# Introduction to Statistical Analysis

Cancer Research UK Cambridge Institute - 7th of November 2024

Luca Porcu and Chandra Chilamakuri (Bioinformatics core)

CANCER RESEARCH UK | Cambridge Institute

UNIVERSITY OF CAMBRIDGE

# Timeline

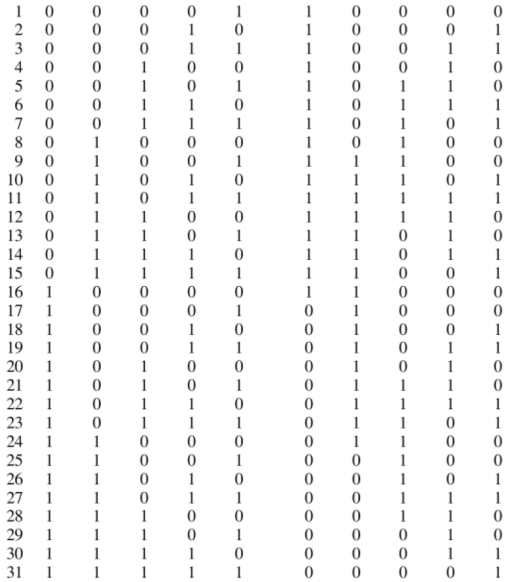
- Morning (9.30-12.00)
  - Data types, data analysis and descriptive statistics → Online quiz
  - Central limit theorem (CLT) → Simulations
  - Inferential statistics: estimation → Simulations
- Lunch
- Afternoon (13.00-17.00)
  - Inferential statistics: one-sample tests → Exercises
  - Inferential statistics: two-sample tests → Exercises

# The Scope of Statistics

1. Study of populations

2. Study of variation

3. Study of methods of the reduction of data



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 4 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 7 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 8 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 10 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 11 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 13 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 14 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| 15 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 16 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 17 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 18 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 20 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 21 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 22 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 23 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 24 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 25 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 26 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 27 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 28 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 29 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 30 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 31 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

R.A. Fisher, Statistical Methods for Research Workers - Introduction, First Edition 1925

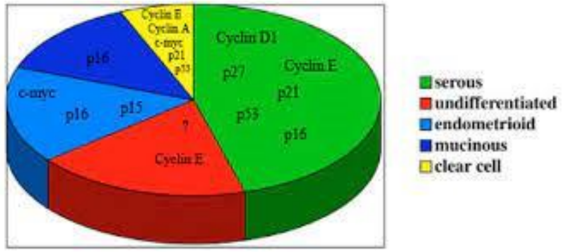
# Data types

c o m p l e x i t y

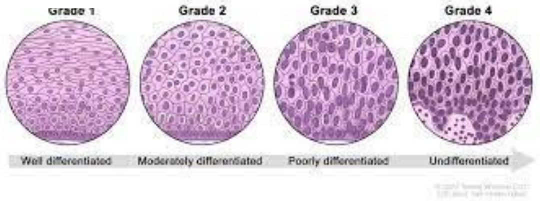
## Qualitative

- Binary/dichotomous

- Nominal

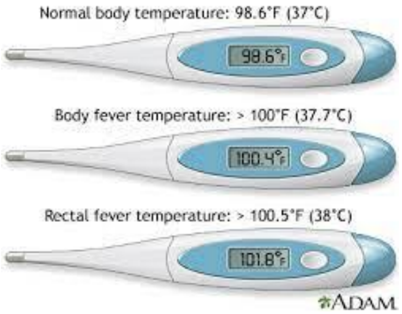


- Ordinal



## Quantitative

- Interval scale (arbitrary zero point)


Normal body temperature: 98.6°F (37°C)
98.6°F
Body fever temperature: > 100°F (37.7°C)
100.4°F
Rectal fever temperature: > 100.5°F (38°C)
101.8°F
ADAM

- Ratio scale (meaningful zero point)

# Data types

complexity

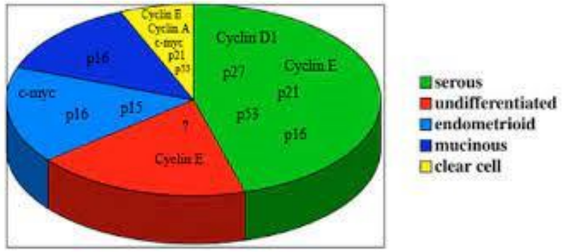
## Qualitative

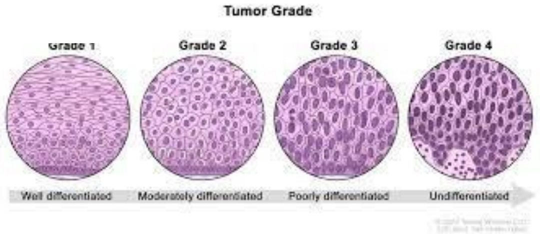
- Binary/dichotomous
- Nominal
- Ordinal

