Types of Data

Descriptive Statistics

Table 3: Topics for the first FRCR medical statistics module

Topic	Further guidance
Types of data	▪ Present and summarise individual variables ▪ Recognise categorical data (nominal, ordinal) ▪ Recognise discrete and continuous numerical data ▪ Recognise symmetric and skewed distribution ▪ Describe the normal distribution ▪ Interpret bar charts and histograms ▪ Define and apply measures of central tendency and spread

What is Statistics?

Statistics is the science of

collecting data

summarizing data

presenting data

and interpreting data

Using data to estimate the magnitude of associations

and test hypotheses

What is Statistics?

"**Statistics**…is the core science of evidence-based practice"

Martin Bland


Population
Random Sample
Select random sample
*Use sample to make an **inference** about the population*

Types of Analysis

- Descriptive (epidemiology, pure statistics)
  - “What happened?”
  - Focus: Past/Historical
  - Action: Reporting and Visualisation

- Predictive (Statistical inference)
  - “What will happen?”
  - Focus: Future/Probabilities
  - Action: Forecasting and Risk

- Causal (Causal Inference)
  - “Why did it happen?”
  - Focus: Cause and Effect
  - Action: Intervention and Strategy

Frequentist vs Bayesian

Feature	Frequentist Approach	Bayesian Approach
Probability is...	A fixed, long-term frequency.	A personal degree of certainty.
Parameters	Fixed but unknown (e.g., "The average height *is* X").	Random variables (e.g., "The average height is a *range*").
Testing Goal	To reject the Null Hypothesis.	To update the probability of a hypothesis.
Best For...	Large datasets, regulatory trials, objective standards.	Small datasets, "peeking" at results, incorporating expert knowledge.

Descriptive statistics

Summarising and presenting data.

Essential before any predictive or inferential analysis is conducted

Allows "a feel" for the data

Helps form subjective impressions of answers to research questions.

Dataset

A collection of data is called a "***dataset***"

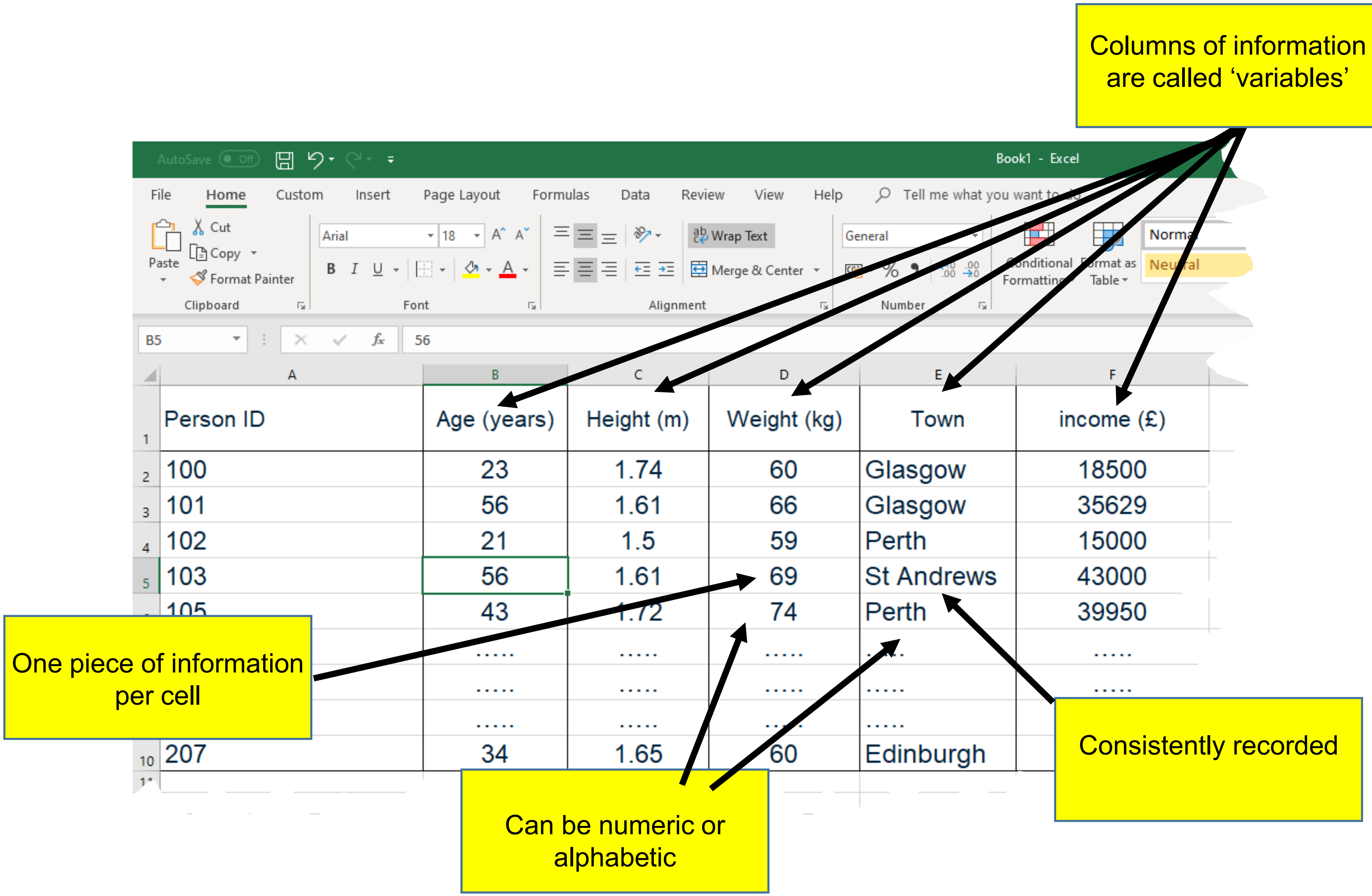
Contains information on ***subjects*** we are interested in:

Individuals: *age, height, weight, town, PS, time since diagnosis*

Hospitals: *number of nurses, number of beds, death rate, LOS*

Countries: *population size, literacy rates, GDP, life expectancy*

For computer analysis the data must have a clear <u>structure</u>

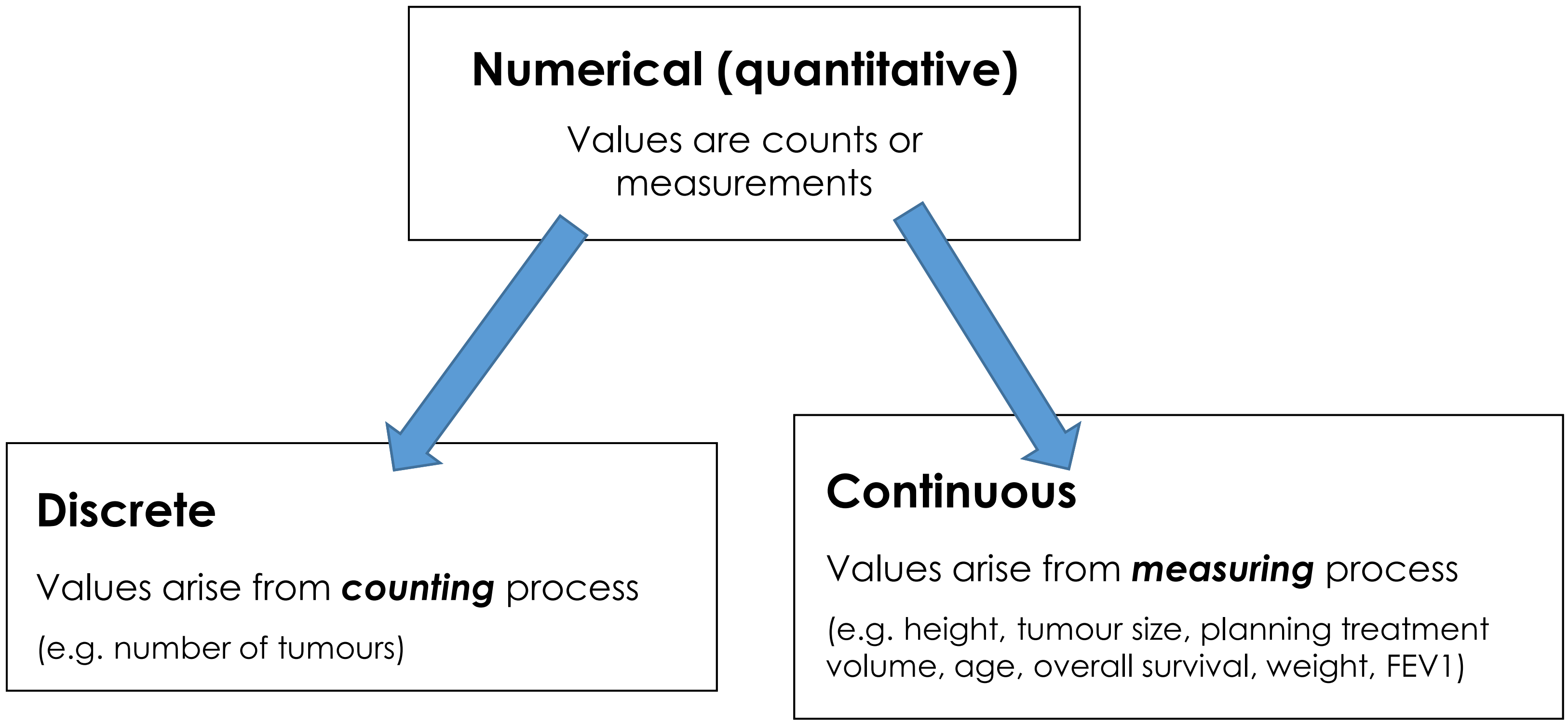


Person ID	Age (years)	Height (m)	Weight (kg)	Town	income (£)
100	23	1.74	60	Glasgow	18500
101	56	1.61	66	Glasgow	35629
102	21	1.5	59	Perth	15000
103	56	1.61	69	St Andrews	43000
105	43	1.72	74	Perth	39950
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	
207	34	1.65	60	Edinburgh	

Types of Data

To produce descriptive statistics appropriately requires knowledge of different **data types**.

Broadly, data are either ***numerical or categorical***


Numerical (quantitative)
Values are counts or measurements
Discrete
Values arise from ***counting*** process
(e.g. number of tumours)
Continuous
Values arise from ***measuring*** process
(e.g. height, tumour size, planning treatment volume, age, overall survival, weight, FEV1)

Categorical (qualitative)

tells us which category an individual belongs to

Nominal scale

Categories distinguished by a **name** with **no intrinsic ordering**
(e.g. sex, histology, cancer type, postcode)

Ordinal scale

Categories distinguished by name with intrinsic ordering
(e.g performance status, toxicity)


A sample of 6 week old Burmese kittens
Identify 2 continuous variables and 2 categorical variables

Binary variables

A categorical measure with only two categories (for example alive or dead) is called **dichotomous**.

