APPLIED PHYSICS AND BIOSTATISTICS

MODULE MATH7 — Statistics: Describing and Comparing Biomedical Data

Biological Variability, Summary Measures, Normal Distribution and Inference

Prof. Valentina Notarstefano | Università degli Studi di Teramo

MODULE MATH7 — BIOSTATISTICS

Why Statistics Matters for a Biotechnologist

Biology is variable: repeat a measurement and you get a spread of values, not one number.

Descriptive statistics summarise that spread — mean, median, standard deviation, quartiles.

The normal distribution turns a spread into a reference interval (mean ± 2 SD = central 95%).

Inferential statistics ask whether a difference is real: treated vs control (the t-test).

And they tell us how good a diagnostic test is: its sensitivity and specificity.

Numerical Data

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Categorical Data

- Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Population, Statistical Unit, Statistical Sample

- Statistical population: set of elements to which the statistical investigation refers:
  - opinions of Americans regarding a new presidential election: all USA citizens
  - genes overexpressed in individuals affected by obesity: all individuals with obesity
  - …

- Statistical unit: each element of the statistical population, the smallest unit from which data are collected:
  - A citizen, an individual with obesity….

- Statistical sample (sample): any set of statistical units drawn from the entire population. A sample is therefore a subset of measurements selected from the population
  - 50 individuals with obesity-related conditions (randomly selected).

Quantitative and Qualitative Variables

- **Quantitative variable**: when it takes numerical values:
  - Continuous: it assumes continuous values within an interval (a person's weight and height, intensity levels of samples on microarray, level of gene expression, etc.)
  - Discrete: it assumes discrete values such as number of samples, number of over-expressed genes, number of patients, etc.

- **Qualitative variable**: when it takes non-numerical values
  - Ordinal: the data are ordered (good–average–poor, cold–lukewarm–hot…)
  - Categorical: male/female, phenotype, groups of diseased/healthy patients, etc.

Absolute and Percentage Frequencies

- When the sample for which we intend to describe the statistical variables is very large, rather than considering all values, one may report only distinct values and indicate, for each value, how many times it occurs.
- Let $Y = (y_1, y_2, \ldots, y_N)$ be a discrete statistical variable. We define as modalities the distinct values among $y_1, \ldots, y_N$ and as absolute frequency of a modality the number of times it is observed in the expression of the statistical variable.

Y=(60, 80, 92, 100, 83, 96, 74, 63, 80, 100, 90, 75, 74, 92)

Modality	60	80	92	100	83	96	74	63	90
Absolute frequency	1	2	2	2	1	1	2	1	1

Absolute and Percentage Frequencies

Let Y be a statistical variable, and f the absolute frequency of modality z.

We define the relative frequency of modality z as the ratio f/N, where N is the number of elements in the sample.

The percentage frequency is obtained by multiplying the relative frequency by 100.

Mean

- Mean of a population: sum of all values of the population variables divided by the number of units in the population (N)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Where N is the number of elements in the population, X_i is the i-th observation of the variable X_i

- Mean of a sample: sum of all values of the variables in a subset of the population divided by the number of units in that sample (n)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where n is the number of elements in the sample, X_i is the i-th observation of the variable X_i

Given the following set of measurements of gene expression levels:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

mean of the population:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

mean of the sample: (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

The mean of any sample $\bar{X}$ may differ substantially from that of the entire population μ. The larger the sample, the closer the sample mean will be to the population mean

Mean

- **Weighted mean of a population**: a weight is assigned to each variable; all variable values are summed, each multiplied by its weight, and the resulting total is divided by the sum of the weights

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

- The expected value of a variable X, denoted by $E[X]$, is defined as the mean of X computed over a large number of experiments

Exercise 1 – Arithmetic and Weighted Mean

a) Expression level of a gene (relative units) in n = 8 samples:

12, 15, 11, 14, 18, 13, 16, 13

Calculate the sample mean.

b) HbA1c (%) is measured in three groups of patients of different size:

Group	Patients (n)	Mean HbA1c (%)
A	10	5.2
B	30	6.0
C	20	5.5

Calculate the overall mean of the 60 patients. Why is it not enough to average the three means?

Reminders

sample mean: $\bar{x} = \Sigma\, x_i / n$

weighted mean: $\Sigma\,(p_i \cdot x_i) / \Sigma\, p_i$

(weights p_i = group sizes)

Exercise 1 – Arithmetic and Weighted Mean: Solution

a) Sample mean

$\bar{x} = (12 + 15 + 11 + 14 + 18 + 13 + 16 + 13) / 8 = 112 / 8 =$ **14**

b) Weighted mean (the weight of each group is its number of patients)

- Sum of the weights: $\Sigma p_i = 10 + 30 + 20 = 60$
- Sum of the products: $\Sigma p_i \cdot x_i = 10 \cdot 5.2 + 30 \cdot 6.0 + 20 \cdot 5.5 = 52 + 180 + 110 = 342$
- Weighted mean = 342 / 60 = **5.7 %**

Why not the mean of the means? (5.2 + 6.0 + 5.5) / 3 = 5.57 %: it would give the same weight to a group of 10 patients and to one of 30. Group B, being larger, must count more.

Mode

- The mode is the most frequent value of a distribution, or more precisely, the most recurrent category of the variable (i.e., the one associated with the highest frequency).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

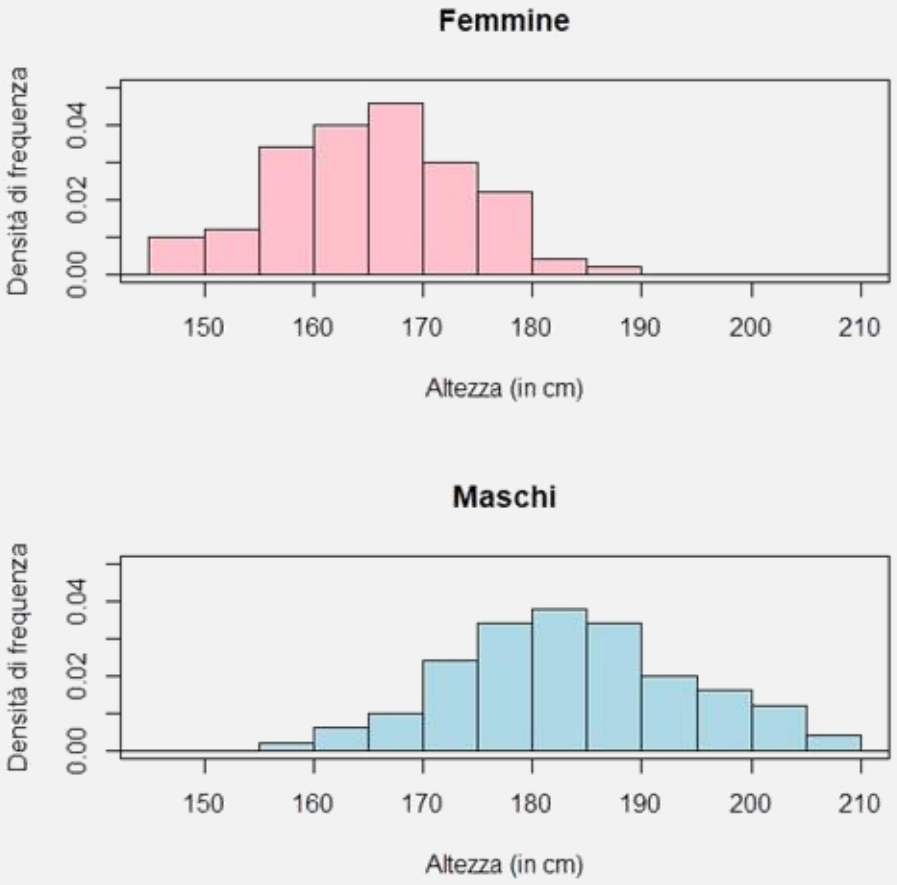
- The mode of this sample is 1005, as it occurs 3 times.

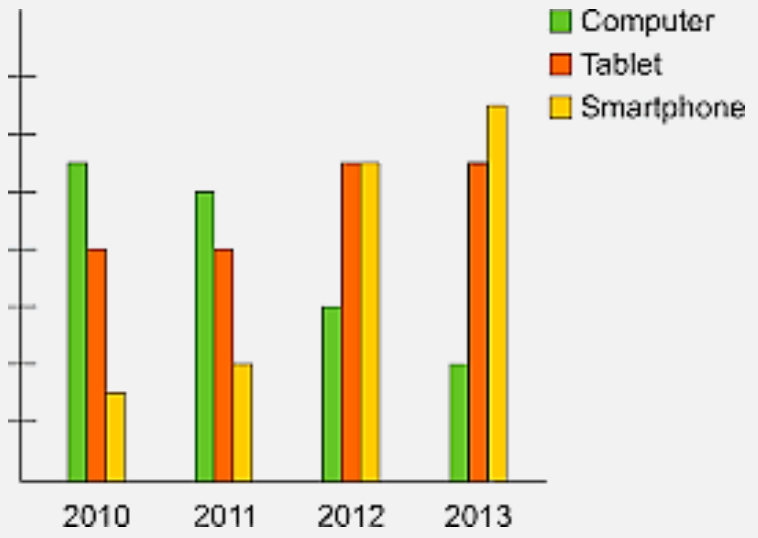
Characteristics of the mode:

- it is used solely for descriptive purposes, because it is less stable and less objective than other measures of central tendency
- to identify the mode of a distribution, graphical methods can be used, such as histograms
- it may differ within the same data series when distribution classes (intervals) of different widths are formed
- to identify the mode within a frequency class, without knowing how the data are distributed, the assumption of a uniform distribution is adopted.

Histograms

- A histogram describes the relative frequency of data within an interval (a, b) and is used to visualize the data distribution.





Unimodal/Bimodal Distributions

- A distribution may exhibit multiple modes:
  - **Unimodal distributions**: frequency distributions that have a single mode, that is, a single point of maximum (representing both the relative and the absolute maximum)
  - **Bimodal or k-modal distributions**: frequency distributions that exhibit two or more modes, that is, two (or k) relative maxima
    - Example: by measuring the heights of a group of adolescents in which the larger part is female and the smaller part male, one obtains a bimodal distribution, with a primary and a secondary mode.