## Quantitative

- Discrete
- Continuous

OPPOSITES
Alive
Dead





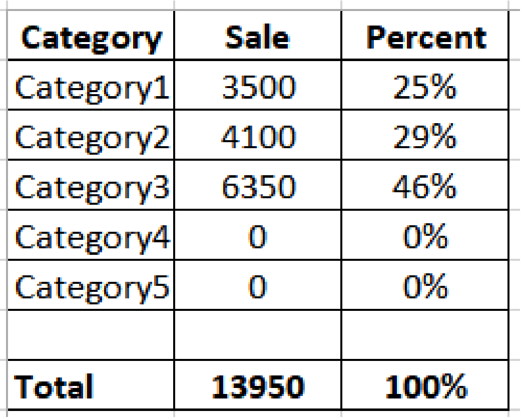
Number of metastases

# Data analysis: descriptive statistics

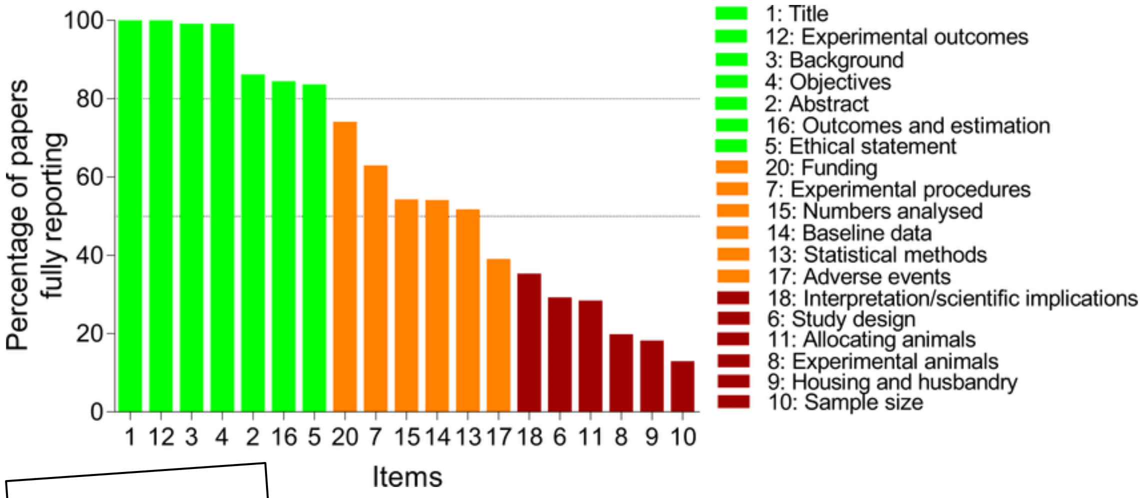
Here, the data are analyzed on their own terms, essentially without extraneous assumptions.

The principal aim is the organization and summarization of the data in ways that bring out their main features and clarify their underlying structure.

Median



| Category | Sale | Percent |
|---|---|---|
| Category1 | 3500 | 25% |
| Category2 | 4100 | 29% |
| Category3 | 6350 | 46% |
| Category4 | 0 | 0% |
| Category5 | 0 | 0% |
| | | |
| **Total** | **13950** | **100%** |



Range

E.L. Lehmann, George Casella, Theory of Point Estimation, Second Edition

# Descriptive statistics in preclinical research

- Baseline data

  (e.g. strain, sex, age, weight, housing)

[†] Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, et al. (2009) Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLoS ONE 4(11): e7824. doi:10.1371/journal.pone.0007824

- Experimental design

  (e.g. sample size, blocking, treatment)

- Outcomes (e.g. distribution,

  length of follow-up, number of events)