Sometimes the categories of a dichotomous variable are labeled 0 and 1, and called a **binary variable**.

Paired data

Usually, we compare 2 separate groups of individuals.

Sometimes data may consist of paired of observations from the same individuals

Tumour size before or after treatment

PS recorded on same patients by 2 separate clinicians

Pain score recorded on 2 separate occasions

It is important to be able to recognise 'paired' data

Graphics and descriptive statistics

Identify the main features of data

Detect outliers

Identifying data which has been recorded incorrectly

Includes frequency tables, histograms, bar charts, pie charts, scattergraphs, contingency tables, box plots

Frequency Distribution Tables

Frequency count - a count of the number of times something occurs

Within a data set we list the data values and count how many times each value occurs.

A **frequency distribution table** lists data values and the frequency each value occurs

Frequency Distribution Tables

Number of brain metastases	Number of patients (n)	percent of patients (%)
1	19	59.4
2	4	12.5
3 or more	9	28.1
Total	32	100

The ***most*** *frequent* category is called the **mode**.

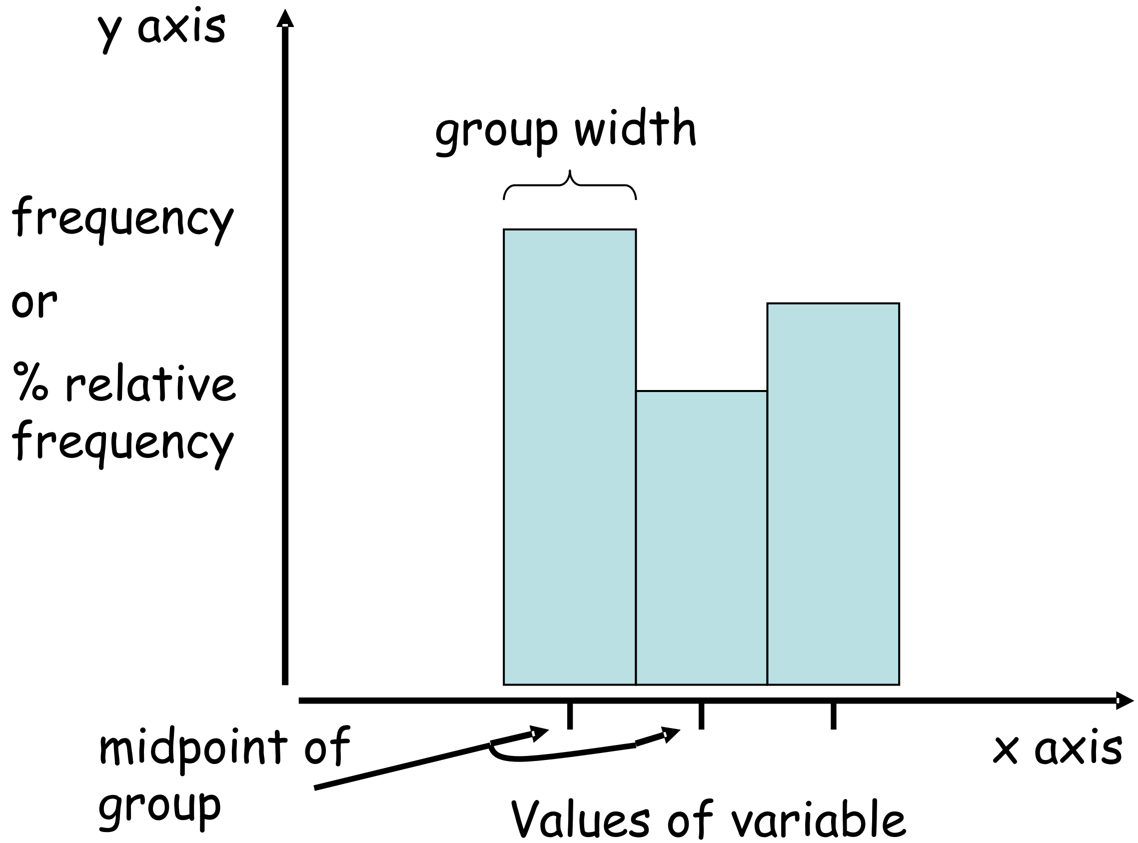
The percentages sometimes expressed **proportions** (0.594, 0.125, 0.281).

Grouped Frequency Distribution Tables

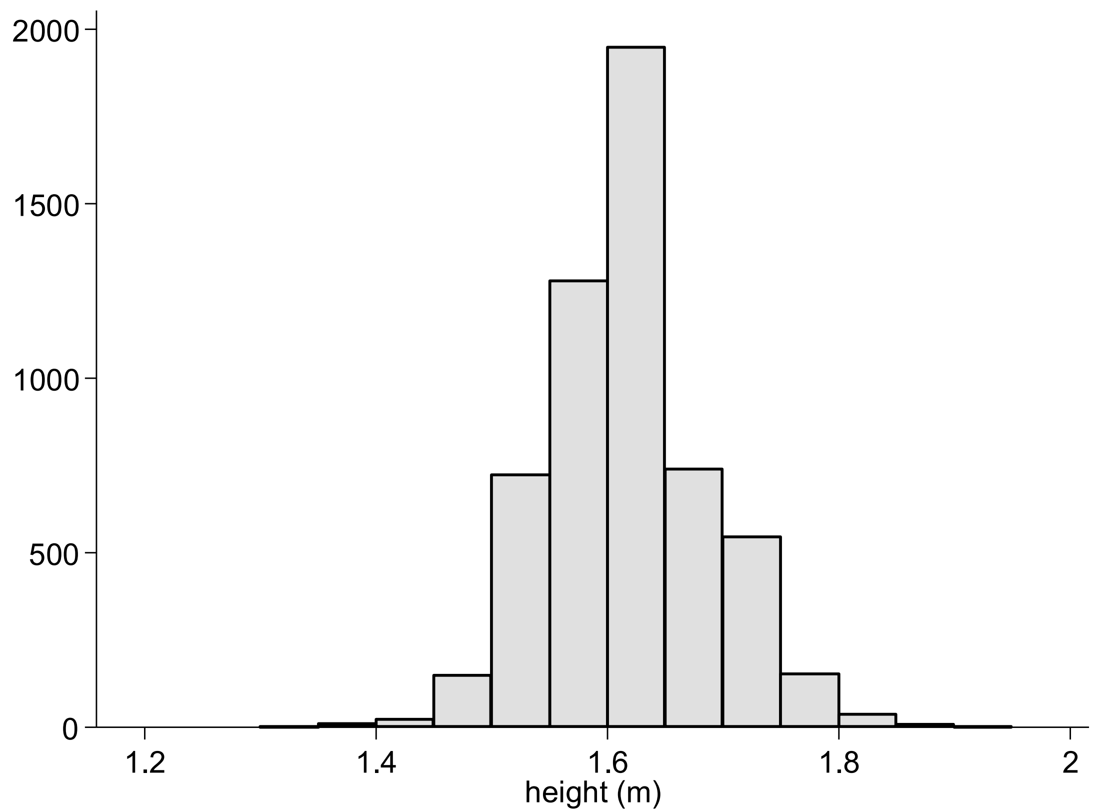
When there are many data values - data are grouped (e.g – age groups 20-29, 30-39, 40-49, 50-59, 60-69, etc).

The frequency distribution table then includes data groupings and the frequency of each group.

Histogram (numeric data)

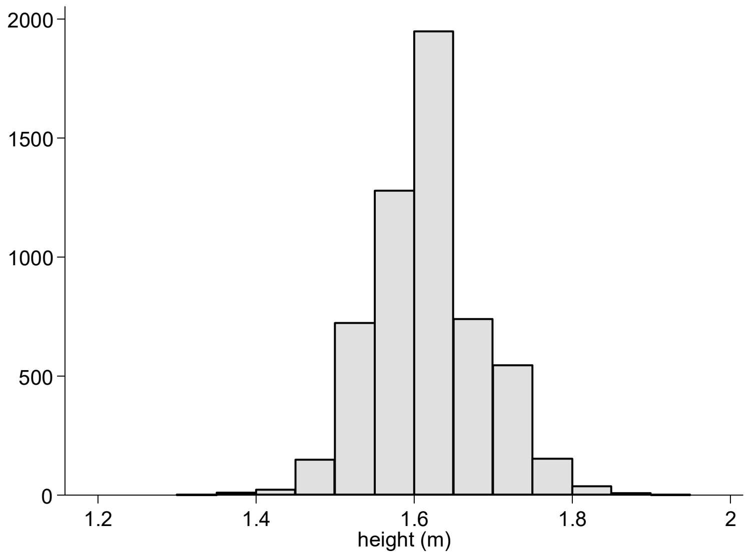


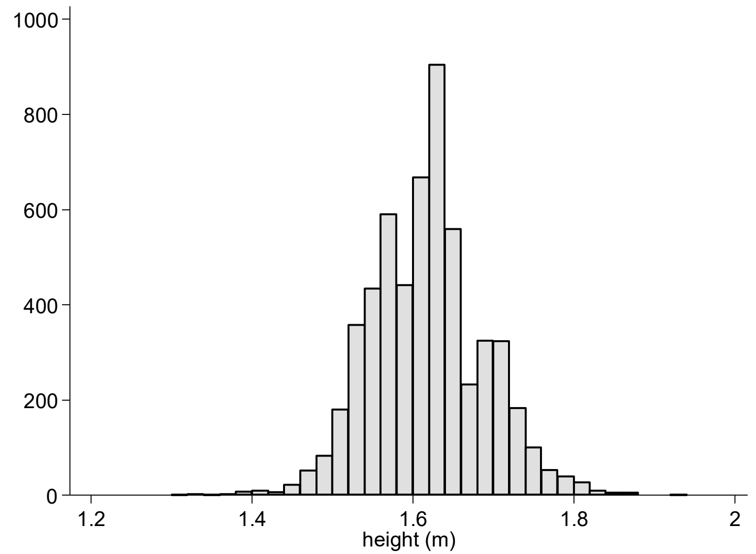
Histogram (numeric data)



Height of 5,628 women aged 25-64

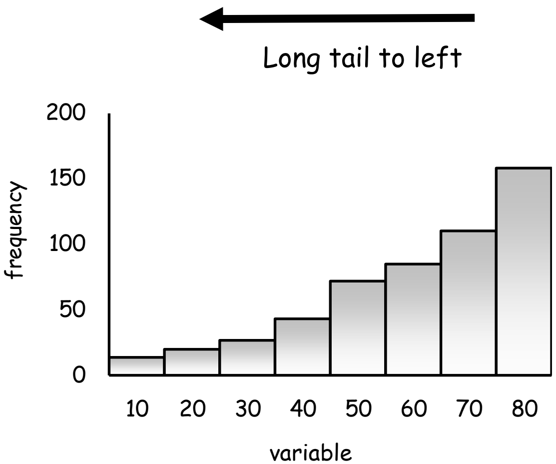
Histogram (numeric data)