Zero-Modal Distribution

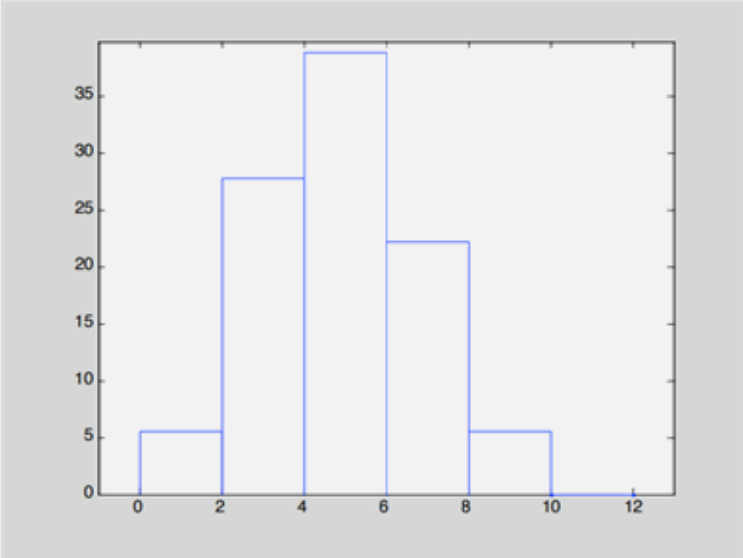
- No value has a higher frequency than the others:

$$A = \{1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6\}$$

Unimodal distribution

C'è un solo valore con una frequenza più elevata degli altri

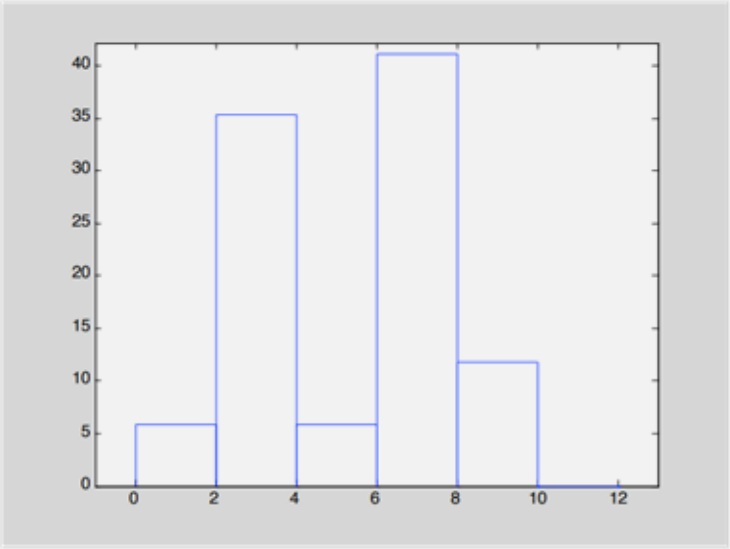
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}

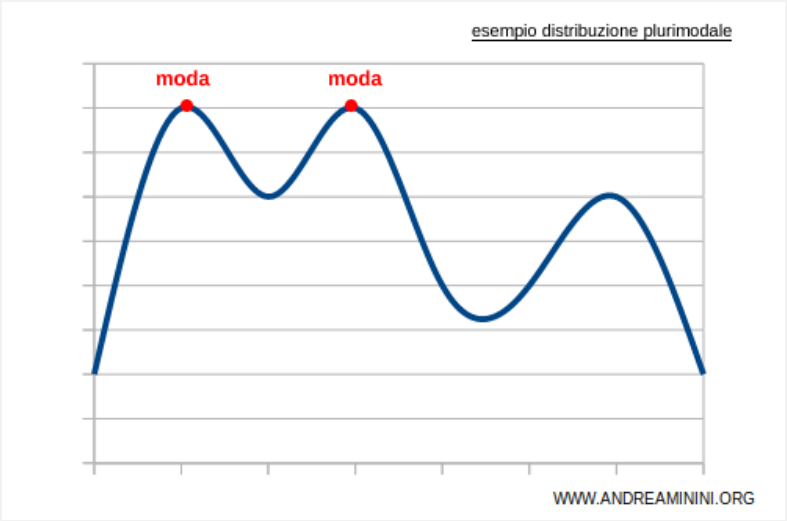


Bimodal distribution

Ci sono due valori con una frequenza più elevata degli altri.

A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}





Exercise 2 – Frequencies and Mode

In a clinic, the number of visits in one year is recorded for N = 20 patients:

3, 2, 2, 3, 1, 1, 4, 1, 2, 0, 1, 1, 2, 1, 2, 0, 0, 2, 1, 1

Tasks:

- build the table with value, absolute, relative and percentage frequency;
- find the mode;
- calculate the percentage of patients with at least 2 visits.

Reminders

relative frequency = f / N

percentage frequency = f / N × 100

mode = value with the highest absolute frequency

Exercise 2 – Frequencies and Mode: Solution

Step 1. Count how many times each value appears: this is the absolute frequency f. Check: Σ f = N = 20.

Step 2. Divide each f by N = 20 (relative frequency) and multiply by 100 (percentage). Example, value 1: 8 / 20 = 0.40 → 40%.

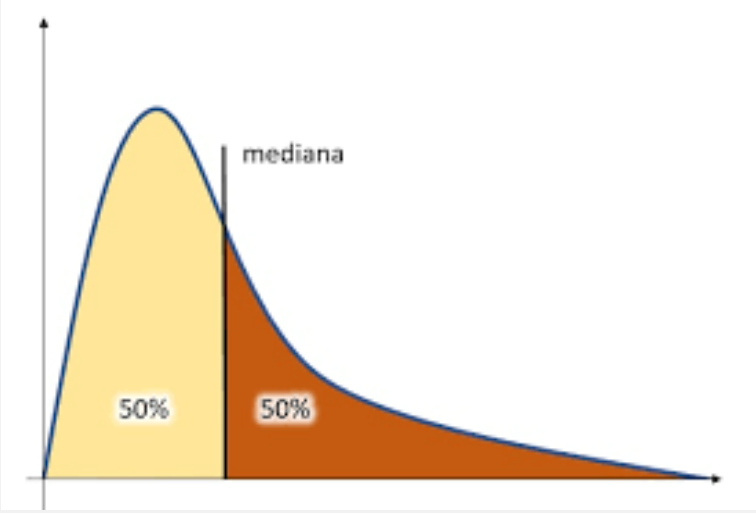
b) Mode: the highest frequency is 8, so the mode is **1 visit**.

c) At least 2 visits: values 2, 3 and 4 → 6 + 2 + 1 = 9 patients → 9 / 20 = 0.45 → **45%**.

Value	f	f / N	%
0	3	0.15	15%
1	**8**	**0.40**	**40%**
2	6	0.30	30%
3	2	0.10	10%
4	1	0.05	5%
Total	**20**	**1.00**	**100%**

Median

- The median is the value that occupies the central position in an ordered set of data.
- It is a robust measure, as it is minimally influenced by the presence of outliers.
- Characteristics:
  - it is used when one seeks to mitigate the effect of extreme values;
  - in a distribution or data series, each value drawn at random has the same probability of being below or above the median.



Calculation of the Median

To calculate the median of a dataset, one must:

1. arrange the values in ascending or descending order and count the total number n of data;
2. if the number (n) of data is odd, the median corresponds to the numerical value of the central datum, i.e., the one occupying position $(n+1)/2$;
3. if the number (n) of data is even, the median is estimated by using the two central values occupying positions $n/2$ and $n/2+1$:
   a. with a small number of observations, the median is taken to be the arithmetic mean of these two intermediate observations;
   b. with a large number of observations grouped into classes, one sometimes resorts to proportions.

Example.
Consider the following sample:

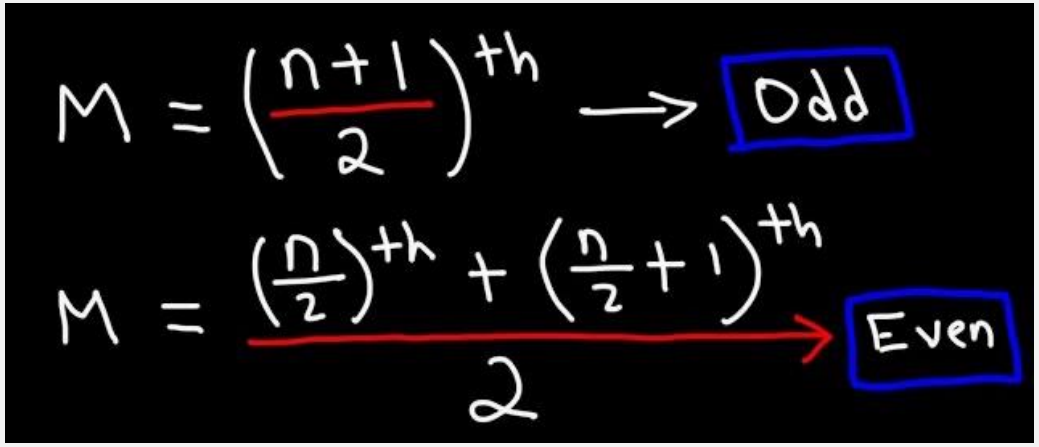
96	78	90	62	73	89	92	84	76	86

Order the samples in ascending order:

62	73	76	78	**84**	**86**	89	90	92	95

1. Since the number of samples is even ($n = 10$), the median is computed as the mean of the two central elements:

$$median = 84 + 862 = 85$$



Exercise 3 – Median and Robustness

C-reactive protein concentration (mg/L) in 7 patients:

6, 35, 4, 7, 5, 5, 8

- Calculate the median and the mean.
- An eighth patient with value 6 is added: recalculate the median.
- If the value 35 were actually 350, how would the mean and the median change? Which index is more robust?

Reminders

median: sort the data;

n odd → position (n + 1)/2

n even → mean of positions n/2 and n/2 + 1

Exercise 3 – Median and Robustness: Solution

a) n = 7. Sorted data: 4 · 5 · 5 · **6** · 7 · 8 · 35

- n is odd → position (7 + 1)/2 = 4 → median = **6 mg/L**
- Mean = (4 + 5 + 5 + 6 + 7 + 8 + 35) / 7 = 70 / 7 = **10 mg/L**

b) n = 8. Sorted: 4 · 5 · 5 · **6** · **6** · 7 · 8 · 35

- n is even → positions n/2 = 4 and n/2 + 1 = 5 → median = (6 + 6) / 2 = **6 mg/L**

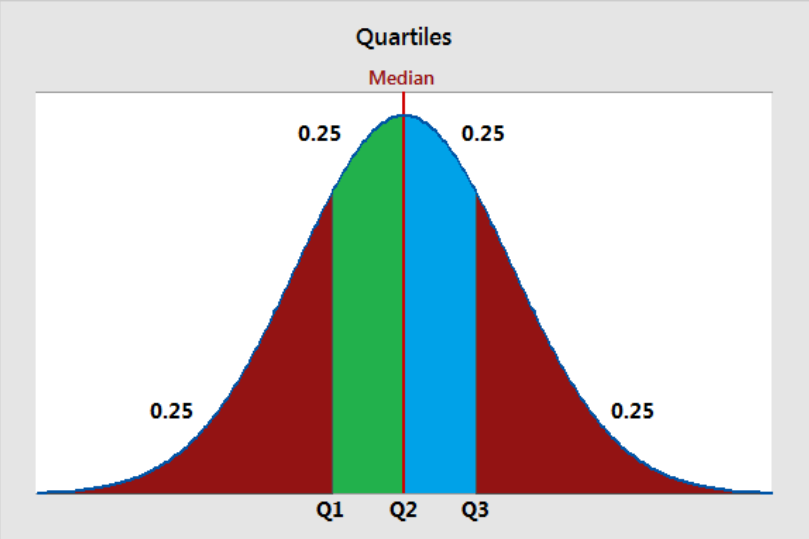
c) Outlier 350 instead of 35 (n = 7)

- Sum = 70 − 35 + 350 = 385 → mean = 385 / 7 = **55 mg/L**
- Sorted data: 4 · 5 · 5 · 6 · 7 · 8 · 350 → the 4th value is still 6 → median = **6 mg/L**

The median does not change, the mean does: the median is robust to outliers.