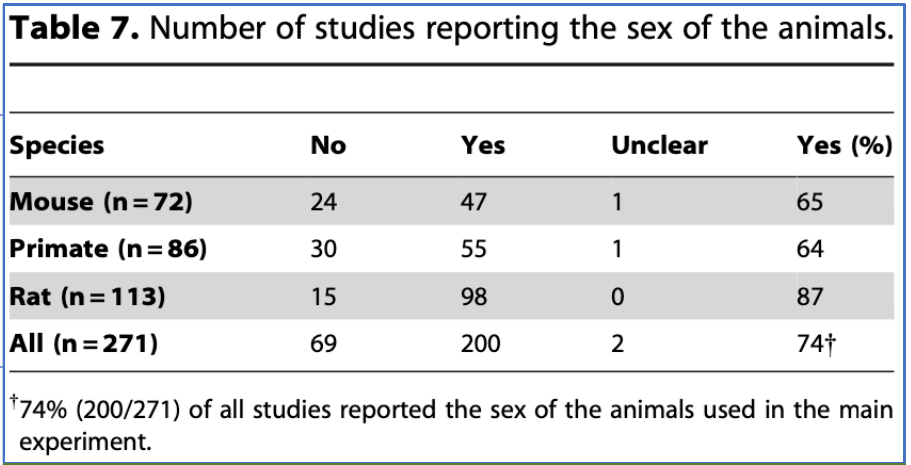
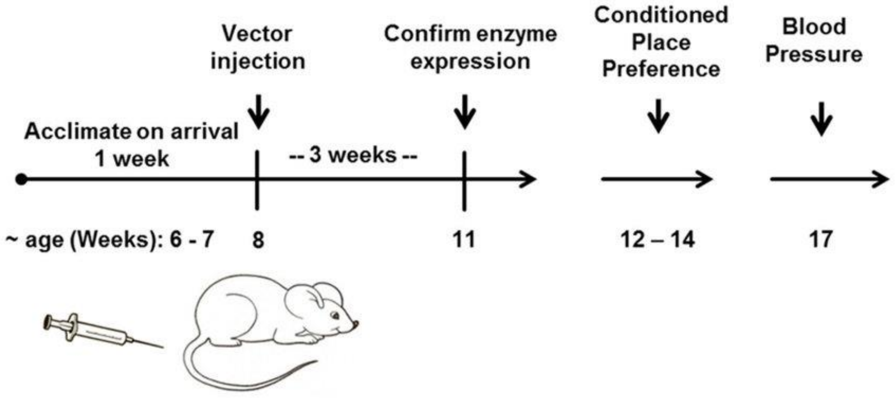
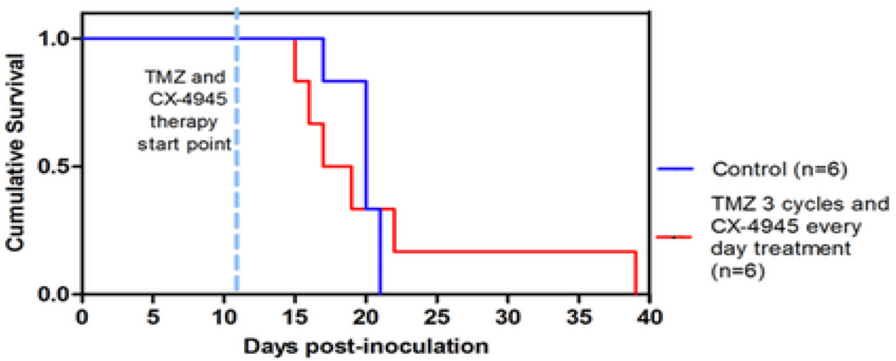


**Table 7.** Number of studies reporting the sex of the animals.

| Species | No | Yes | Unclear | Yes (%) |
|---|---|---|---|---|
| **Mouse (n = 72)** | 24 | 47 | 1 | 65 |
| **Primate (n = 86)** | 30 | 55 | 1 | 64 |
| **Rat (n = 113)** | 15 | 98 | 0 | 87 |
| **All (n = 271)** | 69 | 200 | 2 | 74† |

†74% (200/271) of all studies reported the sex of the animals used in the main experiment.


Vector injection
Confirm enzyme expression
Conditioned Place Preference
Blood Pressure
Acclimate on arrival 1 week
-- 3 weeks --
~ age (Weeks): 6 - 7
8
11
12 – 14
17



# Data analysis: frequentist inference

## Random events.

Empirical phenomena which have the following two features:

1. They do not have deterministic regularity (i.e. observations of them do not always yield the same outcome)
2. They possess some statistical regularity, indicated by the statistical stability of their frequencies.

Shiryaev Albert N., Probability-1, Third Edition

# Data analysis: frequentist inference

## Random events.

“It is our task to detect regularity in the presence of confusion, order in the presence of chaos.”