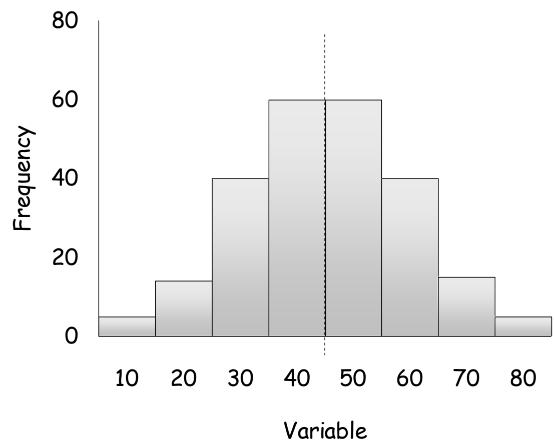


Height of 5,628 women aged 25-64

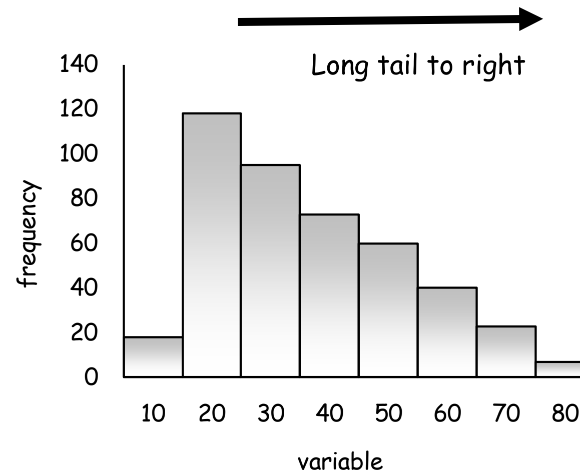
Negatively skewed
(long tail to the left)



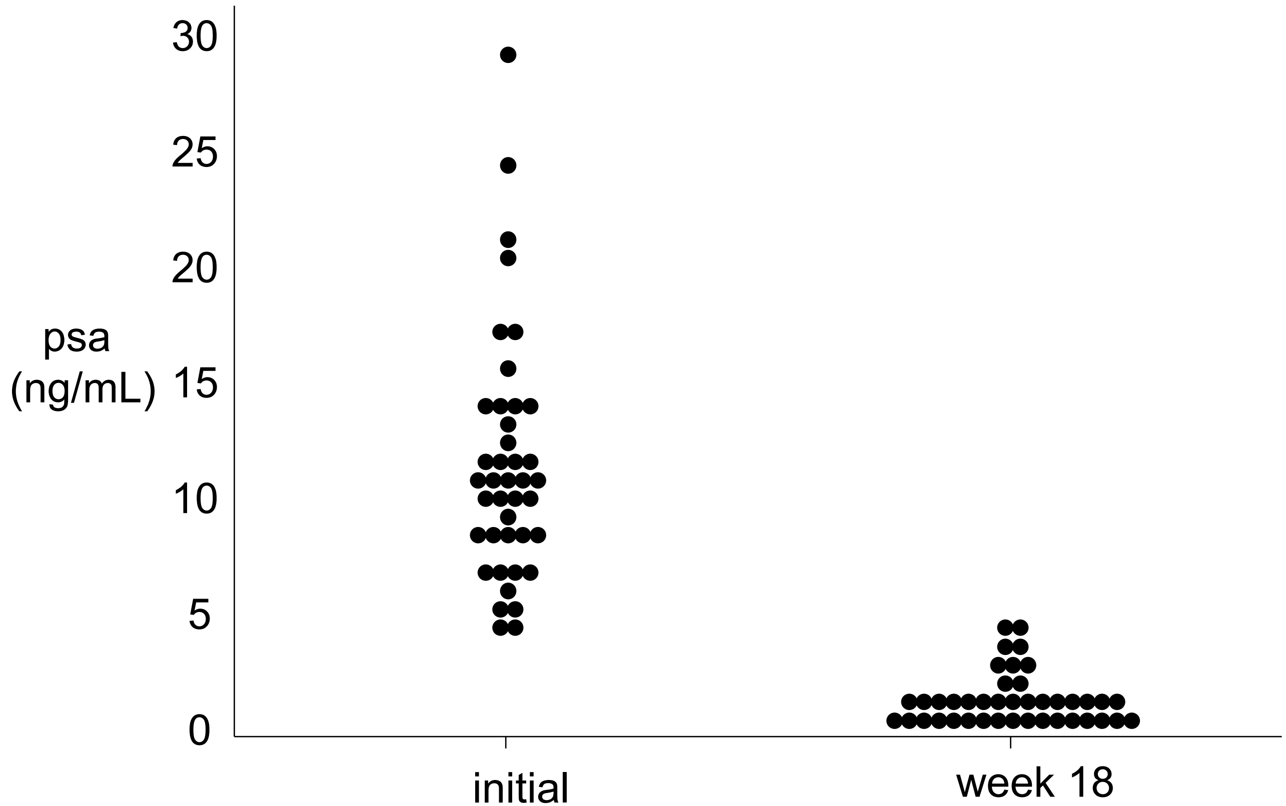
Symmetric (bell shaped) histogram



Positively skewed
(long tail to the right)

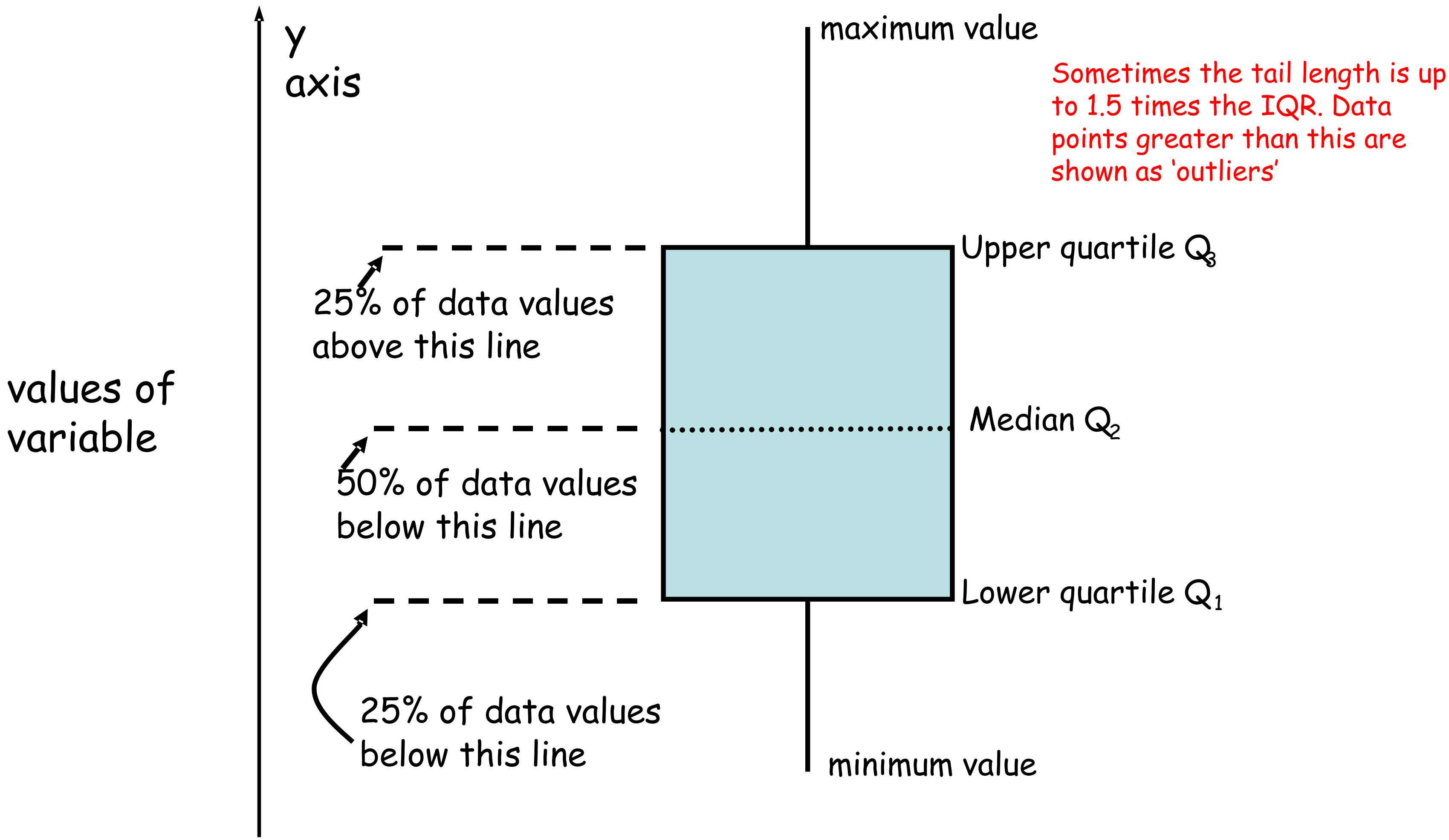


Dotplot (numeric data)