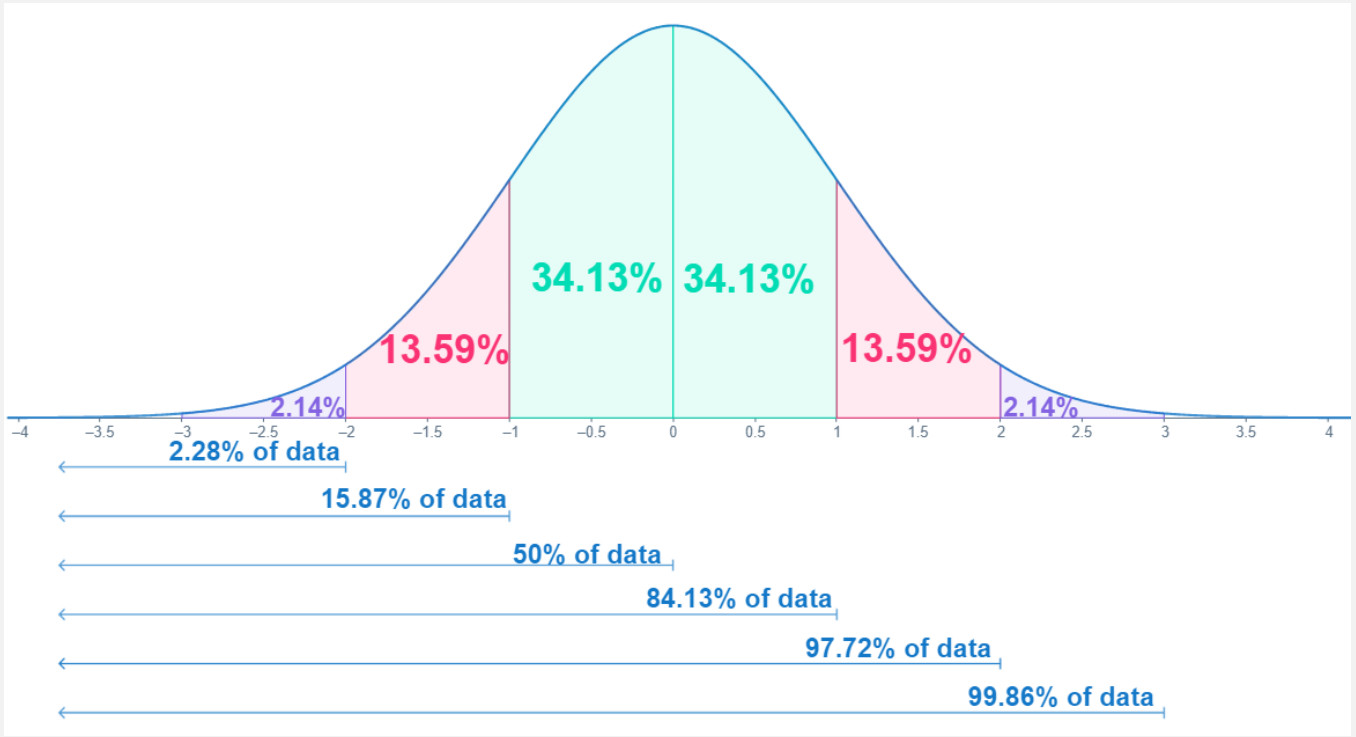
Quantiles

- The quantiles are a family of measures, to which the median also belongs, that are distinguished according to the number of equal parts into which they divide a distribution.
- The quartiles partition the distribution into 4 parts of equal frequency, where each part contains the same fraction of observations:
  - The first quartile is defined as the number $q1$ such that 25% of the statistical data are less than or equal to $q1$.
  - The second quartile is defined as the number $q2$ such that 50% of the statistical data are less than or equal to $q2$. The second quartile corresponds to the median.
  - The third quartile is defined as a number $q3$ such that 75% of the statistical data are less than or equal to $q3$.



Deciles and Percentiles

- Analogously, the following are defined:
  - Deciles: 9 points that divide the ordered distribution into 10 equal parts
  - Percentiles: 99 points that divide the ordered distribution into 100 equal parts



Range of Variation

- The range of a distribution is the difference between the largest and the smallest value in the distribution:

$$C = x_{max} - x_{min}$$

This index is relatively coarse, as it provides no information regarding the variability of intermediate values.

- Example: the range of the following distribution:

$$25 - 26 - 28 - 29 - 30 - 32 \rightarrow C = 32 - 25 = 7$$

Variance

- Variance of the population: optimal measure to characterize the variability of a population. An indicator of the dispersion of a variable or statistical distribution that one obtains by calculating the average of the squares of the deviations from the arithmetic mean (μ).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

where N is the number of observations of the entire population; μ is the mean of the population; X_i is the i-th observation

- Variance of a sample:

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

where n is the number of observations of the sample; $\bar{X}$ is the mean of the sample; x_i is the i-th observation

When n is high enough, the difference between the two formulas are tiny.

Standard Deviation, or Root Mean Square Deviation

- Population standard deviation:

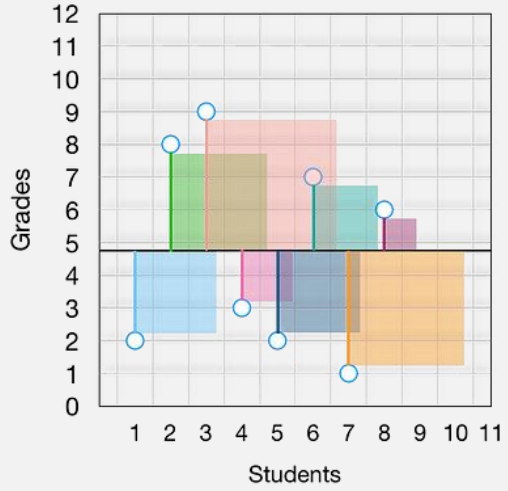
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

where N is the number of observations in the entire population; μ is the population mean; X_i is the i-th observation.

Sample standard deviation:

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

where n is the number of observations in the sample; X is the sample mean; x_i is the i-th observation.



$$\sum (x_n - \bar{x})^2 =$$

$$7.5625 + 10.5625 + 18.0625 + 3.0625 + 7.5625 + 5.0625 + 14.0625 + 1.5625 = 67.5$$

Variance

$$\frac{\sum(x_n-\bar{x})^2}{N} = \frac{67.5}{8} = 8.45 \text{ points}^2$$

Standard deviation

$$\frac{\sum(x_n-\bar{x})^2}{N}$$

Mean ( ) = 8.45

$$\sqrt{} \; 8.45 \; 2.91 = 2.91$$

Example.
Let us consider the grades of two students:

- Anna: $30, 30, 28, 27, 26$
- Stefano: $21, 30, 30, 30, 30$

Both have the same mean of the grades (28.2)

Let us compute the standard deviation:

- $\sigma_{Anna} = 1.78$
- $\sigma_{Stefano} = 4.02$

What does this mean? Anna's grades are more concentrated (closer) than those of Stefano

Exercise 4 – Quartiles

Recovery time (days) in N = 11 patients:

8, 3, 12, 5, 4, 15, 6, 9, 4, 10, 7

- Calculate the median (Q2), the first quartile (Q1) and the third quartile (Q3).
- Calculate the interquartile range and the range.
- Between which values does the central 50% of patients recover?

Reminders

Q1: 25% of the data ≤ Q1

Q2 = median

Q3: 75% of the data ≤ Q3

Method: Q1 and Q3 are the medians of the lower half and of the upper half of the sorted data (the median is excluded if n is odd).

interquartile range = Q3 – Q1

Exercise 4 – Quartiles: Solution

Position	1	2	3	4	5	6	7	8	9	10	11
Sorted value	3	4	**4**	5	6	**7**	8	9	**10**	12	15

Step 1. Sort the data (table above).

Step 2. Median. N = 11 (odd) → position (11 + 1)/2 = 6 → Q2 = **7 days**

Step 3. Q1. Lower half (5 values below the median): 3 · 4 · **4** · 5 · 6 → the 3rd value → Q1 = **4 days**

Step 4. Q3. Upper half (5 values above the median): 8 · 9 · **10** · 12 · 15 → the 3rd value → Q3 = **10 days**

Step 5. Interquartile range = Q3 − Q1 = 10 − 4 = **6 days**

Step 6. Range = $x_{max} - x_{min}$ = 15 − 3 = **12 days**

Interpretation. The central 50% of patients recover between 4 and 10 days; 25% recover in 4 days or less and 25% in 10 days or more.

Exercise 5 – Variability: Variance and Standard Deviation

Resting heart rate (bpm) of n = 5 volunteers:

62, 70, 66, 74, 68

- Calculate the range and the mean.
- Calculate the sample variance and standard deviation.
- If these 5 values were the entire population of interest, what would the variance and standard deviation be?

Reminders

$C = x_{max} - x_{min}$

$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$

$s = \sqrt{s^2}$

$\sigma^2 = \Sigma (x_i - \mu)^2 / N$

$\sigma = \sqrt{\sigma^2}$

Exercise 5 – Variability: Solution

Range. C = 74 − 62 = **12 bpm**

Mean. $\bar{x}$ = (62 + 70 + 66 + 74 + 68) / 5 = 340 / 5 = **68 bpm**

Deviations from the mean and their squares (table): the sum of the deviations is 0 (check!), the sum of the squares is 80.

Sample (n − 1 = 4): s^2 = 80 / 4 = **20 bpm²**; s = √20 = **4.47 bpm**

Population (N = 5): σ^2 = 80 / 5 = **16 bpm²**; σ = √16 = **4 bpm**

Dividing by n − 1 instead of N gives a slightly larger value; for large n the difference is minimal.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
62	−6	36
70	+2	4
66	−2	4
74	+6	36
68	0	0
Σ	**0**	**80**

Correlation

- It is a more rigorous approach that allows the study the degree of intensity of the linear relationship between pairs of variables

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

Pearson correlation coefficient

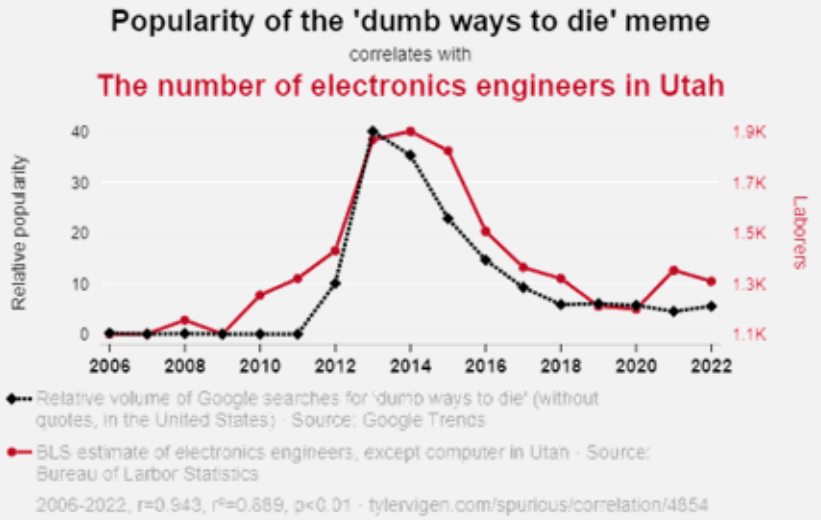
The correlation coefficient allows us to:

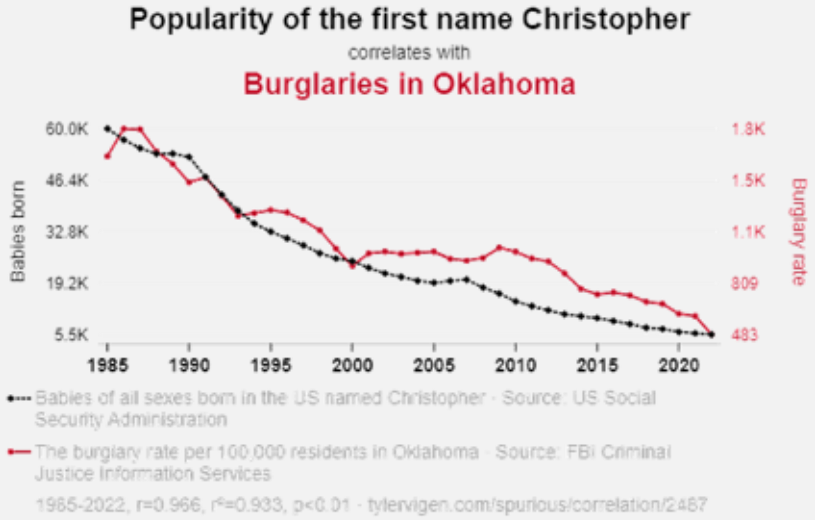
- summarize the strength of the linear relationship between the variables
- assess the apparent association between the variables

