Females	
Adverse event	**Number of Cases**	**Relative Frequency (%)**	**Number of Cases**	**Relative Frequency (%)**
blurring of vision	3	17.6	1	7.7
dry eye sensation	4	13.3	4	30.8
hyperaemia	5	29.4	4	30.8
itchiness	3	17.6	2	15.4
ocular ache	1	5.9	1	7.7
watering	1	5.9	**1**	**7.7**
Total	**17**	**100**	**13**	**100**

Clustered bar chart



Summarizing and Describing Quantitative Variables

- Measures of central tendency
  - Mean
  - Median (50th percentile)
  - Mode

- Measures of variation
  - Range
  - Interquartile range
  - Standard Deviation (Variance)

Sample Mean: The Average or Arithmetic Mean

- Add up data, then divide by sample size (*n*)
- The sample size *n* is the number of observations (pieces of data)

- Example: Five weights (kilograms) (n=5)
  53.5, 65.0, 74.2, 55.6, 51.2

Can be represented with math type notation:

x_1= 53.5, x_2=65.0,…..x_5=51.2

The sample mean is easily computed by adding up the five values and dividing by 5: in stat notation the sample mean is frequently represented by a letter with a line over it: for example (pronounced 'x bar')

Mean, example

- Five weights (kg) (n=5)

53.5, 65.0, 74.2, 55.6, 51.2

$$\bar{x} = \frac{53.5 + 65.0 + 74.2 + 55.6 + 51.2}{5} = 59.9 \text{ kg}$$

Notes on sample mean

- Generic Formula Representation

Where $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots + x_n$

In the formula to find the mean, we use the "summation sign" ∑: This is just mathematical shorthand for "add up all of the observations"

- Also called *sample average* or *arithmetic mean*
- Sensitive to extreme values (in smaller samples)
    - One data point could make a great change in sample mean

title

- The median is the middle number (also called the *50th percentile or Q2). But the data have to be ordered first either ascending or descending*

51.2 53.5 55.6 65.0 74.2

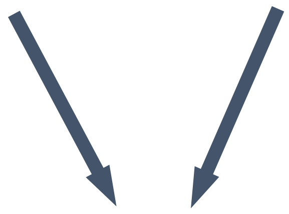
- The median is not sensitive to extreme values
  - For example, if 74.2 became 174.2, the median would remain the same, but the *mean would change from 59.9 kg to 79.9 kg*

51.2 53.5 55.6 65.0 174.2

Median

- If the sample size is an even number

51.2, 53.5, 55.6, 61.4, 65.0, 74.2



Median

$$\frac{55.6+61.4}{2}=58.5\,kg$$

Mode

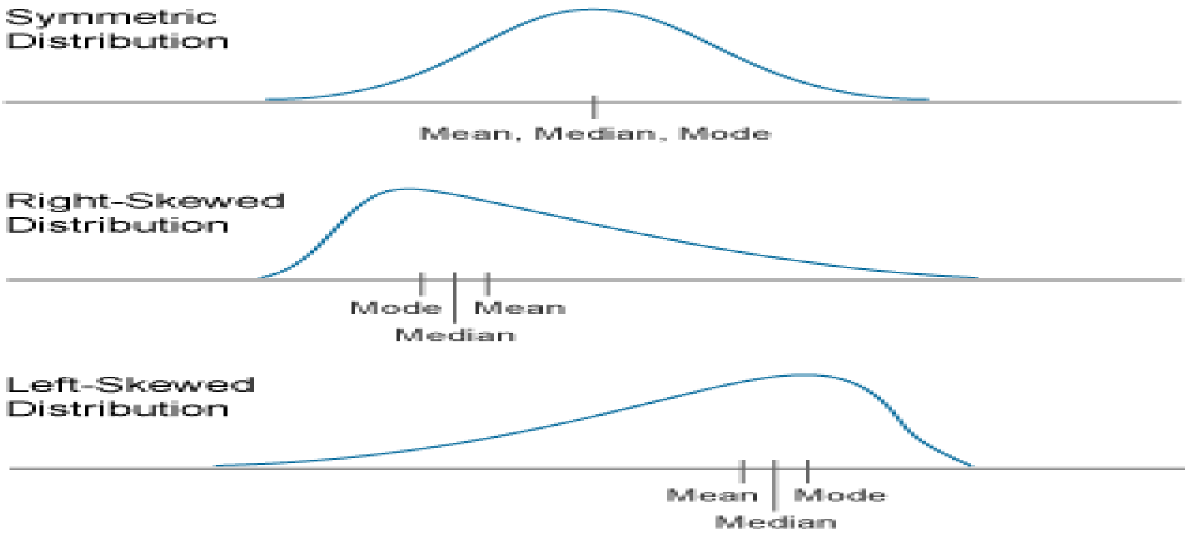
- The mode is the most frequently occurring score in a set of scores.
- 82 82 83 83 84 85 86 **87 87 87** 88 90 95 99 99
- The mode of the above set of numbers is 87 because it appears three times — more than any other number in the set.
- 82 83 84 86 87 **88 88** 89 90 **91 91** 92 94 97 98
- There are two modes above. The numbers 88 and 91 both appear twice. This is a bi-modal (two modes) data set.
- 82 83 84 86 87 88 89 90 91 92 93 94 95 96 97
- There is no mode for this distribution. No score occurs more frequently than any other. The mode is the most frequent score in a set of data.