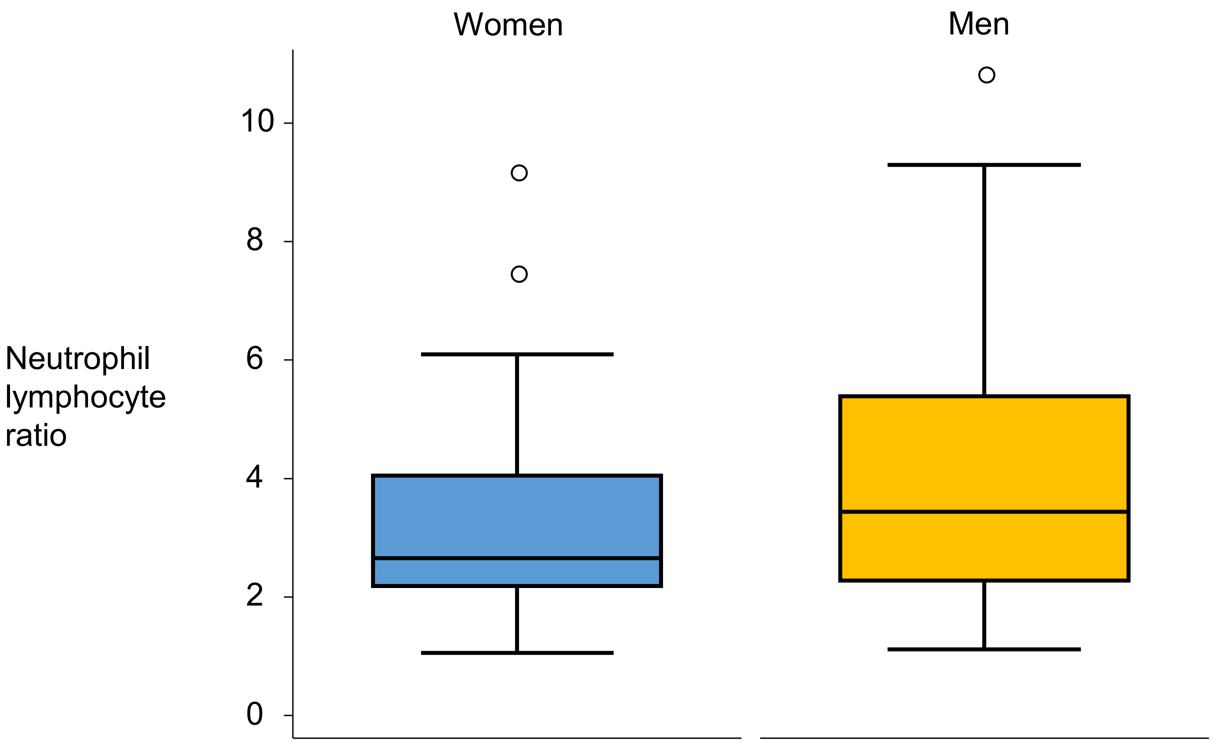


psa levels among prostate cancer patients before and after treatment with SABR

Box plot (numeric data)

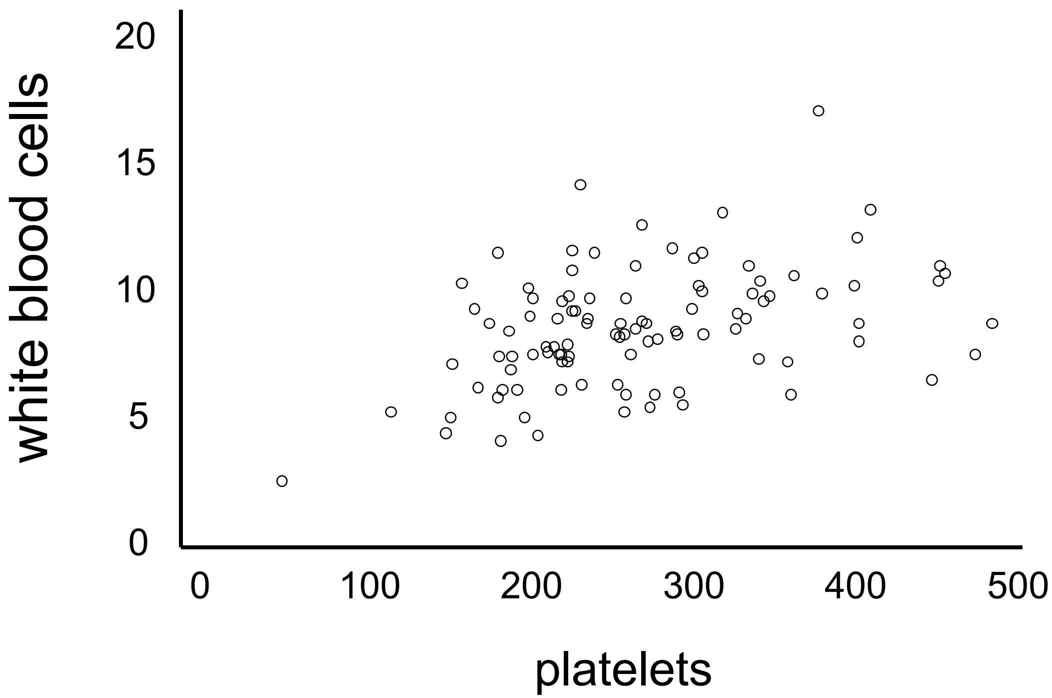


Box plot



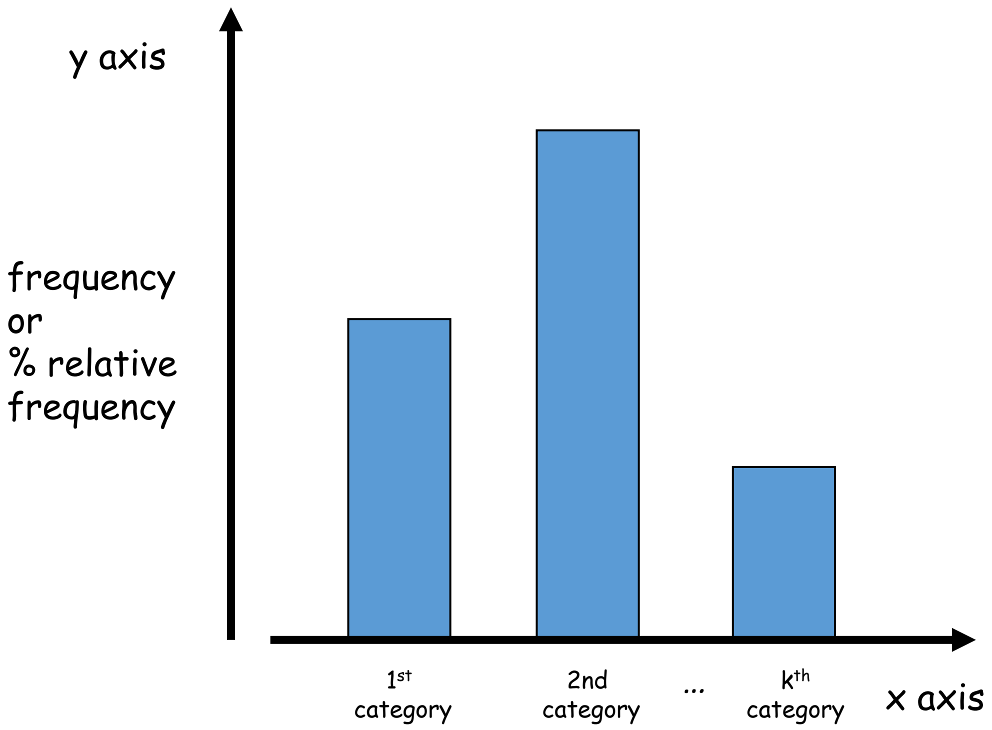
102 patients with NSCLC prior to treatment with SABR

Scattergraph (numeric data)