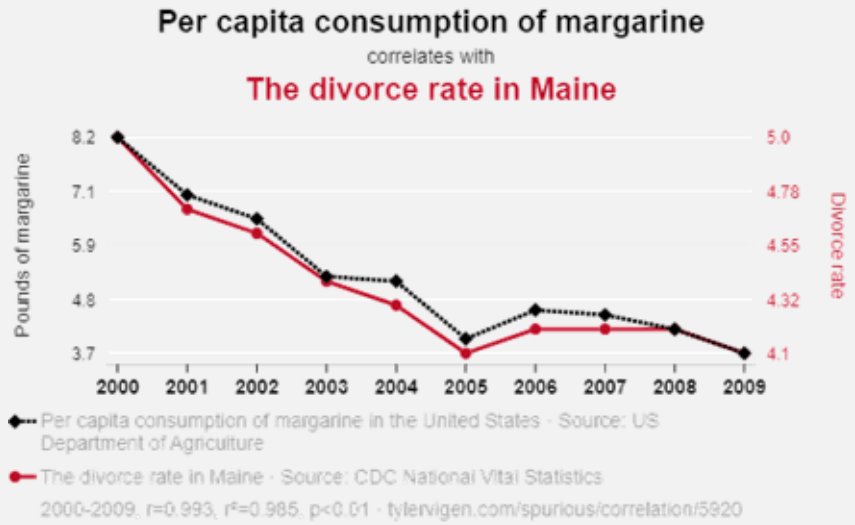
The correlation coefficient:

- ranges from -1 to 1 (if equal to 1 or -1: perfectly correlated)
- is positive when the values of the variables increase together
- is negative when the values of one variable increase as the values of the other decrease
- is not influenced by the units of measurement

CORRELATION IS NOT CAUSATION







Exercise 6 – Pearson Correlation

In 5 hypertensive patients, increasing doses of a drug (x) are administered and the reduction in systolic blood pressure (y) is measured:

Patient	1	2	3	4	5
Dose x (mg)	10	20	30	40	50
Reduction y (mmHg)	4	8	9	10	14

Calculate the Pearson correlation coefficient r and comment on it.

Reminders

$r = \Sigma(dx \cdot dy) / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)}$

with $dx = x - \bar{x}$ and $dy = y - \bar{y}$

$-1 \leq r \leq 1$

Exercise 6 – Pearson Correlation: Solution

Step 1. Means: $\bar{x}$ = 150 / 5 = **30 mg**; $\bar{y}$ = 45 / 5 = **9 mmHg**

Step 2. Deviations from the means, products and squares:

x	y	dx = x − 30	dy = y − 9	dx · dy	dx²	dy²
10	4	−20	−5	100	400	25
20	8	−10	−1	10	100	1
30	9	0	0	0	0	0
40	10	+10	+1	10	100	1
50	14	+20	+5	100	400	25
Σ		**0**	**0**	**220**	**1000**	**52**

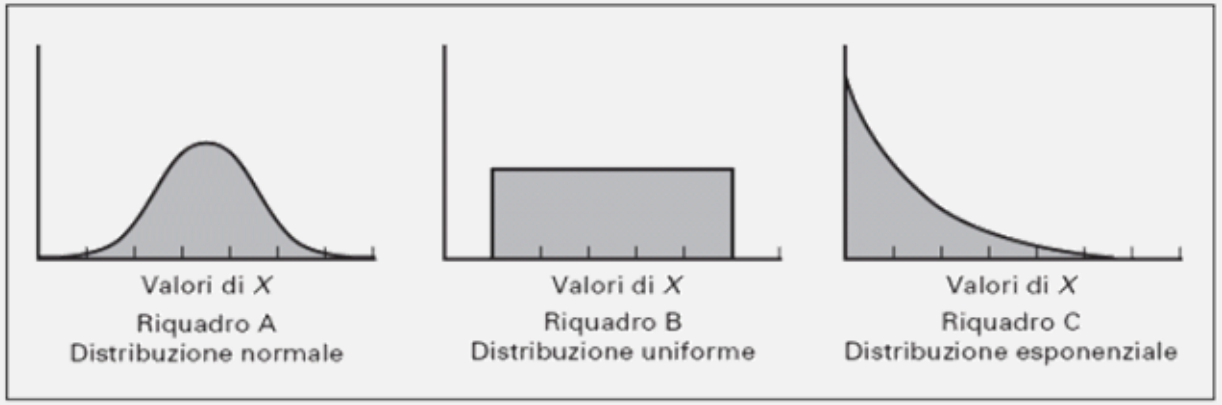
Step 3. r = 220 / √(1000 · 52) = 220 / √52000 = 220 / 228.04 = **0.96**

Step 4. Comment. r is close to +1: a very strong positive linear correlation (higher doses correspond to larger reductions). r does not depend on the units of measurement. **Warning:** correlation does not imply causation.

$$r = \Sigma(dx \cdot dy) / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)}$$

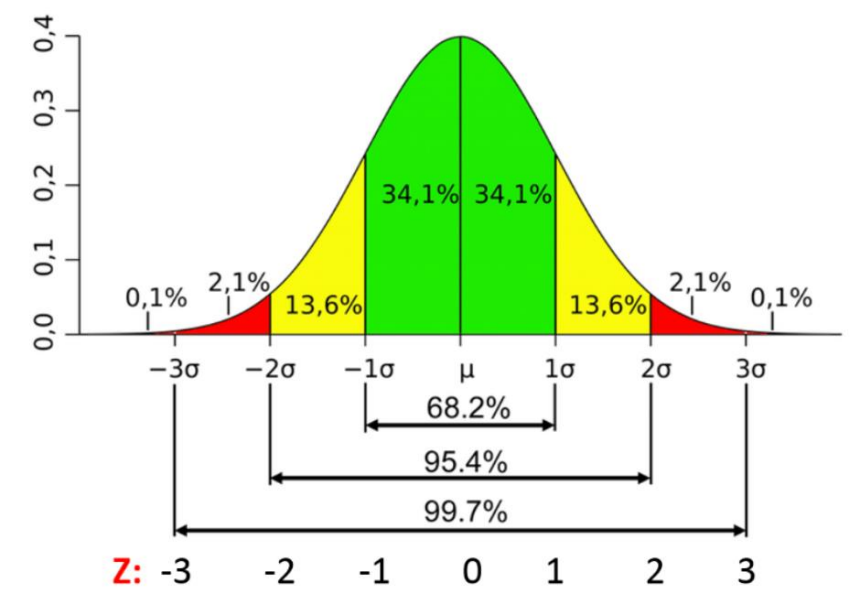
Continuous Probability Distributions

- A continuous probability density function is a model that analytically defines how the values assumed by a continuous random variable are distributed
- When one has a mathematical expression suitable for representing a continuous phenomenon, we are able to calculate the probability that the random variable takes values lying within intervals
- The figure graphically depicts three probability density functions: normal, uniform, and exponential

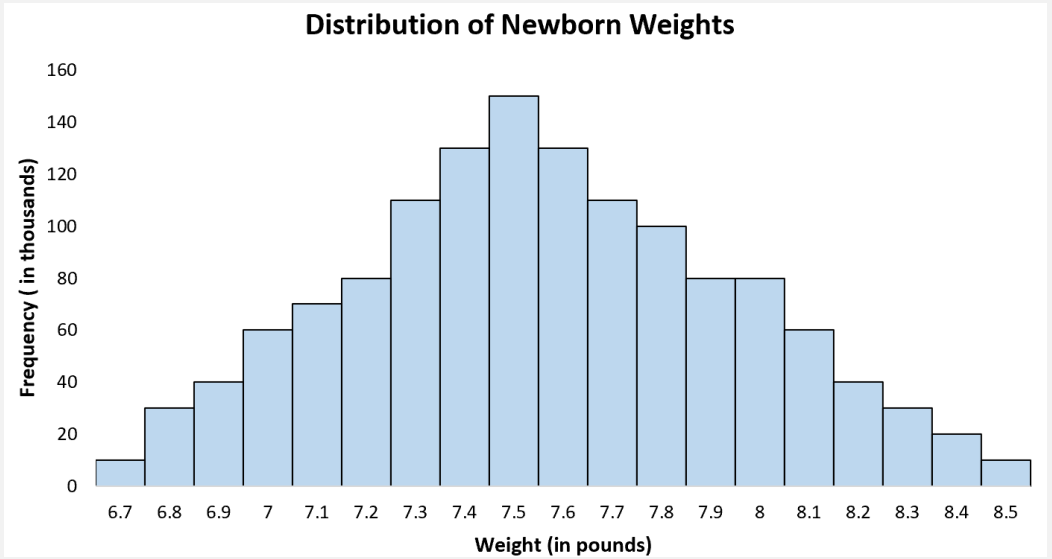


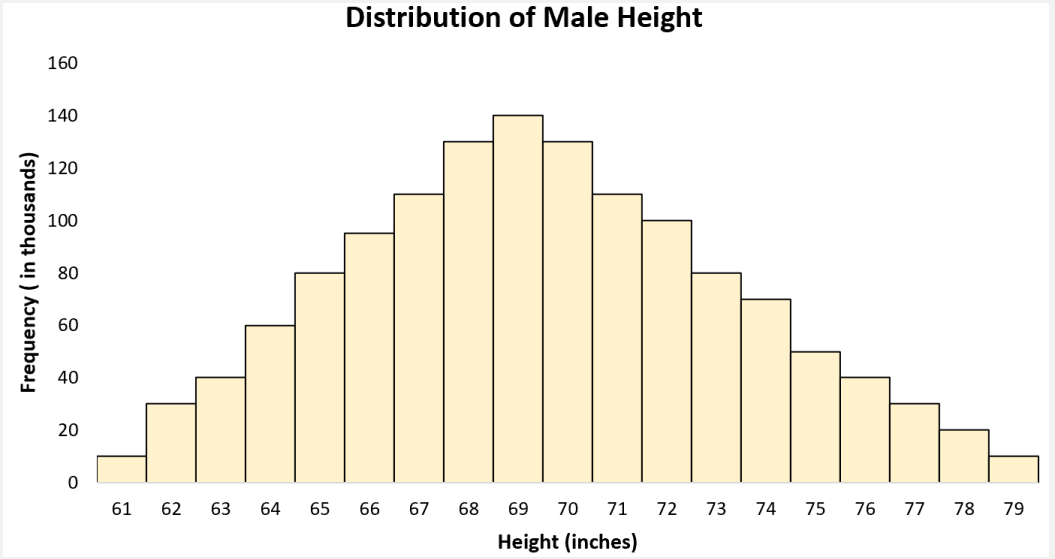
Normal Distribution

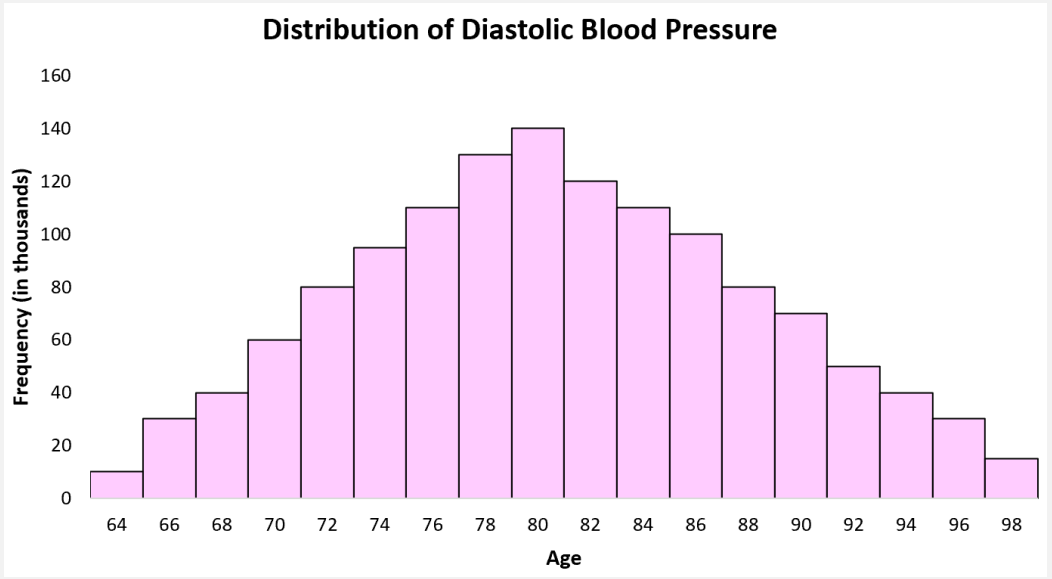
- The normal distribution (or Gaussian distribution) is the most widely used continuous distribution in statistics.
- The normal distribution is important in statistics for three fundamental reasons:
  1. Several continuous phenomena appear to follow, at least approximately, a normal distribution.
  2. The normal distribution can be used to approximate numerous discrete probability distributions.