Properties of the Mean, Median & Mode

- The mean is sensitive to outliers; the others are not.

- The mode may be affected by small changes in the data; the others are not.

- All three measures of location are equal for a symmetric distribution; in a *skewed* distribution they differ (see below).



Measures of variability

Range

Interquartile range (IQR)

Variance (s^2)

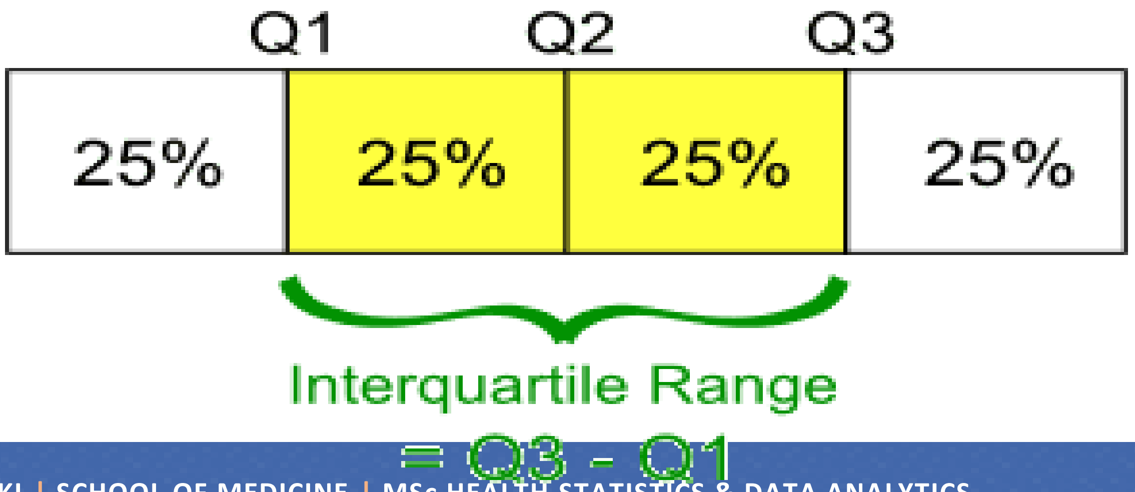
Standard deviation (*s* or *SD*)

Range

- **Range = maximum – minimum**
- Extremely prone to distortion even with few present outliers.
- More useful as an alarm: extreme values indicate potential data import error.
- Weight in kilograms of 11 students:
- 54.6, 56.4, 57.2, 64.6, 67.7, 68.1, 68.7, 69.4**,** 72.3, 74.8, 76.2
- Range : 76.2 - 54.6 = 21.6 kg.

Interquartile range (IQR)

- **IQR = Q3 – Q1**
- What is the quartile? It is one of the four divisions of the values of a ranked variable which are grouped into four equal (in size) parts.



Q1 Q2 Q3

25% 25% 25% 25%

Interquartile Range

= Q3 - Q1

IQR computations

- 54.6, 56.4, 57.2, 64.6, 67.7, **68.1**, 68.7, 69.4, 72.3, 74.8, 76.2

- **Q2** = median = 68.1
- Q2 divides the data into two sets:

A lower one including 54.6, 56.4, **57.2**, 64.6, 67.7

with median 57.2 = **Q1** (25%) and

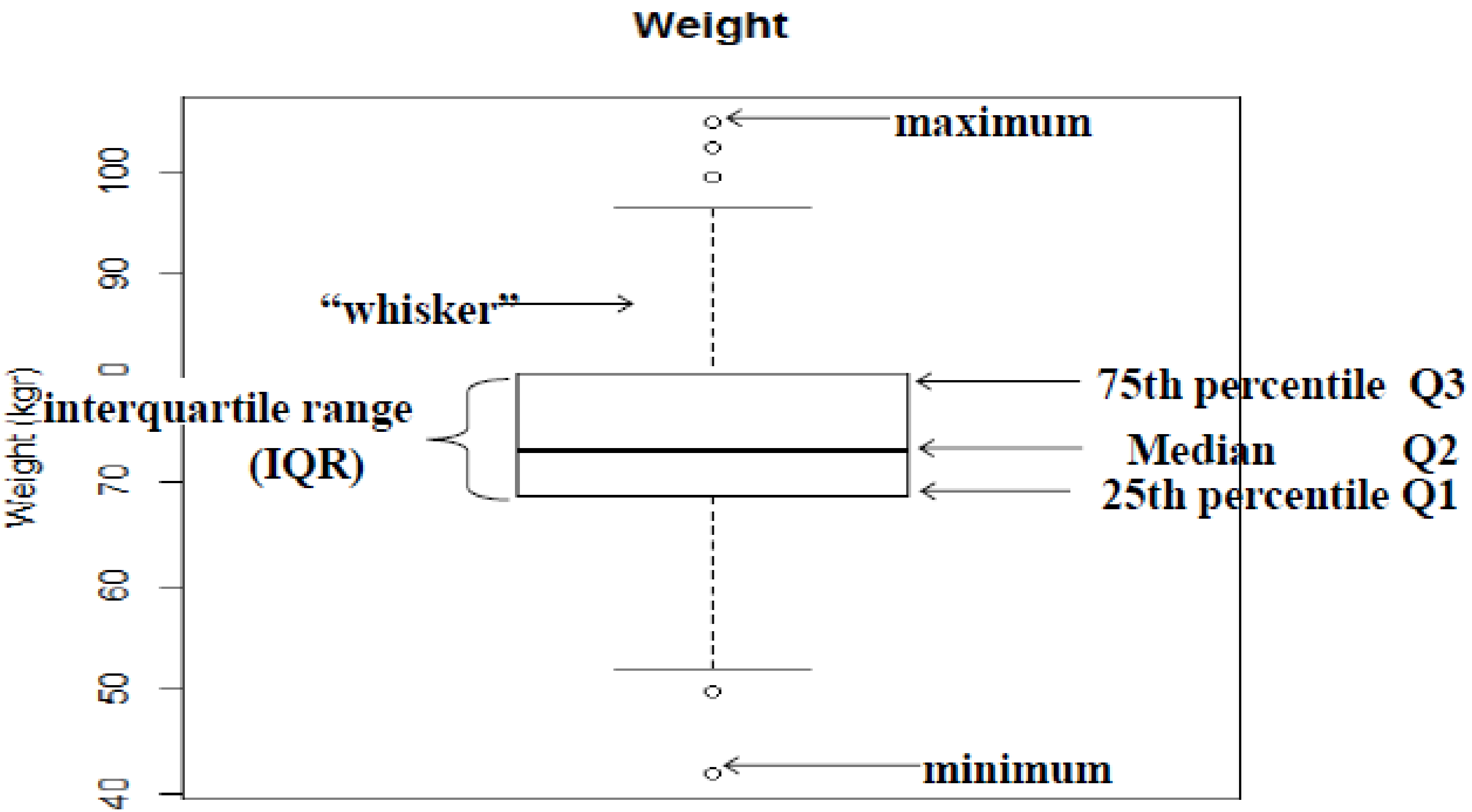
a higher one including 68.7, 69.4, **72.3**, 74.8, 76.2
with median 72.3 = **Q3** (75%)

- **IQR = Q3 - Q1** = 72.3 – 57.2 = 15.1 (50% of the data)

54.6, 56.4, **57.2**, 64.6, 67.7, **68.1**, 68.7, 69.4, **72.3**, 74.8, 76.2

Q1 **Q2** **Q3**

Boxplot


Weight
Weight (kgr)
100
90
80
70
60
50
40
maximum
"whisker"
interquartile range
(IQR)
75th percentile Q3
Median Q2
25th percentile Q1
minimum

Variance and standard deviation

- *The variance is the average of the square of the deviations about the sample mean*

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

- The standard deviation is the square root of *s*^{2}

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

Example : Standard Deviation, on Small Weight Data

- Recall, the 5 weights (kg) with sample mean of 59.9 kg
- Five weights (kg) (n=5)