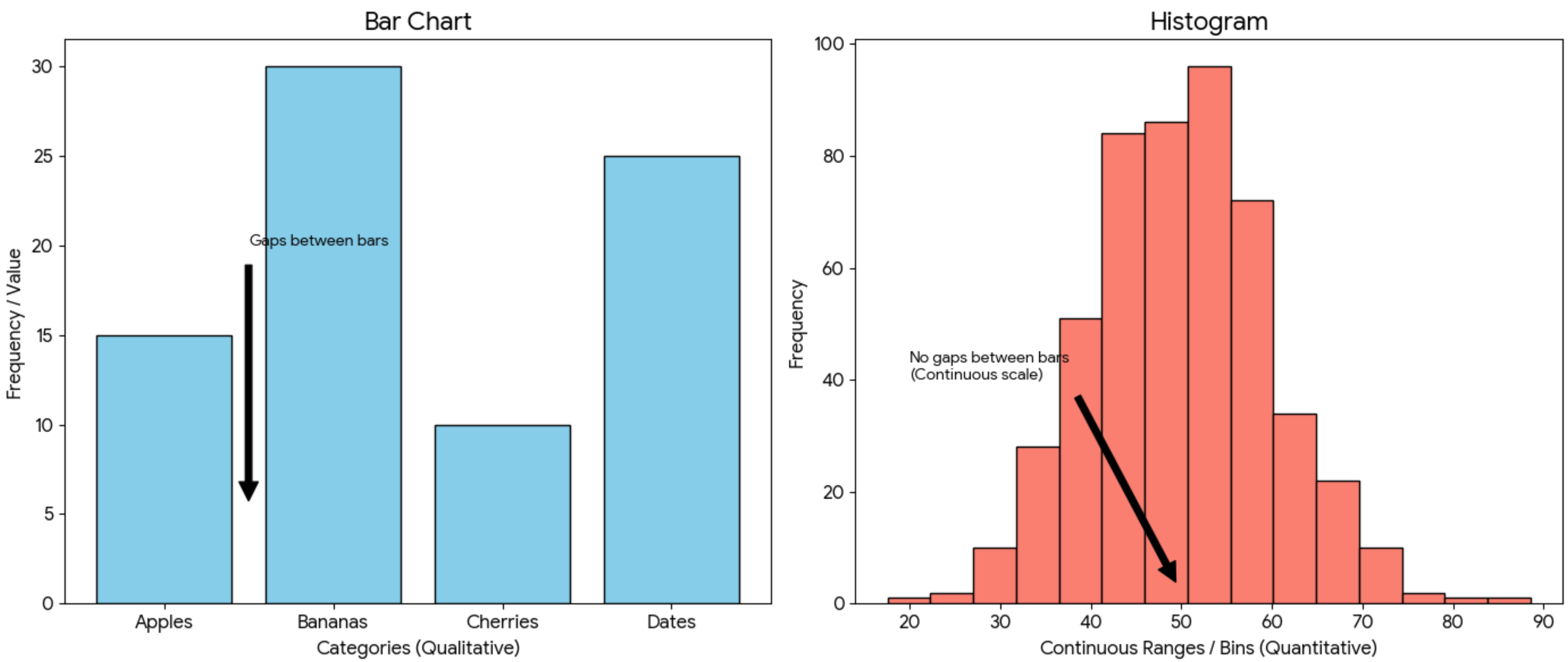


Blood counts among 102 patients with NSCLC prior to treatment with SABR

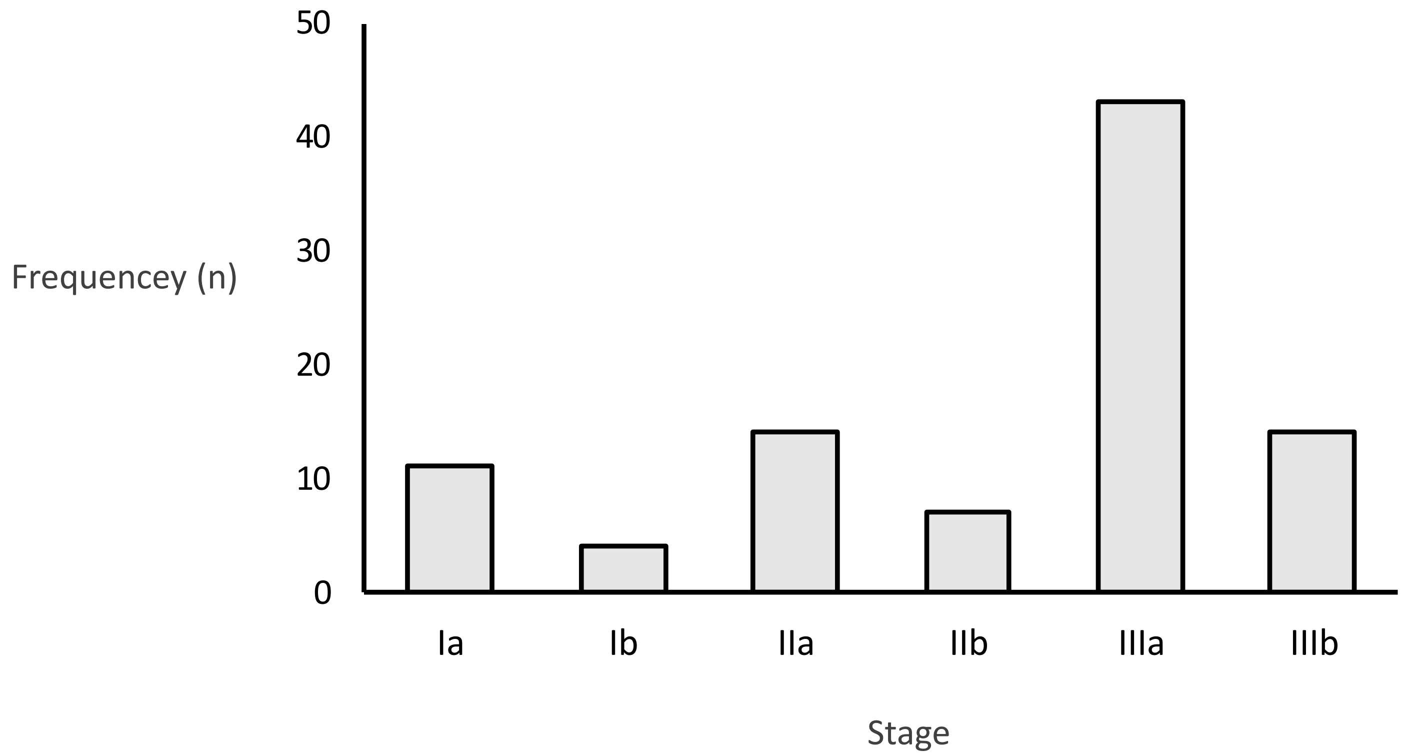
Bar chart (categorical data)



Variable values

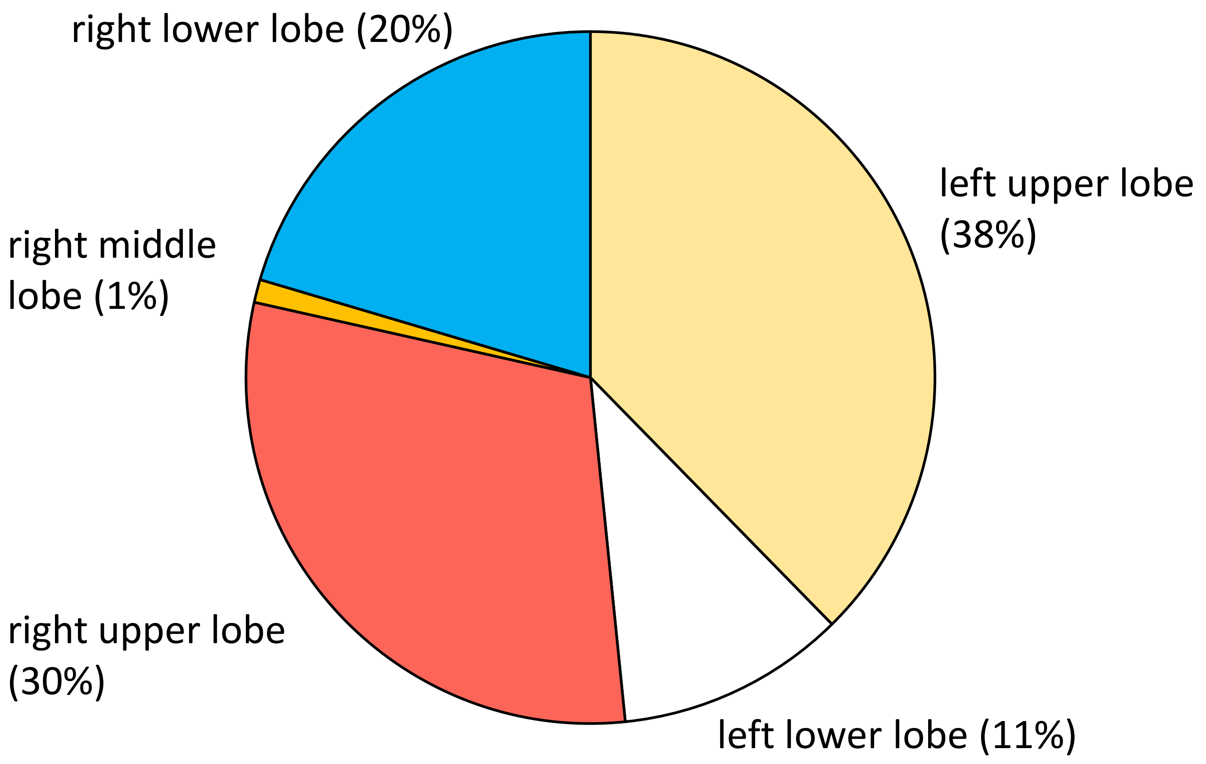

Bar Chart
Gaps between bars
Frequency / Value
Apples
Bananas
Cherries
Dates
Categories (Qualitative)
Histogram
No gaps between bars
(Continuous scale)
Frequency
Continuous Ranges / Bins (Quantitative)

Bar chart



Stage among 93 patients with lung cancer treated with radical radiotherapy

Pie Chart (nominal data)



Location of tumour among 93 patients with lung cancer who received radical radiotherapy

Contingency Tables

relationship between two categorical variables

Sex	Total (n)
Women	60
Men	42
Total	102

Status at two years	Total (n)
Alive	77
Dead	25
Total	102

Status at two years	Women (n)	Men (n)	Total (n)
Alive	48	29	77
Dead	12	13	25
Total	60	42	102

102 patients with NSCLC treated with SABR

Contingency Tables

relationship between two categorical variables

Status at two years	Women (n)	Men (n)	Total (n)
Alive	48	29	77
Dead	12	13	25
Total	60	42	102

Status at two years	Women (n)	Men (n)	Total (n)
Alive	48 (80%)	29 (69%)	77 (75%)
Dead	12 (20%)	13 (31%)	25 (25%)
Total	60 (100%)	42 (100%)	102 (100%)

Status at two years	Women (n)	Men (n)	Total (n)
Alive	48 (80%)	29 (69%)	77 (75%)
Dead	12 (20%)	13 (31%)	25 (25%)
Total	60 (100%)	42 (100%)	102 (100%)

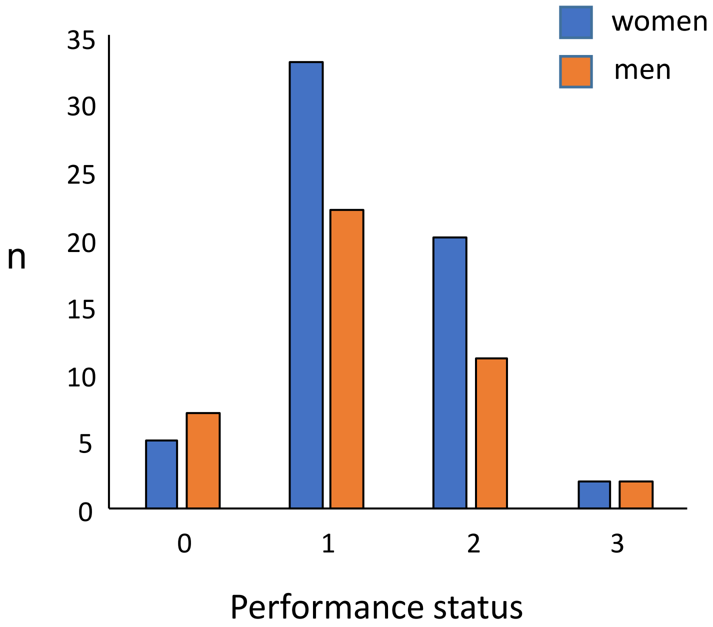
column %

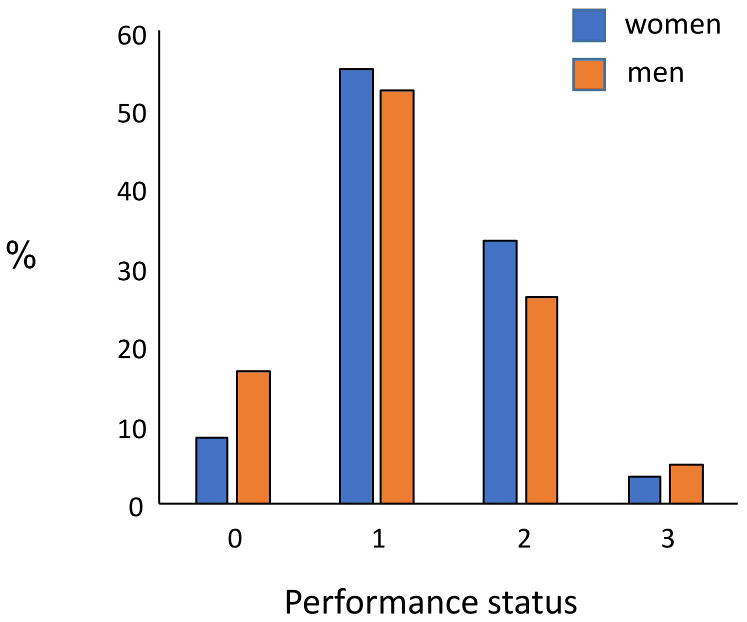
Status at two years	Women (n)	Men (n)	Total (n)
Alive	48 (62%)	29 (38%)	77 (100%)
Dead	12 (48%)	13 (52%)	25 (100%)
Total	60 (59%)	42 (41%)	102 (100%)

row %

Status at two years	Women (n)	Men (n)	Total (n)
Alive	48 (47%)	29 (28%)	77 (75%)
Dead	12 (12%)	13 (13%)	25 (25%)
Total	60 (59%)	42 (41%)	102 (100%)