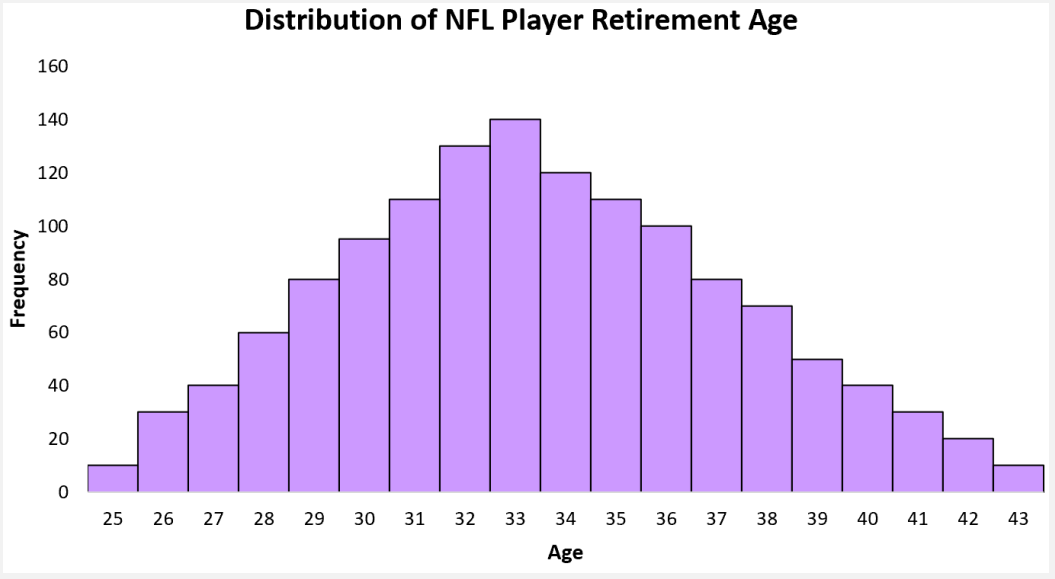





Distribution of Newborn Weights
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
6.7
6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
8.3
8.4
8.5
Weight (in pounds)


Distribution of Male Height
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
Height (inches)


Distribution of Diastolic Blood Pressure
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
Age


Distribution of NFL Player Retirement Age
Frequency
0
20
40
60
80
100
120
140
160
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
Age

Normal Distribution

The normal distribution has several important characteristics:

- The normal distribution has a bell-shaped and symmetric form
- Its measures of central tendency (expected value, median) coincide
- Its interquartile range equals 1.33 times the standard deviation; that is, it covers an interval between $\mu - 2/3\,\sigma$ and $\mu + 2/3\,\sigma$
- The normally distributed random variable takes values between $-\infty$ and $+\infty$

Normal Distribution

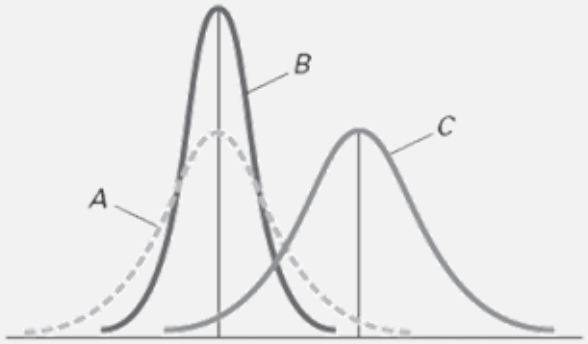
We will use the symbol $f(X)$ to denote the mathematical expression of a probability density function. In the case of the normal distribution, the normal probability density function is given by the following expression:

Normal probability density function

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2}$$

Where μ is the expected value of the population; σ is the standard deviation of the population; X represents the values assumed by the random variable, $-\infty < X < +\infty$

We note that, since e e π are mathematical constants, the probabilities of a normal distribution depend solely on the values assumed by the two parameters μ and σ. By specifying particular combinations of μ and σ, we obtain different normal probability distributions.



Example – Reference Intervals from the Normal Distribution

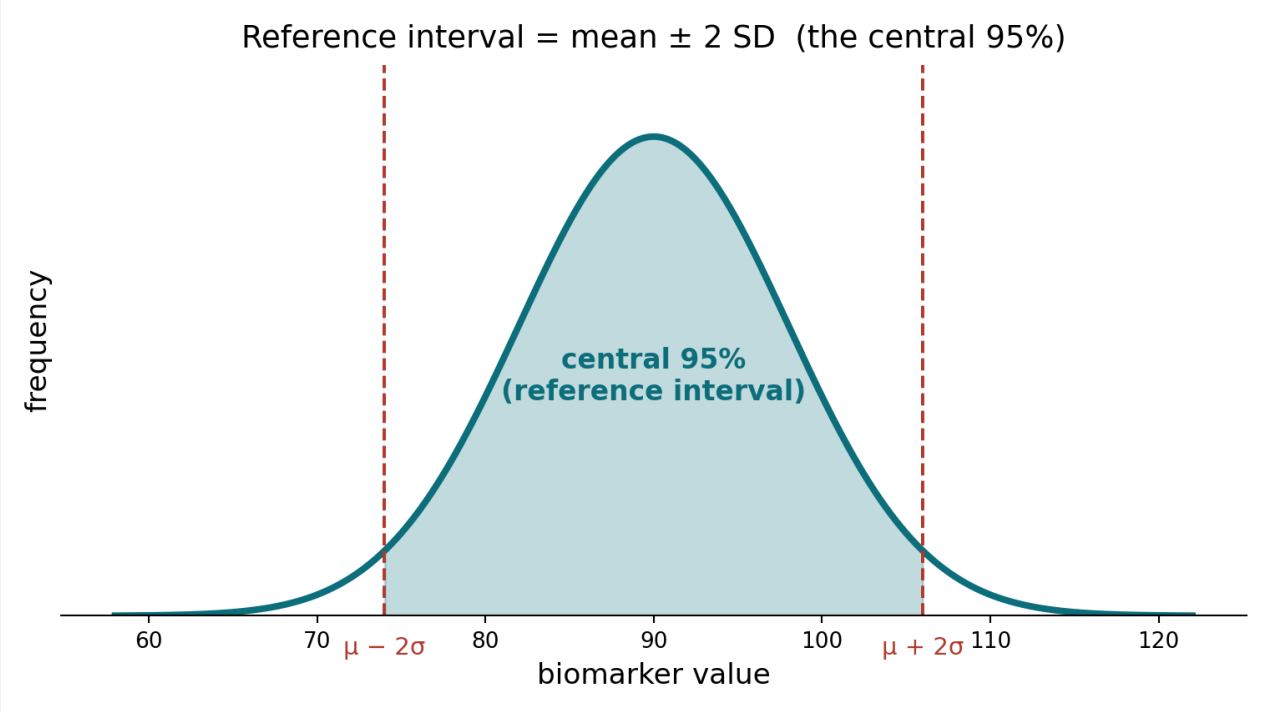
Many biomedical variables are approximately normally distributed around a mean.

The central 95% of the distribution defines the reference (normal) interval:

reference interval ≈ mean ± 2 SD

A value outside it is flagged 'abnormal' — yet ~5% of healthy people fall outside by chance.

This is exactly how laboratory reference ranges are built.



Interval	Share of people inside
μ ± 1σ	about 68%
μ ± 2σ	about 95%
μ ± 3σ	about 99.7%

Exercise 7 – Normal Distribution

Haemoglobin in healthy women is approximately normally distributed, with mean μ = 13.5 g/dL and standard deviation σ = 1.0 g/dL.

- Calculate the reference interval (central 95%).

- Three patients have haemoglobin A = 10.9; B = 14.7; C = 15.8 g/dL. Calculate the **z-score** of each and say whether it falls within the reference interval.

- Out of 400 healthy women, how many do we expect outside the interval?

- Which interval contains about 68% of healthy women?

Reminders

reference interval ≈ μ ± 2σ

μ ± σ ≈ 68%; μ ± 2σ ≈ 95%; μ ± 3σ ≈ 99.7%

z = (x − μ) / σ (number of SDs from the mean)

Exercise 7 – Normal Distribution: Solution

a) $\mu \pm 2\sigma = 13.5 \pm 2 \cdot 1.0 \rightarrow$ **11.5 – 15.5 g/dL**

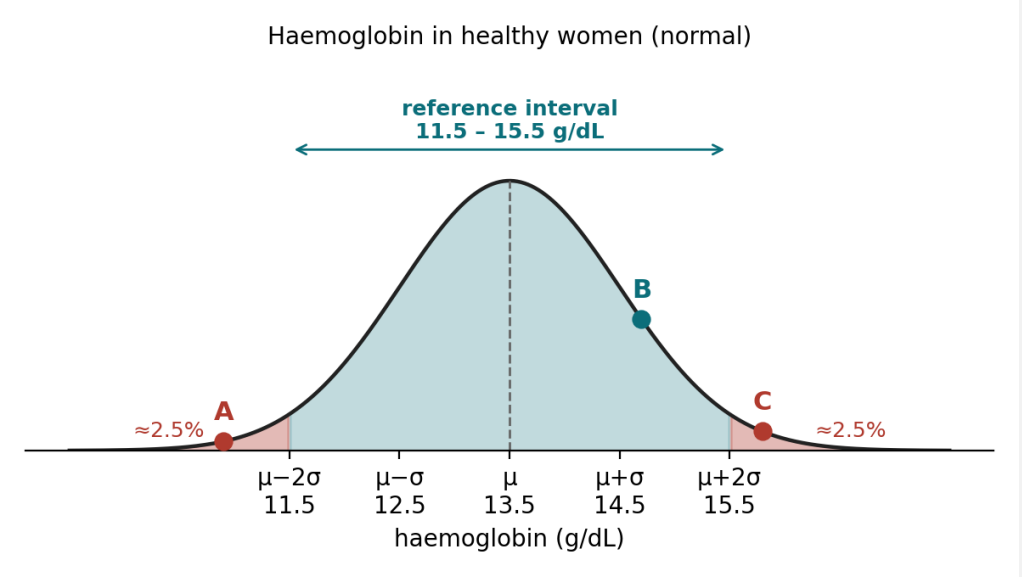
b) $z = (x - \mu) / \sigma$

- A: $z = (10.9 - 13.5) / 1.0 =$ **−2.6** → $|z| > 2$: **outside** (below)
- B: $z = (14.7 - 13.5) / 1.0 =$ **+1.2** → **inside**
- C: $z = (15.8 - 13.5) / 1.0 =$ **+2.3** → **outside** (above)

c) about 5% lie outside: $0.05 \cdot 400 =$ **20 women** (about 10 per tail).