53.5, 65.0, 74.2, 55.6, 51.2

$$\sum_{i=1}^{5}(x_i-\overline{x})^2=(53.5-59.9)^2+(65-59.9)^2+(74.2-59.9)^2$$

$$+(55.6-59.9)^2+(51.2-59.9)^2$$

$$\sum_{i=1}^{5}(x_i-\overline{x})^2=(-6.4)^2+(5.1)^2+(14.3)^2+(-4.3)^2+(-8.7)^2$$

$$=(40.96)+(26.01)+(204.49)+(18.49)+(75.69)$$

$$=365.4\ \ kg^2$$

Example : Standard Deviation, on Small Weight Data

- Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{365.64}{4} = 91.41\ \mathrm{kg}^2$$

- Standard deviation (*s*)

$$\sqrt{\mathrm{s}^2} = \sqrt{91.41\ \mathrm{kg}^2}$$

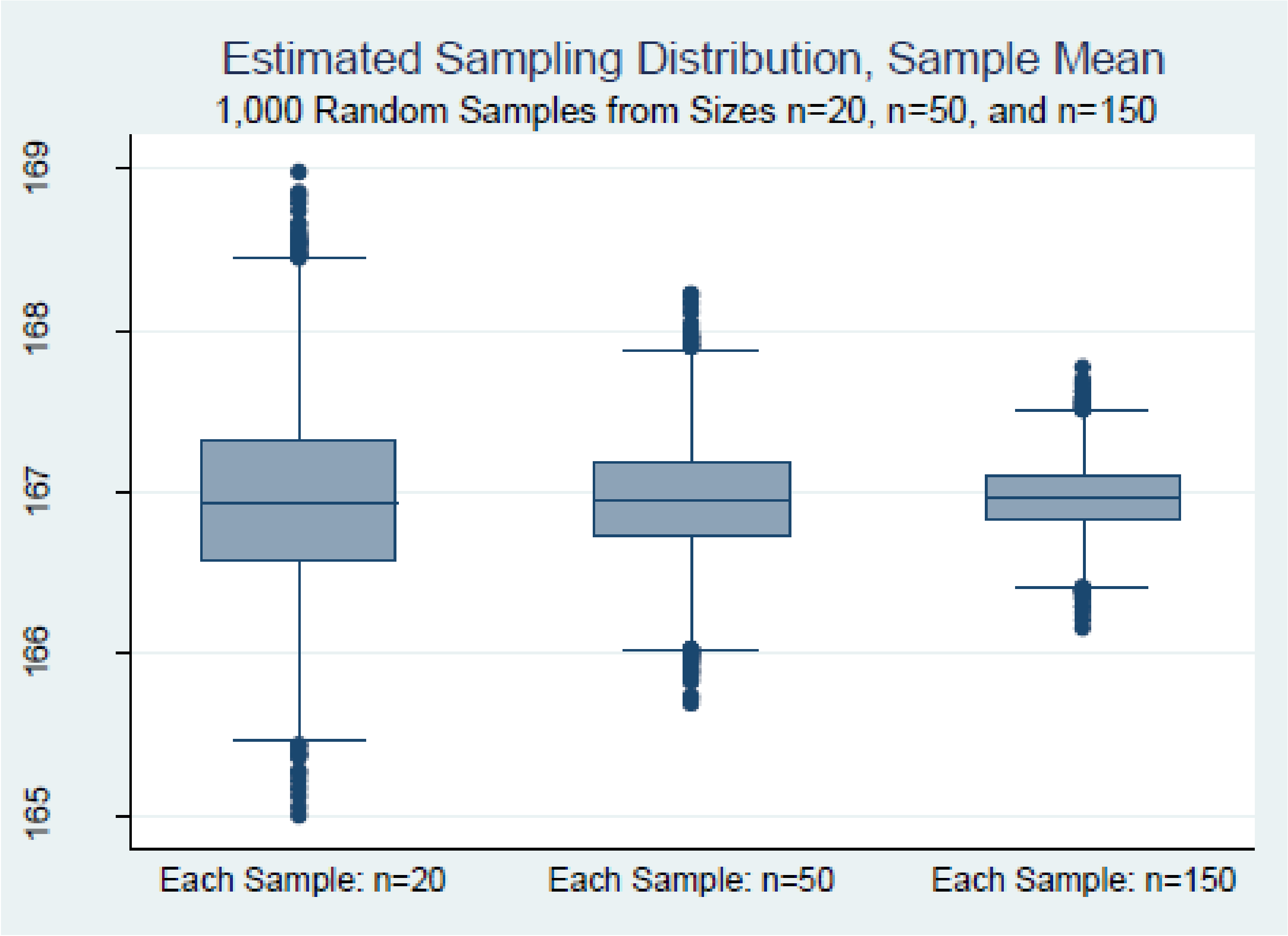
***s* = 9.56 ≈ 10 (kg)**

Which measure of variability

- **SD**
  - Symmetrical data
  - better mathematical properties
  - includes all the information contained in the data
  - Sensitive to extreme values
- **IQR**
  - skewed distributions (those with outliers)
  - not sensitive to extreme values
- **Range**
  - rarely used since it tells us nothing of the dispersion of observations between the maximum and minimum values.
  - Sensitive to extreme values