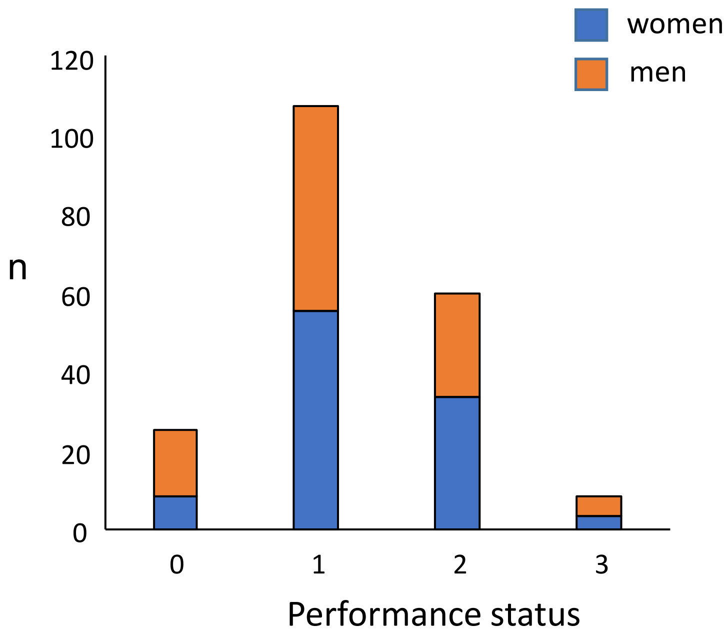
Clustered (grouped) bar charts

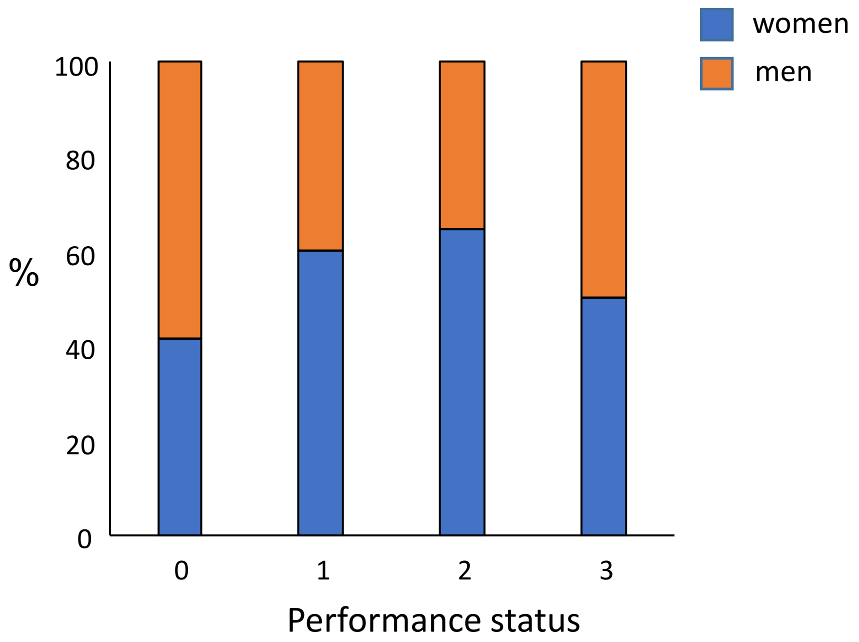




102 patients with NSCLC treated with SABR

Stacked (component), percentage component bar charts





102 patients with NSCLC treated with SABR

Illustrating data

Continuous data

Histograms, scatter plots, box-plots, dot plots

Categorical (nominal) data

Bar charts, pie charts, frequency tables

Descriptive measures

A descriptive measure is a numerical value, which summarises a set of data

number of patients in a sample is **n**

We represent the individual values of a variable **X** with a small letter **x**.

the value for patient 1 is $\mathbf{x_1}$

the value for patient 2 is $\mathbf{x_2}$

the value for patient n is $\mathbf{x_n}$

Measures of Location or Central Tendency

Referred to as ‘average’ values

mean, median, and ***mode***.

Sample mean ($\bar{x}$) is calculated as

Sum of all values

$$mean = \frac{\sum x}{n}$$

$\sum x = x_1 + x_2 + x_3 + \ldots + x_n$ is the sum of all values across the subjects, and n is the total number of subjects.

10 patients with lung cancer had a pretreatment PET scan and SUV_{max} was measured.

Original Data: 1.8, 8.9, 2.7, 9.4, 5.4, 16.0, 5.8, 17.9, 13.1, 6.6

$$\text{The } \textbf{mean } SUV_{max} = \frac{1.8+8.9+2.7+9.4+5.4+16.0+5.8+17.9+13.1+6.6}{10} = 8.8$$

Sample median is the *middle value* when the observations are ranked from lowest to highest.

If n is even, it is the mean of the middle two values.

Its interpretation is that 50% of data values are above the median; 50% are below the median.

Sample mode is the *most frequently* occurring value. This term is seldom used.

Original Data: 1.8, 8.9, 2.7, 9.4, 5.4, 16.0, 5.8, 17.9, 13.1, 6.6

Ordered Data: 1.8, 2.7, 5.4, 5.8, 6.6, 8.9, 9.4, 13.1, 16.0, 17.9 1000.0

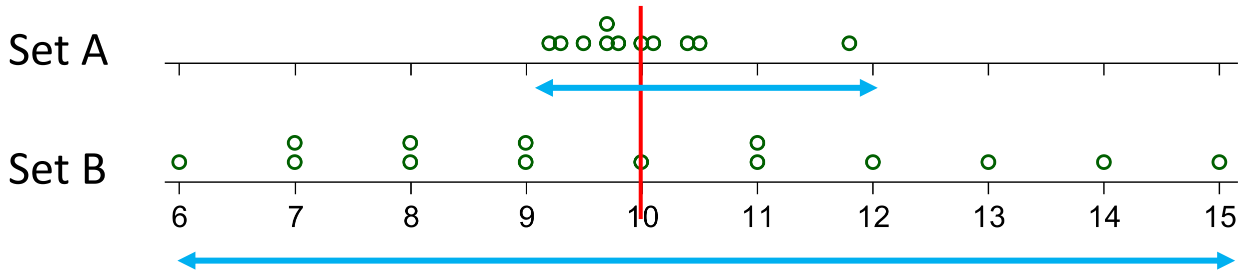
median SUV_{max} = (6.6+8.9)/2 =7.75

Neither the mean nor the median is sufficient as a numerical summary.

Example: - 2 datasets

Set A : 9.2 9.3 9.5 9.7 9.8 10.0 10.1 10.4 10.5 11.8

Set B : 6 7 7 8 8 9 9 10 11 11 12 13 14 15



Each set has a mean of 10.0 units

***Spread of values is different*.**

How should spread be measure?

Measures of Dispersion

How “concentrated” or “spread out” are the data.

They describe the degree to which the data vary about their average value.

range, standard deviation, and ***interquartile range***.

Measures of Dispersion

Range is the difference between the smallest (minimum) and largest (maximum) values

Interquartile range (IQR) is simply the lower quartile and upper quartile (Q1, Q3). Sometimes it is expressed as the value of Q3-Q1.

Original Data: 1.8, 8.9, 2.7, 9.4, 5.4, 16.0, 5.8, 17.9, 13.1, 6.6

Ordered Data: 1.8, 2.7, 5.4, 5.8, 6.6, 8.9, 9.4, 13.1, 16.0, 17.9

median SUV_{max} = (6.6+8.9)/2 =7.75

Q1=5.4 and **Q3**=13.1

Interquartile range (IQR) is (5.4 to 13.1) *or* 13.1-5.4 =7.7

Measures of Dispersion

Standard deviation (**SD or sd**) is a measure of how far away observations are from the sample mean.

distance from the mean

$$SD = \sqrt{\frac{\sum(x - mean)^2}{n - 1}}$$

Original Data: 1.8, 8.9, 2.7, 9.4, 5.4, 16.0, 5.8, 17.9, 13.1, 6.6

standard deviation involves subtracting the mean SUVmax (8.8) from each observation

distance from the mean

-7, 0.1, -6.1, 0.6, -3.4, 7.2, -3.0, 9.1, 4.3, -2.2

squaring

Squared distance from the mean

49.0, 0.01, 37.21, 0.36, 11.56, 51.84, 9.0, 82.81, 18.49, 4.84

summing

Sum of Squared distance from the mean

265.12

dividing by **n-1** $\frac{265.12}{(10-1)}$ = 29.46

take the square root **standard deviation** = √29.46 = 5.43

Reporting means and medians

Means with SD	mean SUV_{max} 8.8 (SD 5.43)
Medians with IQR	median SUV_{max} 7.75 (IQR 5.4 to 13.1)

Do not mix medians with SD, or means with IQR

Which measure to use

The **mode** should be used when describing nominal categorical variables.

When the variable is numeric with a symmetric distribution, then the **mean** is proper measure of center

In the case of skewed distributions, the **median** is better choice for the measure of center.

The **median** is less influenced by **outliers** (extreme values).

1. Categorical data are best illustrated using

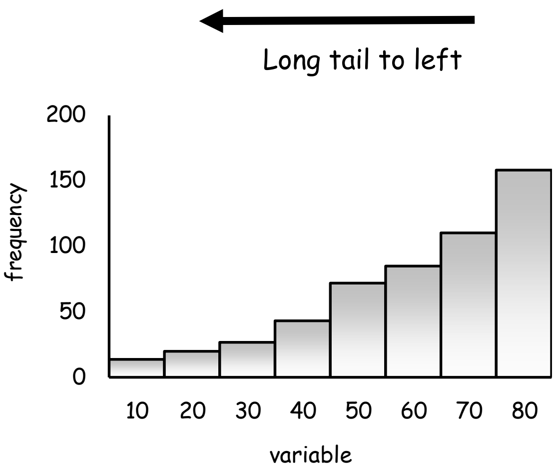
   A. bar charts
   B. box-plots
   C. histograms
   D. means and standard deviations
   E. scatter plots

3. In the case of skewed distributions, which is the best choice of central tendency

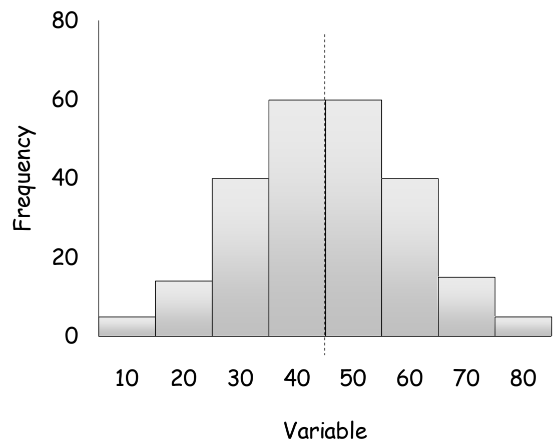
   A. the interquartile range
   B. the mean
   C. the median
   D. the mode
   E. the standard deviation