d) $\mu \pm \sigma = 13.5 \pm 1.0 \rightarrow$ **12.5 – 14.5 g/dL**

Being outside the interval does not mean being ill: about 5% of healthy people fall outside by pure chance.



reference interval ≈ $\mu \pm 2\sigma$

$\mu \pm \sigma \approx 68\%$; $\mu \pm 2\sigma \approx 95\%$; $\mu \pm 3\sigma \approx 99.7\%$

$z = (x - \mu) / \sigma$ (number of SDs from the mean)

From the Sample to the Population: the Standard Error

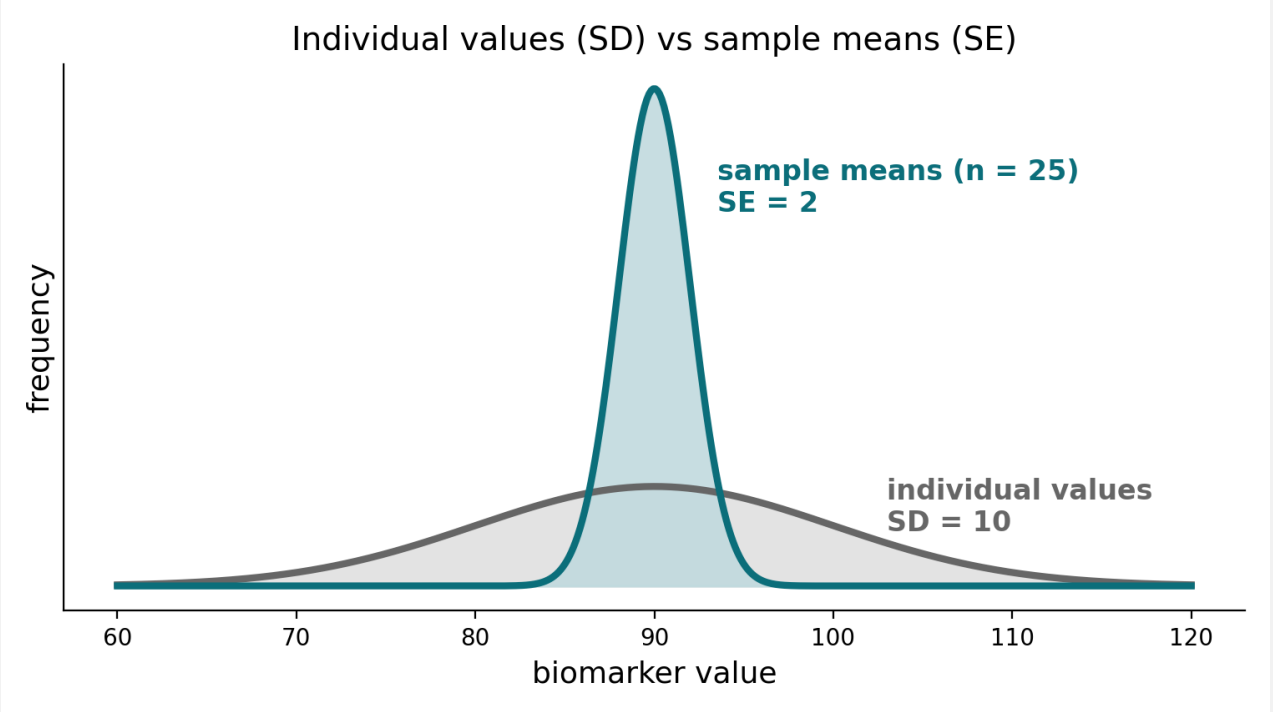
A sample mean is only an estimate of the true (population) mean: another sample would give a slightly different mean.

The spread of the sample means around the true mean is the standard error of the mean (SE):

$$\mathrm{SE} = \mathrm{SD}/\sqrt{n}$$

SD describes the variability of the individual values; SE describes how precisely the mean is estimated.

SD does not shrink as the sample grows; SE does. To halve the SE you need four times as many subjects.



Example – The Standard Error of a Glucose Measurement

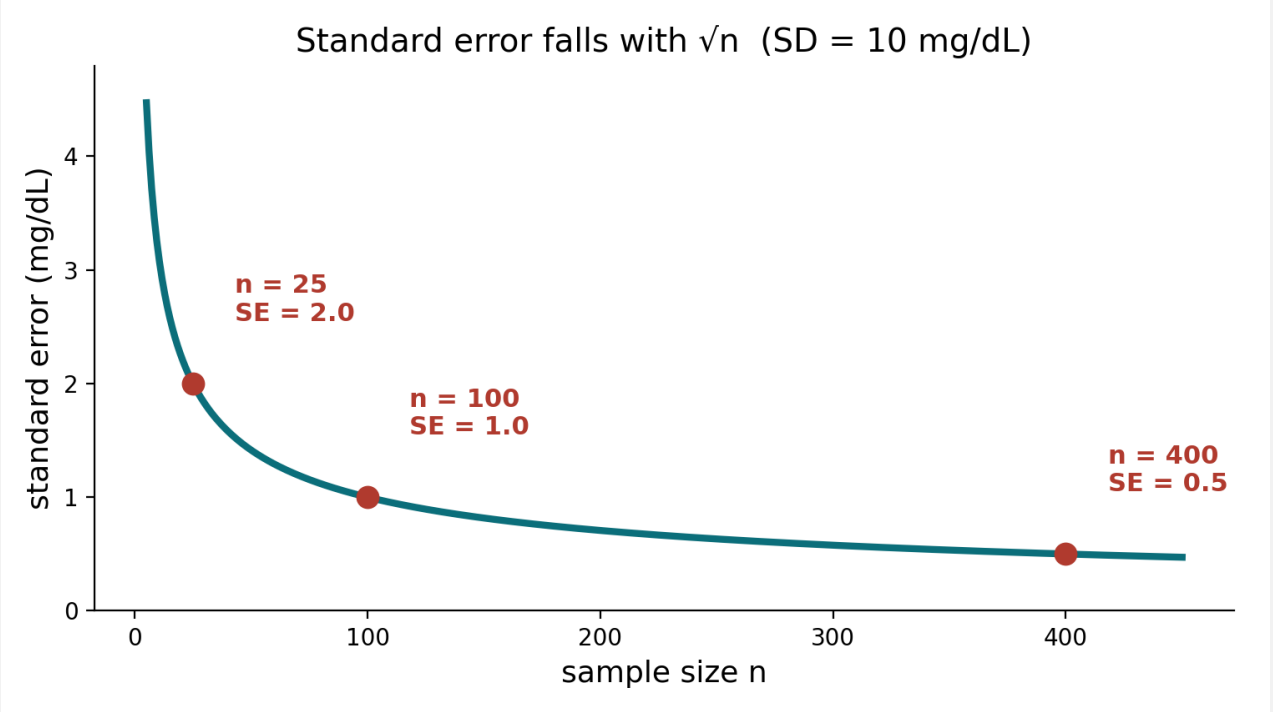
Fasting glucose measured in n = 25 volunteers: mean = 92 mg/dL, SD = 10 mg/dL.

$$ES = 10/\sqrt{25} = 2mg/dL$$

With n = 100 subjects: SE = 10 / √100 = 1 mg/dL. With n = 400: SE = 0.5 mg/dL.

Quadrupling the sample size only halves the SE: precision grows with √n.

The SE is the denominator of the t-test: t = (mean₁ − mean₂) / standard error.



The Confidence Interval of the Mean

Instead of a single number, we report a range of plausible values for the true mean.

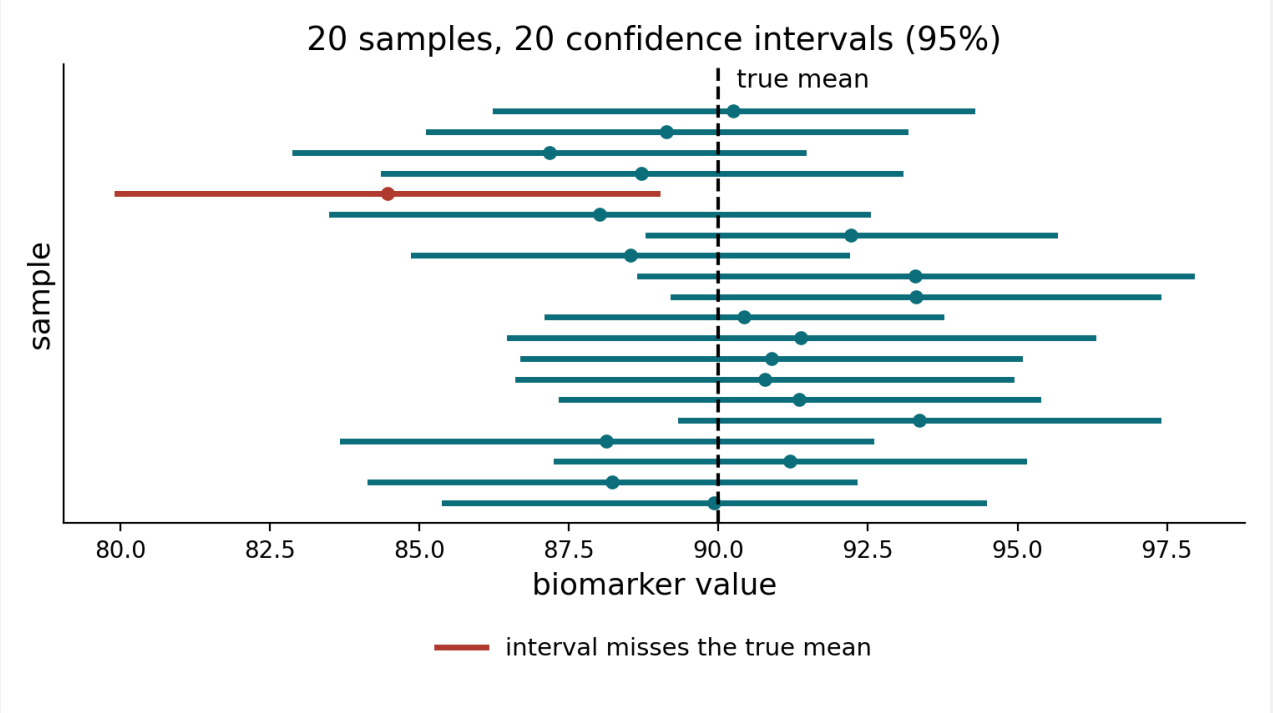
95% confidence interval:

95% CI ≈ mean ± 2 · SE

(exactly mean ± t · SE, with t ≈ 1.96 for large samples and larger for small ones)

If we repeated the study many times, 95% of the intervals built this way would contain the true mean.

Narrow interval = precise estimate. Do not confuse it with the reference interval (mean ± 2 SD), which describes individual values.