With increased sample size

- What do you notice? Mean Height in cm



Exclusion of an outlier

- Endorphin levels (pmol/l) was recorded for 11 marathon runners
- 66 72 79 84 102 110 123 144 162 169 414
- Mean 138.6
- Median 110
- SD 97.97
- Range 348
- IQR 83

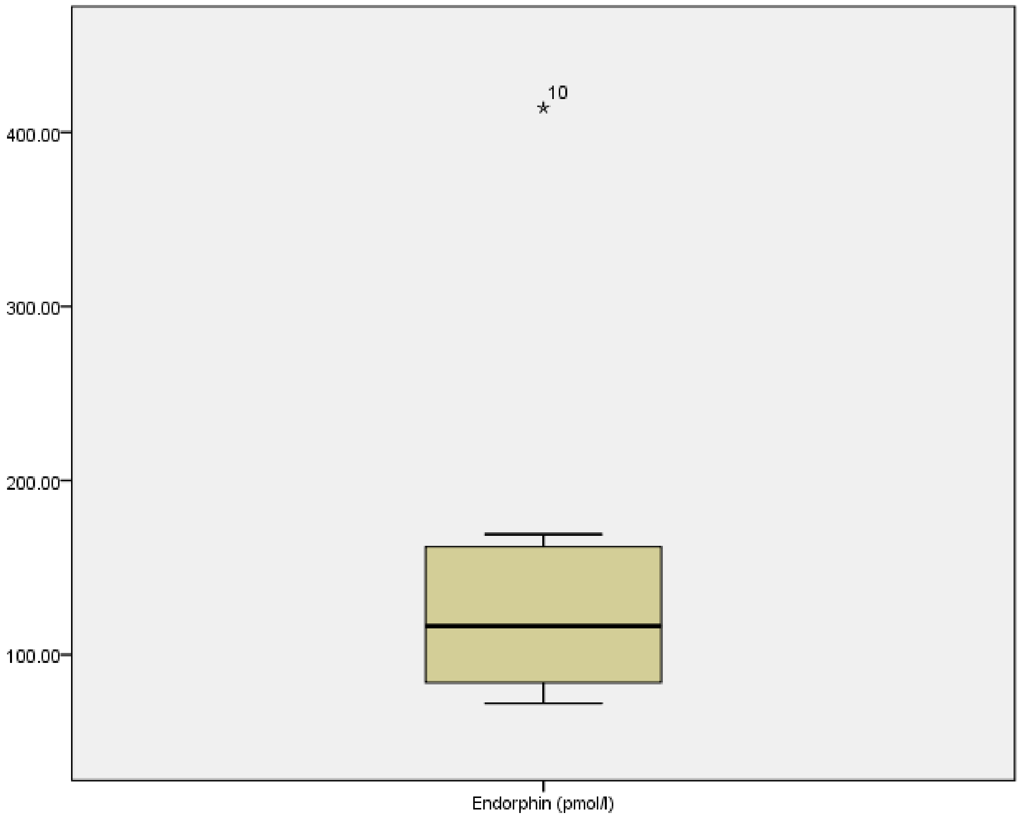
→ 111.1 pmol/l

→ 106 pmol/l

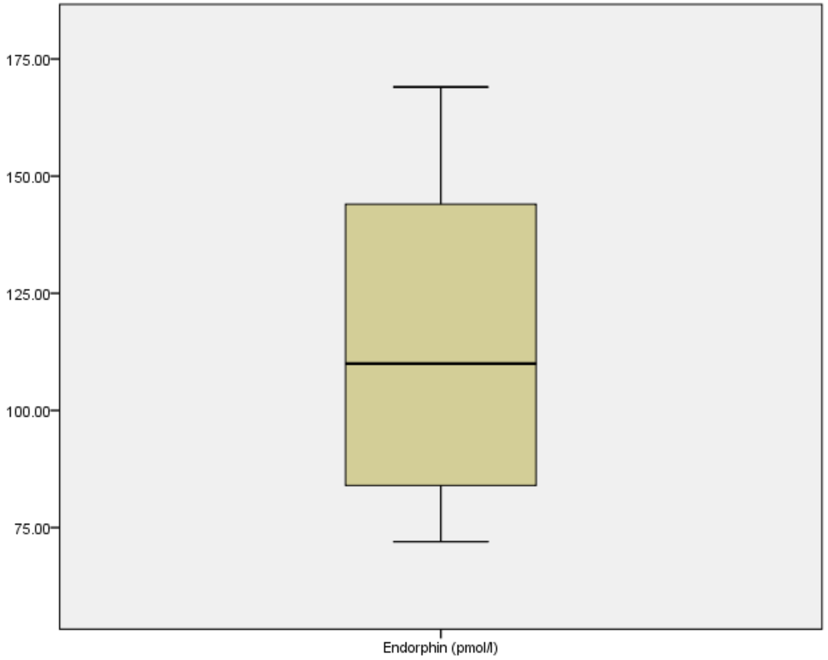