The Normal Distribution

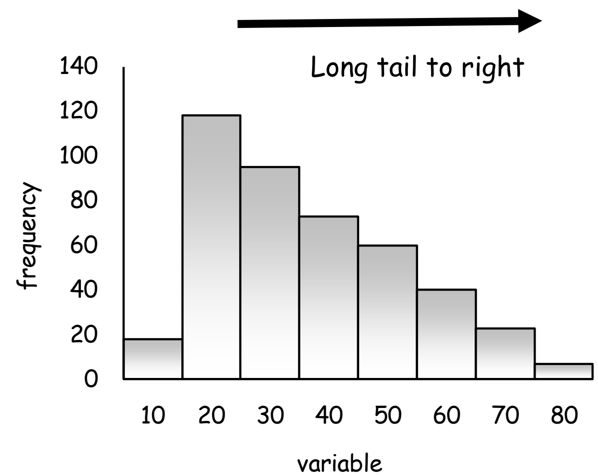
Negatively skewed
(long tail to the left)



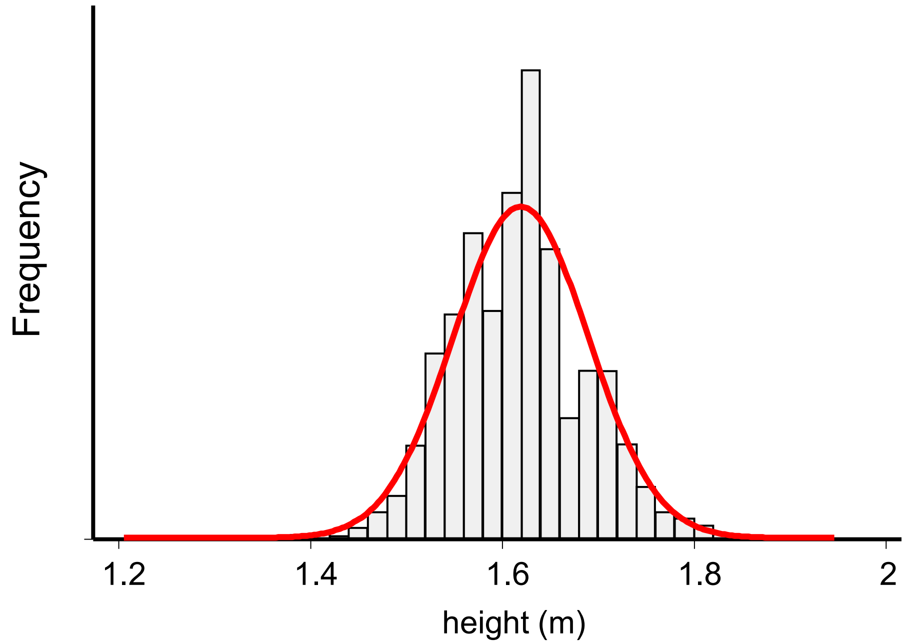
Symmetric (bell shaped) histogram

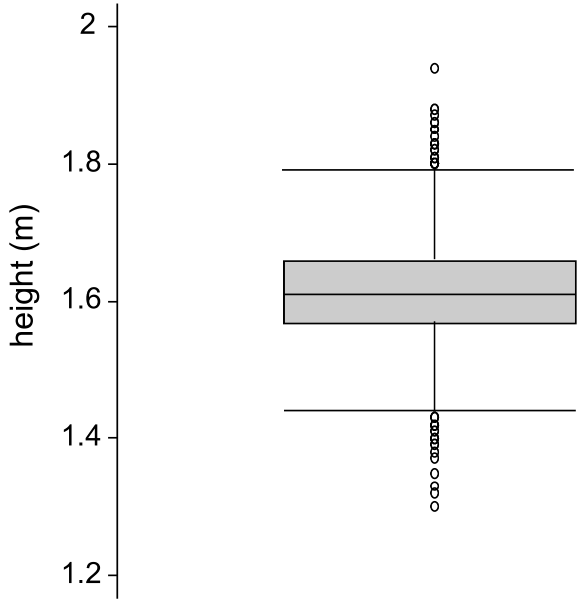


Positively skewed
(long tail to the right)



Normal distribution



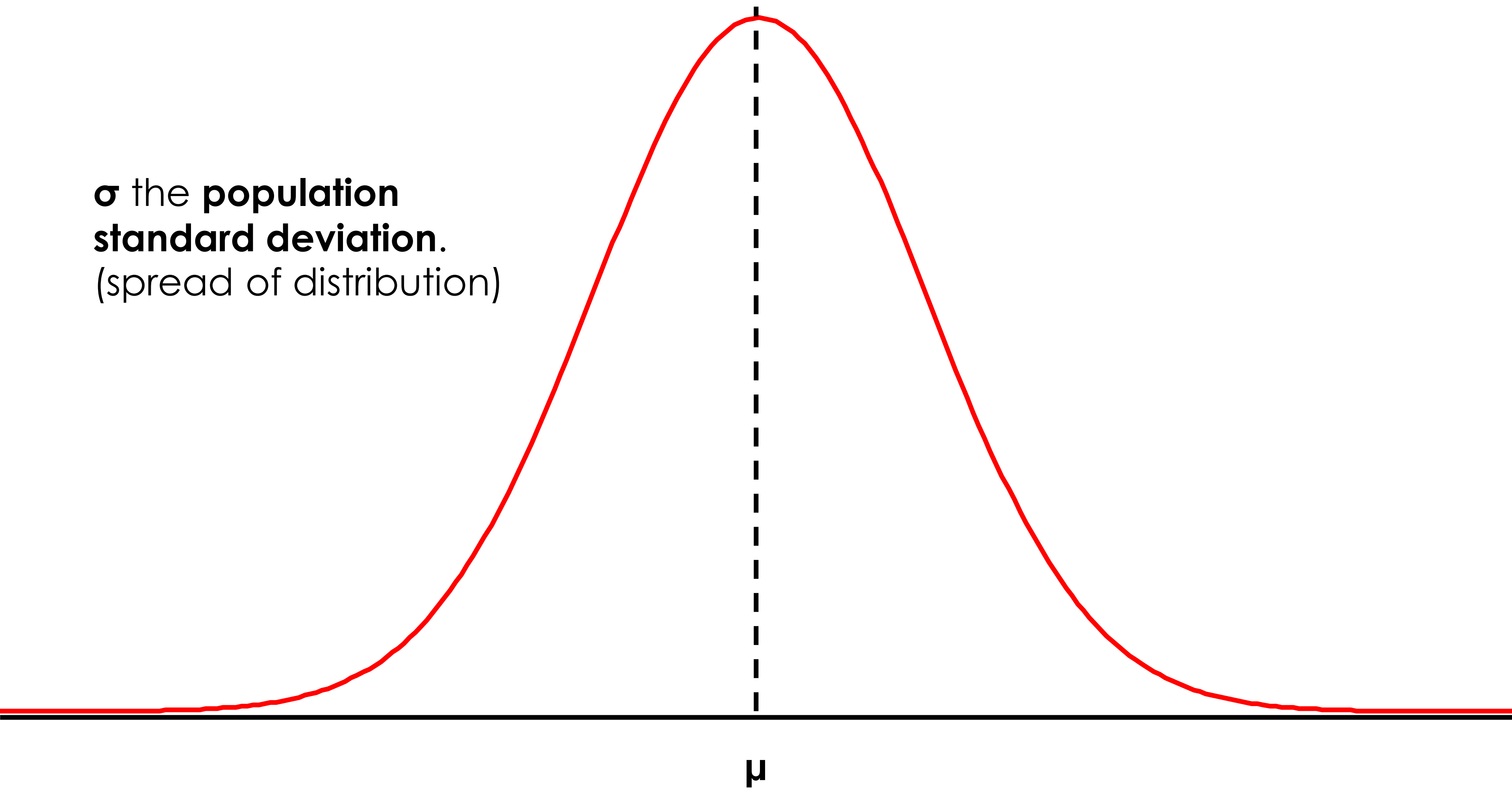


Histogram of height of women attending a weight management service

The characteristic symmetric bell-shape of the Normal distribution.

Normal distribution is completely specified by the mean and standard deviation

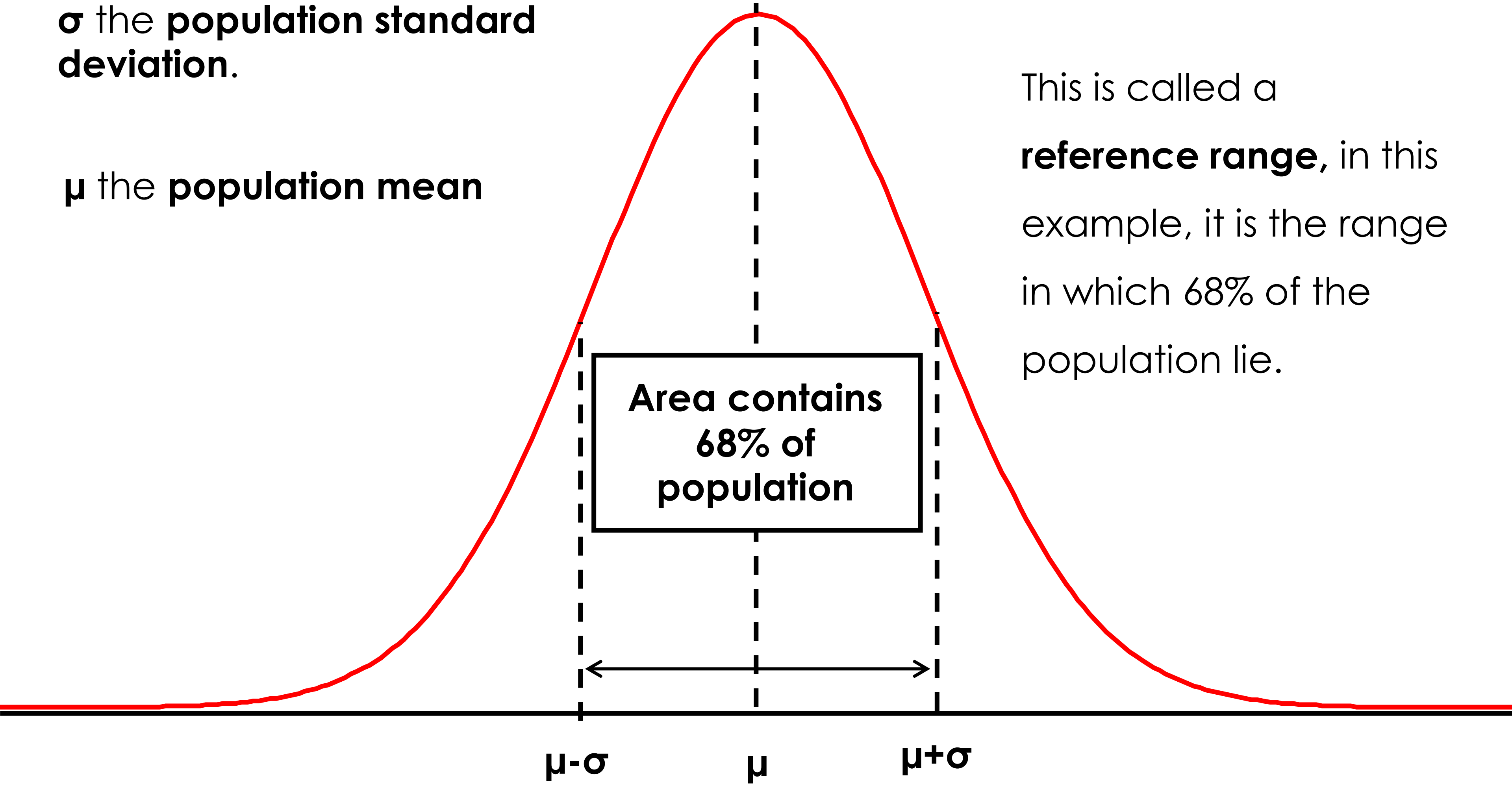
The Normal distribution is described by parameters **μ** and **σ**,



The Normal distribution is described by parameters **μ** and **σ**,

σ the **population standard deviation**.