Example – 95% Confidence Interval of Glucose

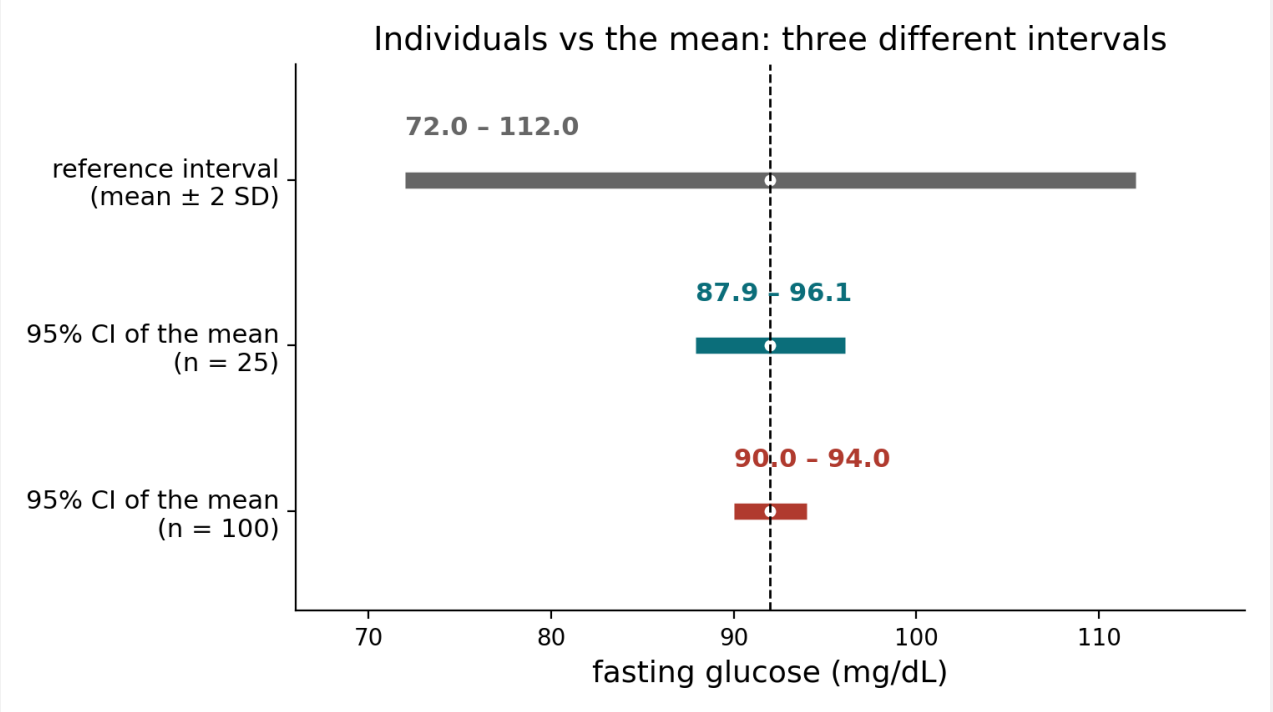
n = 25, mean = 92 mg/dL, SD = 10 mg/dL → SE = 2 mg/dL. For n = 25, t = 2.06.

95% CI = 92 ± 2.06 · 2 = 87.9 – 96.1 mg/dL

We are 95% confident that the true mean glucose of the population lies between about 88 and 96 mg/dL.

With n = 100 (SE = 1 mg/dL) the interval shrinks to 90.0 – 94.0 mg/dL.

The reference interval (mean ± 2 SD = 72 – 112 mg/dL) is much wider: it describes individuals, not the mean.



The *p*-Value

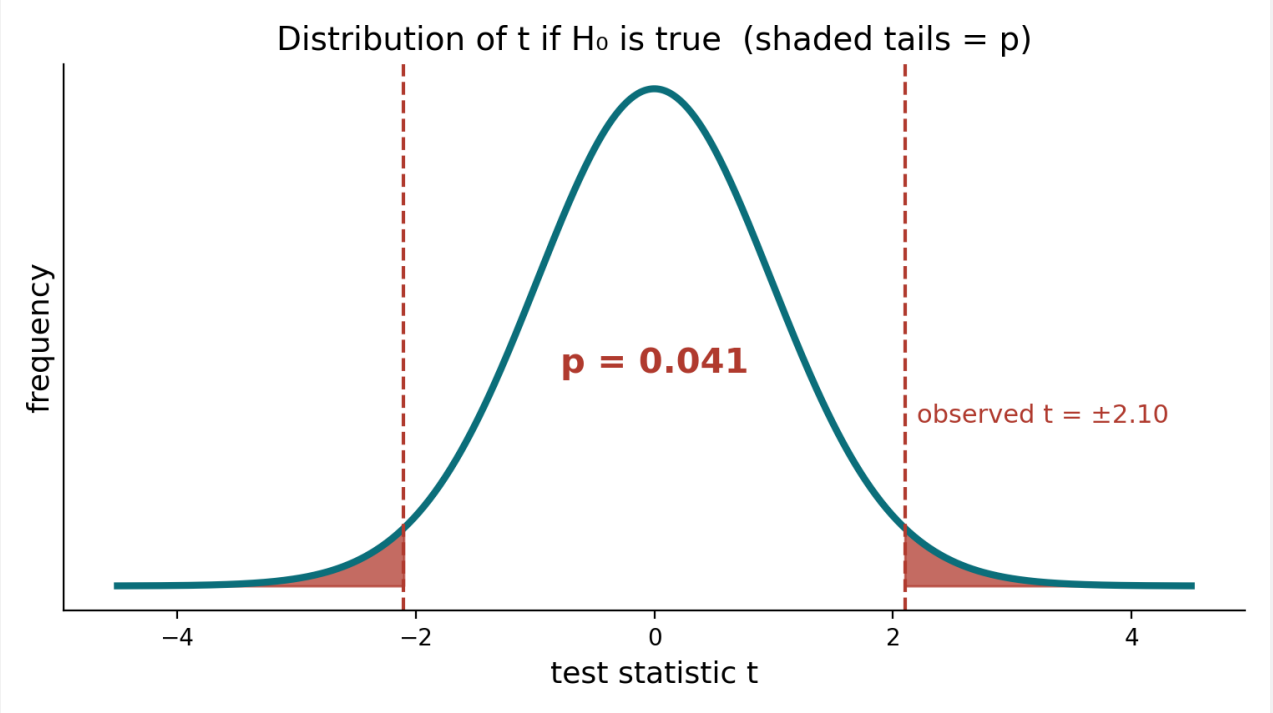
Start from the null hypothesis H_0: there is no real effect (e.g. treated and control have the same mean).

The p-value is the probability of observing a difference at least as large as ours if H_0 were true.

$p < 0.05 \rightarrow$ statistically significant

Small p (by convention $p < 0.05$): data hard to explain by chance alone, so we reject H_0.

Large p: data compatible with chance; we cannot reject H_0 (this does not prove that there is no effect).



Reading the *p*-Value Correctly

p is NOT the probability that H_0 is true, and 0.05 is a convention, not a law of nature: p = 0.04 and p = 0.06 tell nearly the same story.

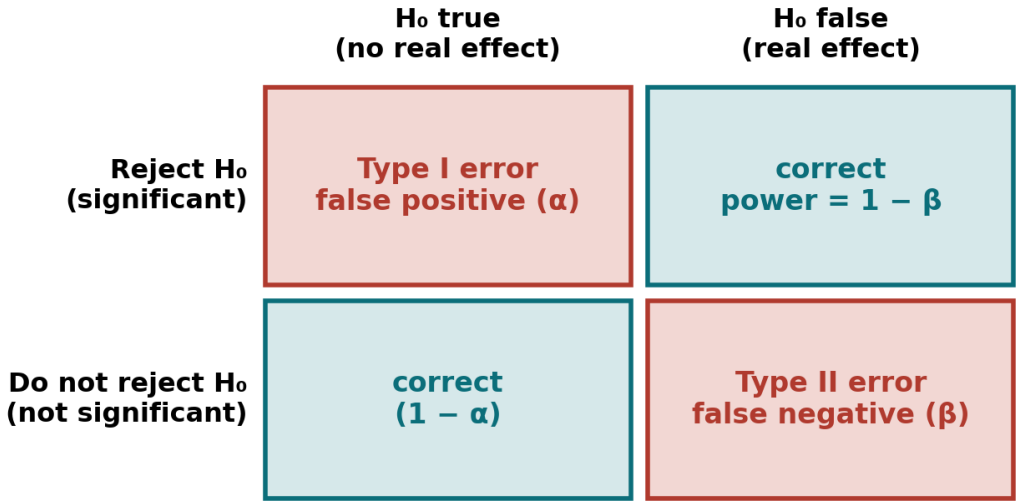
Significant ≠ important.

With very large samples a tiny difference can give p < 0.001: always report the size of the effect and its confidence interval.

p and CI go together: if the 95% CI of the difference does not contain 0, then p < 0.05.

Two ways of being wrong: Type I error (false positive, probability α = 0.05) and Type II error (false negative, probability β).

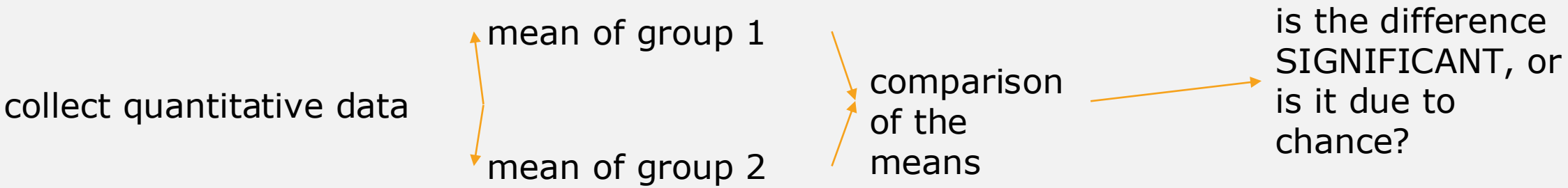
The two ways of being wrong



Comparing Two Means: the Student *t* Test

<u>is the difference between the means of the two samples significant?</u>

Can you state that the observed difference is not due to chance but that, instead, there is genuinely a difference between the means of the two populations?



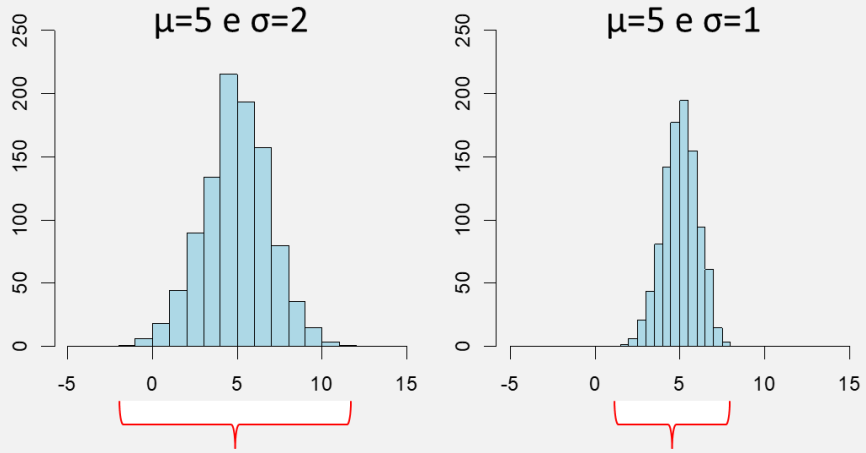


Comparing Two Means: the Student *t* Test

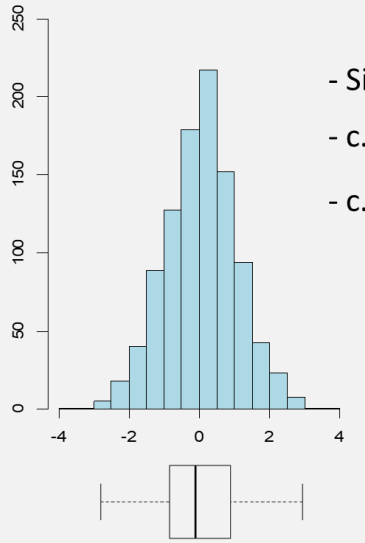
comparison of the means → ZERO HYPOTHESIS: the difference is due to chance! → accept or reject the zero hypothesis? → Student's *t*-test

Conditions:

1. Independence of observations
2. Normality of the populations being compared
3. Homogeneity of variance



Analisi dell'istogramma



- Simmetria (media ≈mediana)
- c. 2/3 dei dati in un intervallo μ±σ
- c. 95% dei dati in un intervallo μ±2σ