→ 37.39 pmol/l → 103pmol/l

→ 71.25 pmol/l

Boxplot



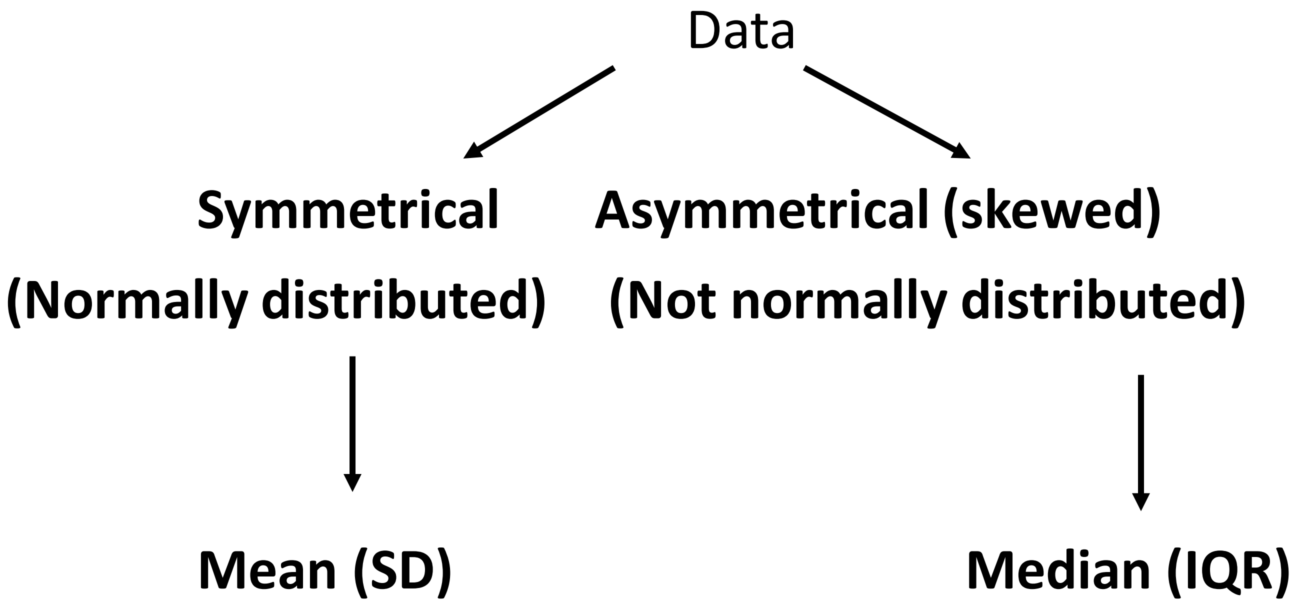
10
400.00
300.00
200.00
100.00
Endorphin (pmol/l)



175.00
150.00
125.00
100.00
75.00
Endorphin (pmol/l)

Take home message

Summary measures



Data

Symmetrical

(Normally distributed)

Asymmetrical (skewed)

(Not normally distributed)

Mean (SD)

Median (IQR)