μ the **population mean**



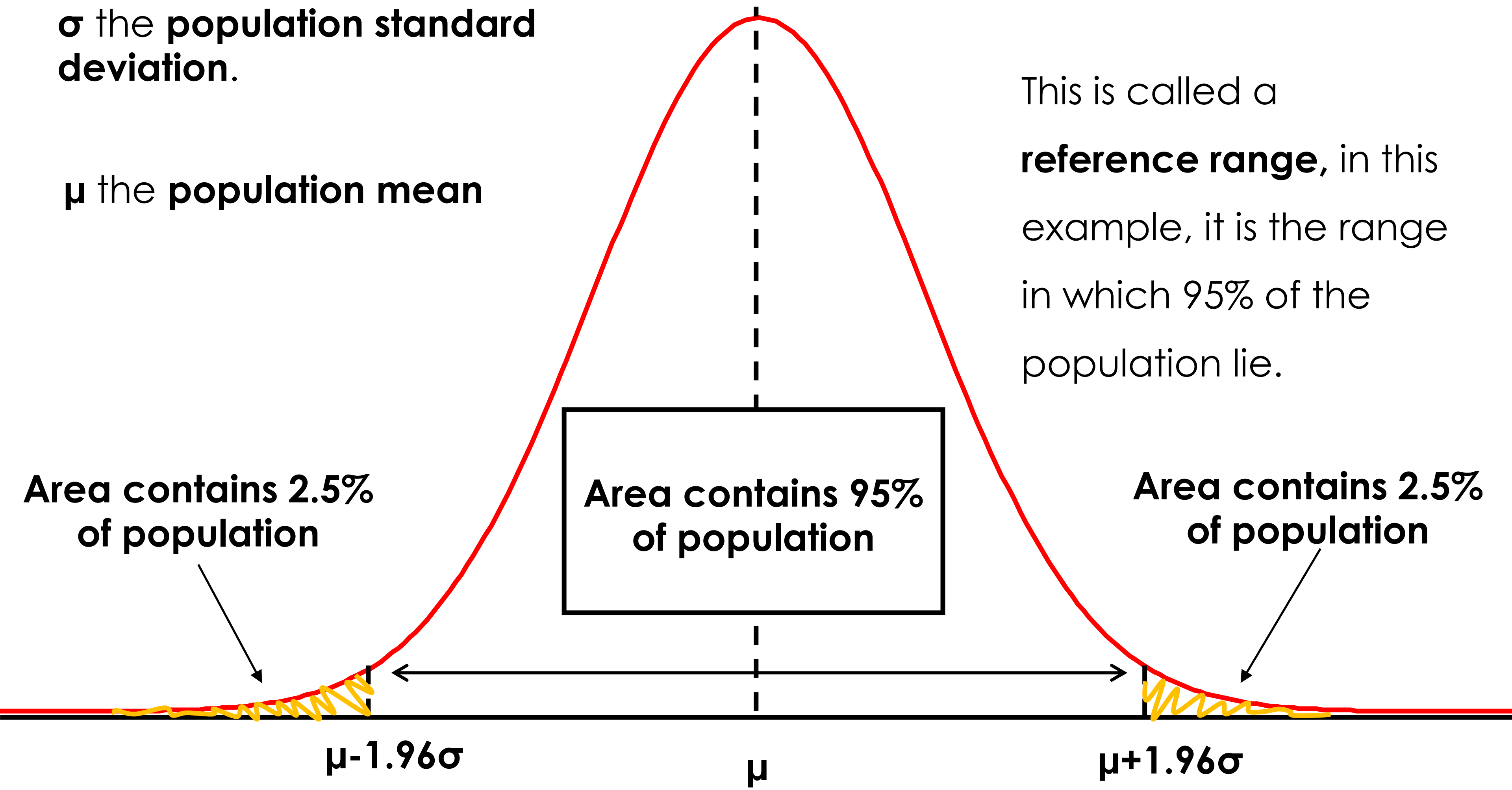
This is called a **reference range,** in this example, it is the range in which 68% of the population lie.

The Normal distribution is described by parameters **μ** and **σ**,

σ the **population standard deviation**.

μ the **population mean**

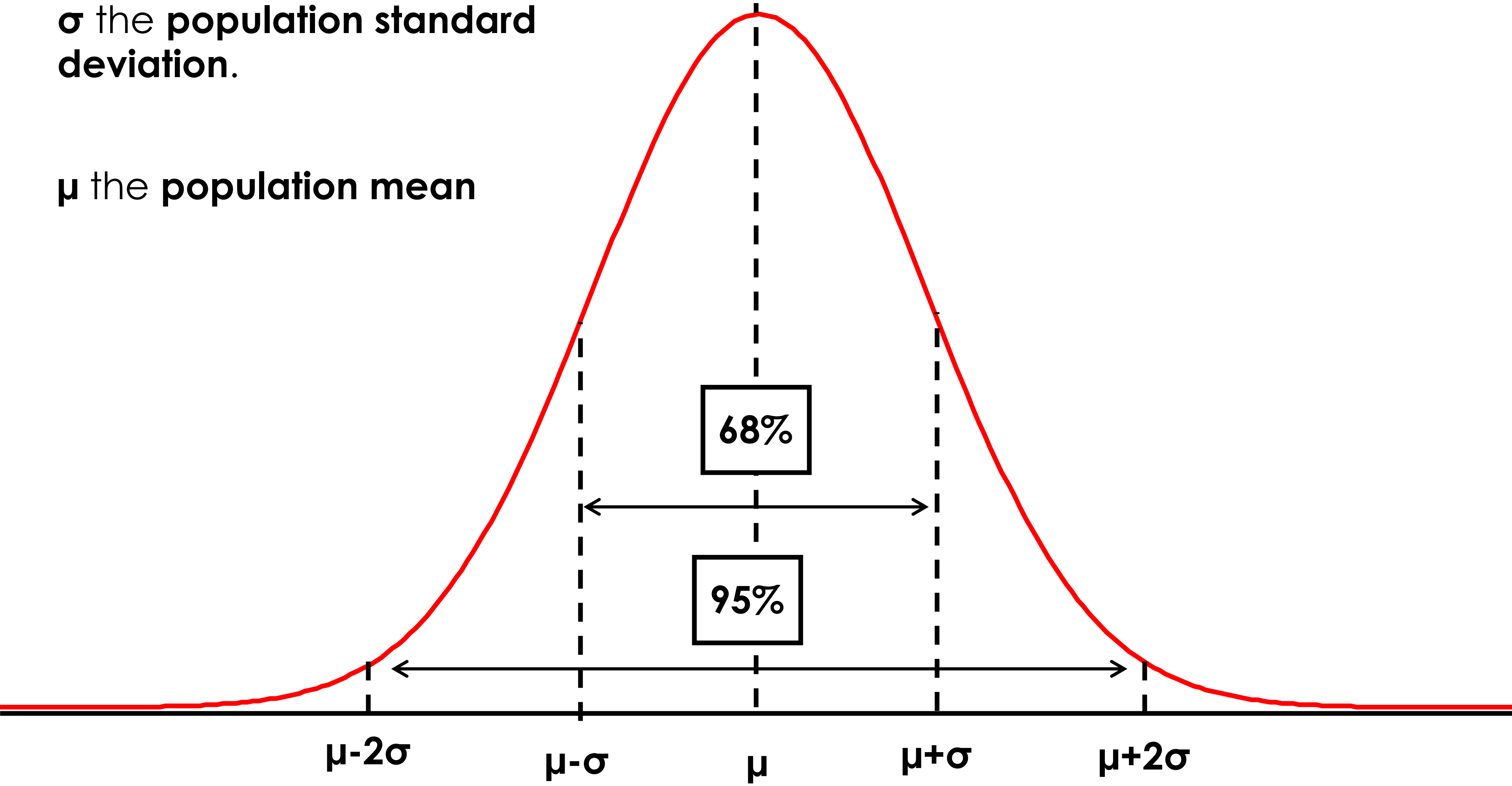
This is called a **reference range,** in this example, it is the range in which 95% of the population lie.



The Normal distribution is described by parameters **μ** and **σ**,

σ the **population standard deviation**.

μ the **population mean**



1. In a normal distribution 95% of values lie within
   a) the range
   b) the interquartile range
   c) ± 1 standard deviation from the mean
   d) ± 1.5 standard deviations from the mean
   e) ± 2 standard deviations from the mean

2. In a normal distribution it is expected that
   a) the median and mean will be the same
   b) the median will be greater than the mean
   c) the median will be smaller than the mean
   d) the median cannot be calculated
   e) the mean and median will not be the same

Exercise

- In excel, calculate the mean and standard deviation

	SUV_{max}
	1.8
	8.9
	2.7
	9.4
	5.4
	16
	5.8
	17.9
	13.1
	6.6
sum=	87.6
mean=	8.76

	SUV_{max}	$(SUV_{max}\text{-mean})$
	1.8	-6.96
	8.9	0.14
	2.7	-6.06
	9.4	0.64
	5.4	-3.36
	16	7.24
	5.8	-2.96
	17.9	9.14
	13.1	4.34
	6.6	-2.16
sum=	87.6	0
mean=	8.76	

	SUV_{max}	$(SUV_{max}$-mean)		$(SUV_{max}$-mean$)^2$
	1.8	-6.96		48.44
	8.9	0.14		0.02
	2.7	-6.06		36.72
	9.4	0.64		0.41
	5.4	-3.36		11.29
	16	7.24		52.42
	5.8	-2.96		8.76
	17.9	9.14		83.54
	13.1	4.34		18.84
	6.6	-2.16		4.67
sum=	87.6	0		265.10
mean=	8.76		SD=	5.43