Comparing Two Means: the Student *t* Test

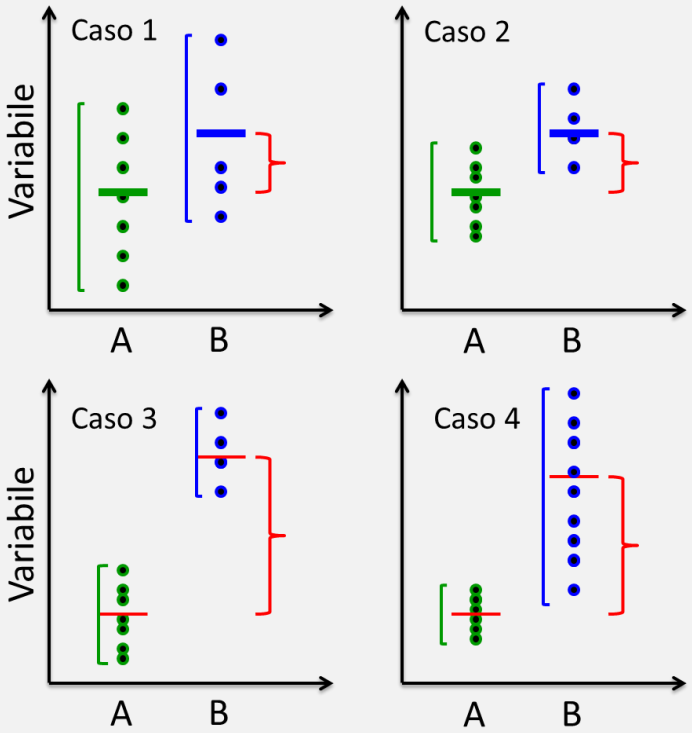
difference between means

$$t = \frac{m_a - m_b}{S} \sqrt{\frac{n_a\, n_b}{n_a + n_b}}$$

dimensional factor

standard error

t = (mean₁ − mean₂) / standard error



difference between means

variability A

variability B

Example – Comparing Two Groups with the Student *t* Test

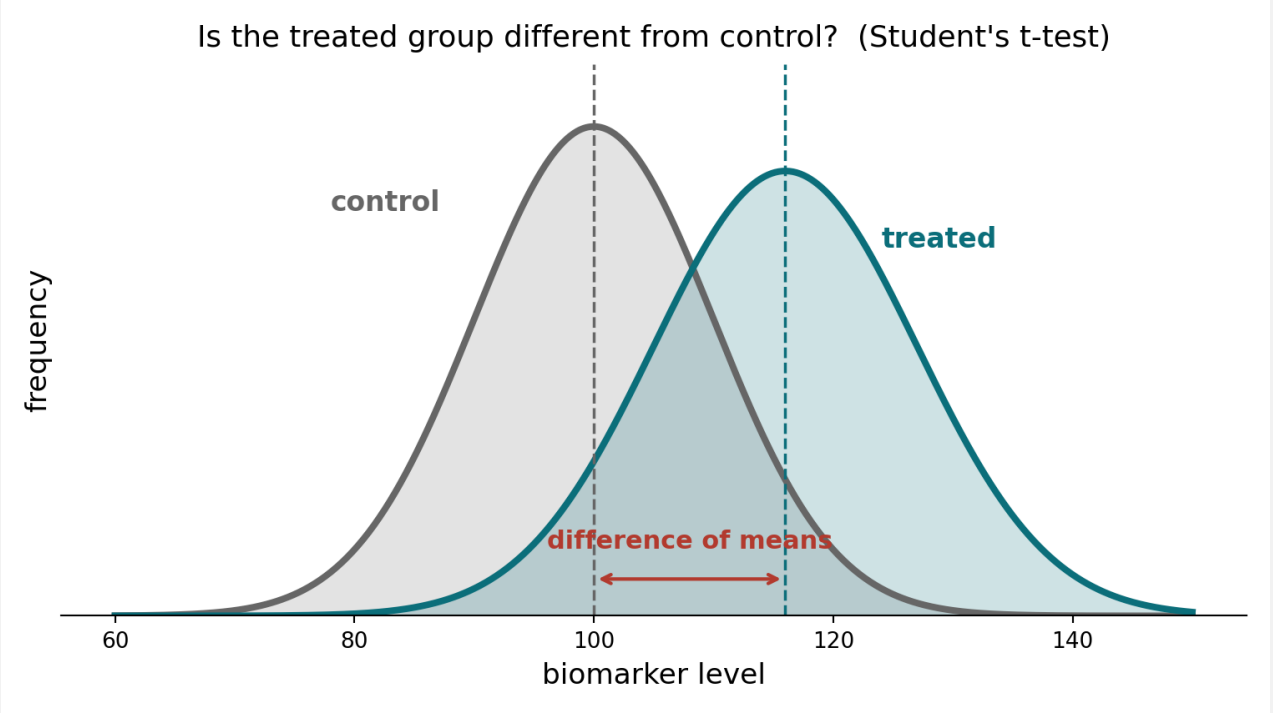
Does a treatment change a biomarker? Compare the treated group with a control group.

The t-test weighs the difference between the two means against the spread within the groups:

$$t = (\text{mean}_1 - \text{mean}_2) / \text{standard error}$$

A small p-value means the observed difference is unlikely to be due to chance alone.

This is the logic behind most 'treated vs control' experiments.



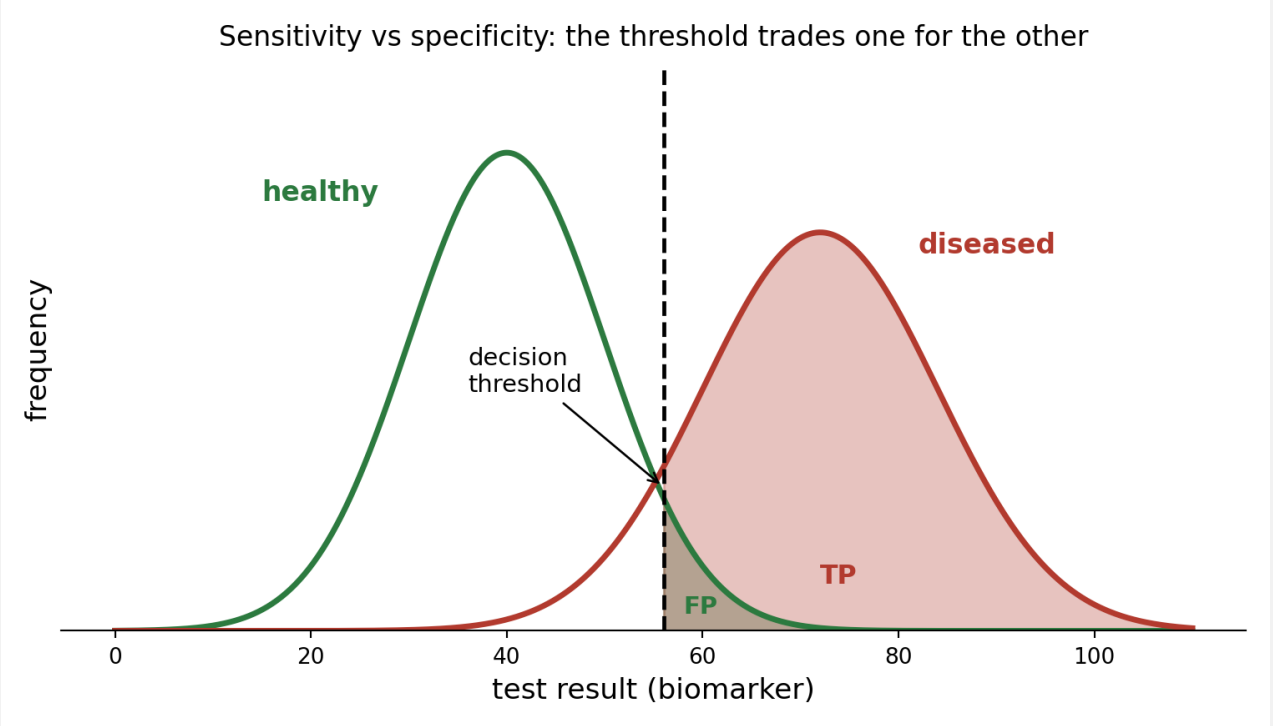
Diagnostic Accuracy — Sensitivity and Specificity

A diagnostic test must separate diseased from healthy people, but their values overlap.

Sensitivity = TP / (TP + FN): the fraction of diseased people correctly detected.

Specificity = TN / (TN + FP): the fraction of healthy people correctly cleared.

Moving the decision threshold trades sensitivity against specificity.



Exercise 10 – Diagnostic Tests

A rapid test is applied to 1000 people. The prevalence of the disease is 10%: 100 diseased and 900 healthy. The test is positive in 90 of the 100 diseased and in 90 of the 900 healthy.

- Build the 2 × 2 table (TP, FN, FP, TN).
- Calculate sensitivity and specificity.

Reminders

Se = TP / (TP + FN)

Sp = TN / (TN + FP)

Exercise 10 – Diagnostic Tests: Solution (a – c)

	Diseased	Healthy	Total
Positive test	**TP = 90**	FP = 90	180
Negative test	FN = 10	**TN = 810**	820
Total	100	900	1000

a) FN = 100 – 90 = 10 (diseased with a negative test); TN = 900 – 90 = 810 (healthy with a negative test).

b) Se = TP / (TP + FN) = 90 / 100 = **0.90 = 90%**

Sp = TN / (TN + FP) = 810 / 900 = **0.90 = 90%**

MODULE MATH7 — BIOSTATISTICS

Statistics Exercises

Ten exercises with step-by-step worked calculations (a pocket calculator is enough). Each exercise has one slide with the problem and one with the solution.

No.	Topic	Indices used
1	Arithmetic and weighted mean	sample mean; weighted mean
2	Frequencies and mode	absolute, relative and percentage frequency; mode
3	Median and robustness	median (odd and even n); effect of outliers
4	Quartiles	Q1, Q2, Q3; interquartile range; range
5	Variability	variance and standard deviation (sample and population)
6	Correlation	Pearson correlation coefficient
7	Normal distribution	z-score; reference interval mean ± 2 SD
8	Standard error and confidence interval	SE; 95% CI of the mean; effect of n
9	Comparing two means	Student t-test; p-value; CI of the difference
10	Diagnostic tests	sensitivity, specificity, PPV, NPV; effect of prevalence

MODULE MATH7 — BIOSTATISTICS

Formula Sheet

Index	Formula
Relative frequency and %	f / N $f / N \times 100$ (f = absolute frequency, N = number of data)
Sample mean	$\bar{x} = \Sigma x_i / n$
Weighted mean	$\Sigma (p_i \cdot x_i) / \Sigma p_i$ (p_i = weights)
Median	ordered data: n odd → position $(n + 1)/2$; n even → mean of positions $n/2$ and $n/2 + 1$
Quartiles	Q1, Q2 (= median), Q3; interquartile range = Q3 − Q1
Range	$C = x_{max} - x_{min}$
Sample variance and SD	$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$ $s = \sqrt{s^2}$
Population variance and SD	$\sigma^2 = \Sigma (x_i - \mu)^2 / N$ $\sigma = \sqrt{\sigma^2}$
Pearson correlation	$r = \Sigma(dx \cdot dy) / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)}$, with $dx = x - \bar{x}$, $dy = y - \bar{y}$
z-score; reference interval	$z = (x - \mu) / \sigma$ reference interval $\approx \mu \pm 2\sigma$ (central 95%)
Standard error; 95% CI	$SE = s / \sqrt{n}$ 95% CI $= \bar{x} \pm t \cdot SE$ ($df = n - 1$)
t-test (two samples)	$t = (\bar{x}_1 - \bar{x}_2) / SE_{diff}$, $SE_{diff} = \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$, $df = n_1 + n_2 - 2$
Diagnostic tests	Se = TP/(TP+FN) Sp = TN/(TN+FP) PPV = TP/(TP+FP) NPV = TN/(TN+FN)