Article table

A.-B. Haidich et al. / Journal of Clinical Epidemiology 66 (2013) 371–378

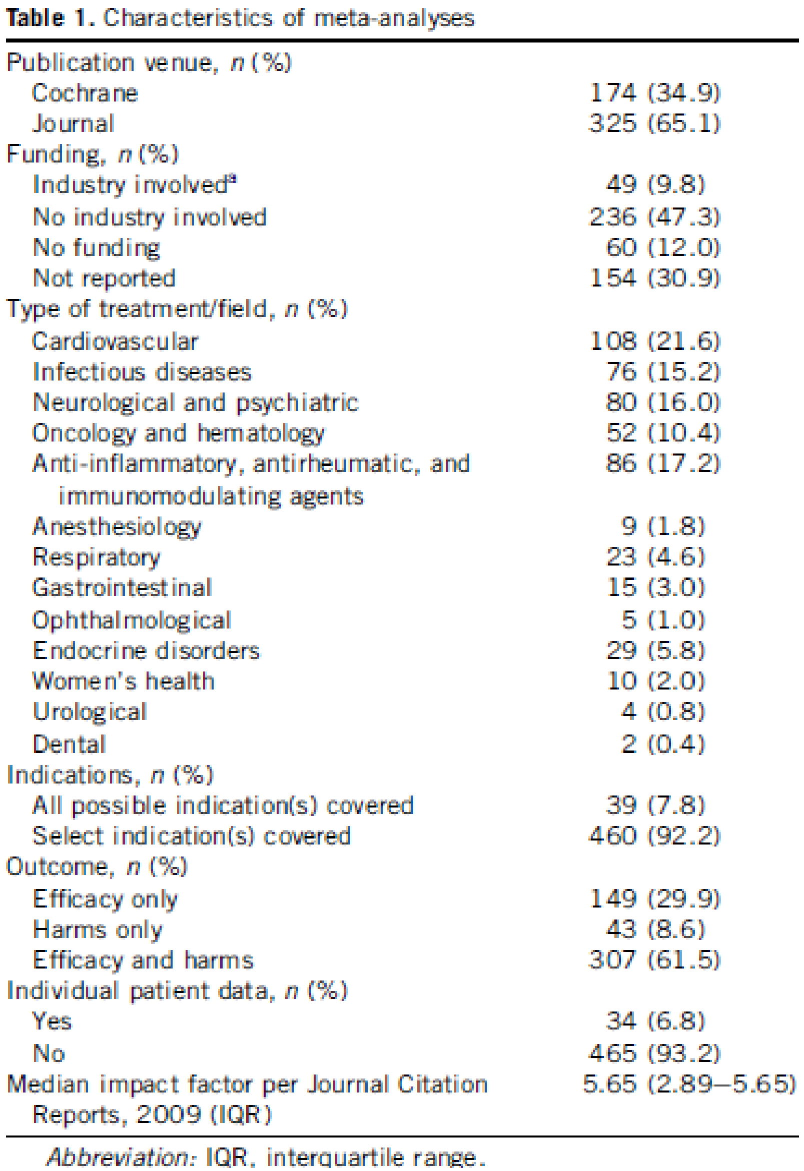
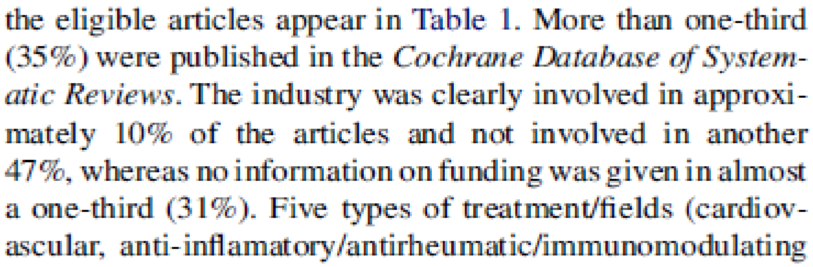


Table 1. Characteristics of meta-analyses

Publication venue, *n* (%)	
Cochrane	174 (34.9)
Journal	325 (65.1)
Funding, *n* (%)	
Industry involved[a]	49 (9.8)
No industry involved	236 (47.3)
No funding	60 (12.0)
Not reported	154 (30.9)
Type of treatment/field, *n* (%)	
Cardiovascular	108 (21.6)
Infectious diseases	76 (15.2)
Neurological and psychiatric	80 (16.0)
Oncology and hematology	52 (10.4)
Anti-inflammatory, antirheumatic, and immunomodulating agents	86 (17.2)
Anesthesiology	9 (1.8)
Respiratory	23 (4.6)
Gastrointestinal	15 (3.0)
Ophthalmological	5 (1.0)
Endocrine disorders	29 (5.8)
Women's health	10 (2.0)
Urological	4 (0.8)
Dental	2 (0.4)
Indications, *n* (%)	
All possible indication(s) covered	39 (7.8)
Select indication(s) covered	460 (92.2)
Outcome, *n* (%)	
Efficacy only	149 (29.9)
Harms only	43 (8.6)
Efficacy and harms	307 (61.5)
Individual patient data, *n* (%)	
Yes	34 (6.8)
No	465 (93.2)
Median impact factor per Journal Citation Reports, 2009 (IQR)	5.65 (2.89–5.65)

Abbreviation: IQR, interquartile range.



the eligible articles appear in Table 1. More than one-third (35%) were published in the *Cochrane Database of Systematic Reviews*. The industry was clearly involved in approximately 10% of the articles and not involved in another 47%, whereas no information on funding was given in almost a one-third (31%). Five types of treatment/fields (cardiovascular, anti-inflamatory/antirheumatic/immunomodulating

Internet Addiction in Greek Medical Students: an Online Survey

Zoi Tsimtsiou • Anna-Bettina Haidich • Dimitris Spachos • Stamatia Kokkali •
Panagiotis Bamidis • Theodoros Dardavesis • Malamatenia Arvanitidou

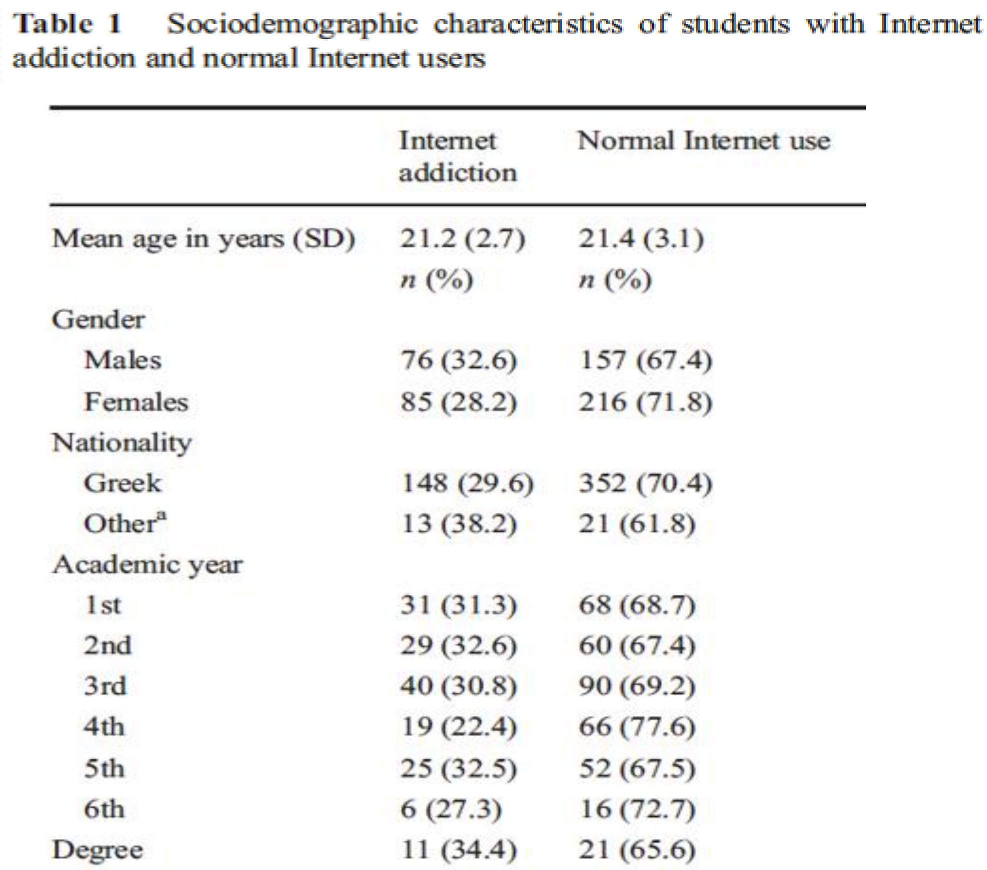


Table 1 Sociodemographic characteristics of students with Internet addiction and normal Internet users

	Internet addiction	Normal Internet use
Mean age in years (SD)	21.2 (2.7)	21.4 (3.1)
	n (%)	*n* (%)
Gender		
Males	76 (32.6)	157 (67.4)
Females	85 (28.2)	216 (71.8)
Nationality		
Greek	148 (29.6)	352 (70.4)
Other[a]	13 (38.2)	21 (61.8)
Academic year		
1st	31 (31.3)	68 (68.7)
2nd	29 (32.6)	60 (67.4)
3rd	40 (30.8)	90 (69.2)
4th	19 (22.4)	66 (77.6)
5th	25 (32.5)	52 (67.5)
6th	6 (27.3)	16 (72.7)
Degree	11 (34.4)	21 (65.6)

Ερωτήσεις