Edwards A.W.F., Likelihood (Expanded Edition), 1992

# Data analysis: frequentist inference

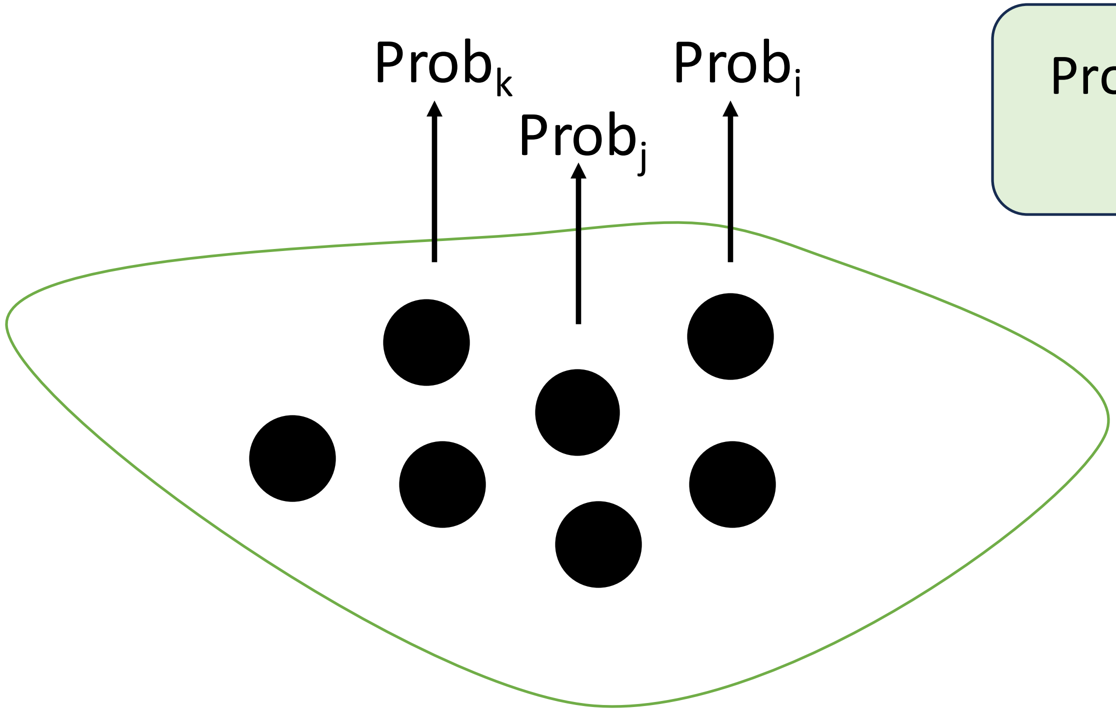
## Mathematical analysis of random events.

- Random events are **more or less adequately** described by statistical distributions (e.g. normal distribution).

- Statistical regularity is captured by **parameters of statistical distributions** (e.g. mean and standard deviation of the normal distribution).

# Probability distribution

**Def:** In probability theory and statistics, a **probability distribution** is the *mathematical function* that gives the *probabilities* of occurrence of different possible *outcomes* for an experiment.



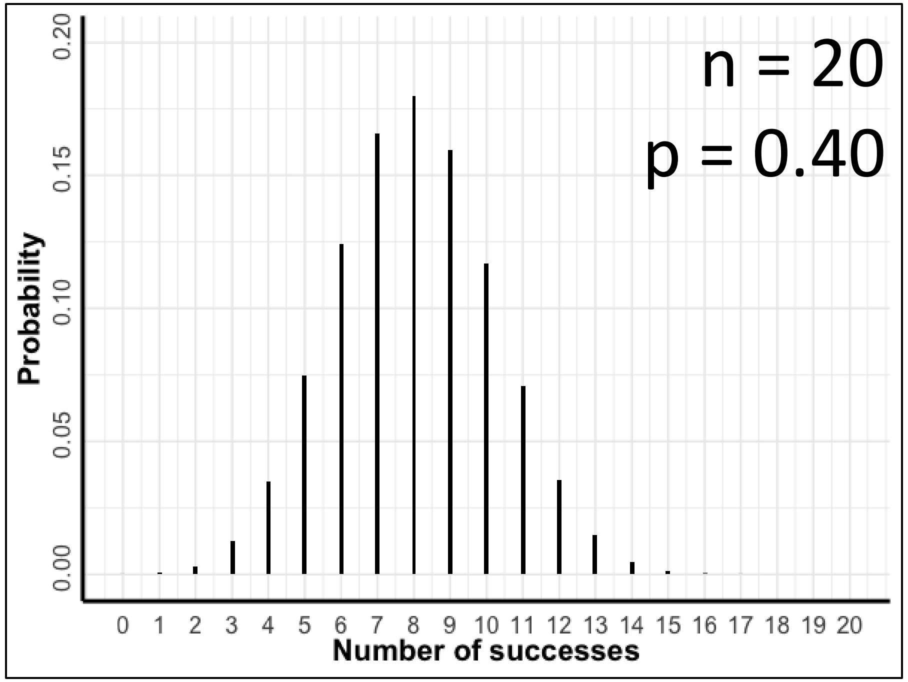
$\text{Prob}_i$ could be interpreted as parameters of a general statistical distribution

● Outcome

https://en.Wikipedia.org/wiki/Probability_distribution, 15th August 2024

# Binomial distribution

**Def:** the **binomial distribution** with parameters *n* and *p* is the discrete statistical distribution of the number of successes in a sequence of *n* independent trials. Each trial (Bernoulli trial) has a binary outcome: success with probability *p* and failure with probability 1-*p*.



**Assumptions of the binomial distribution**

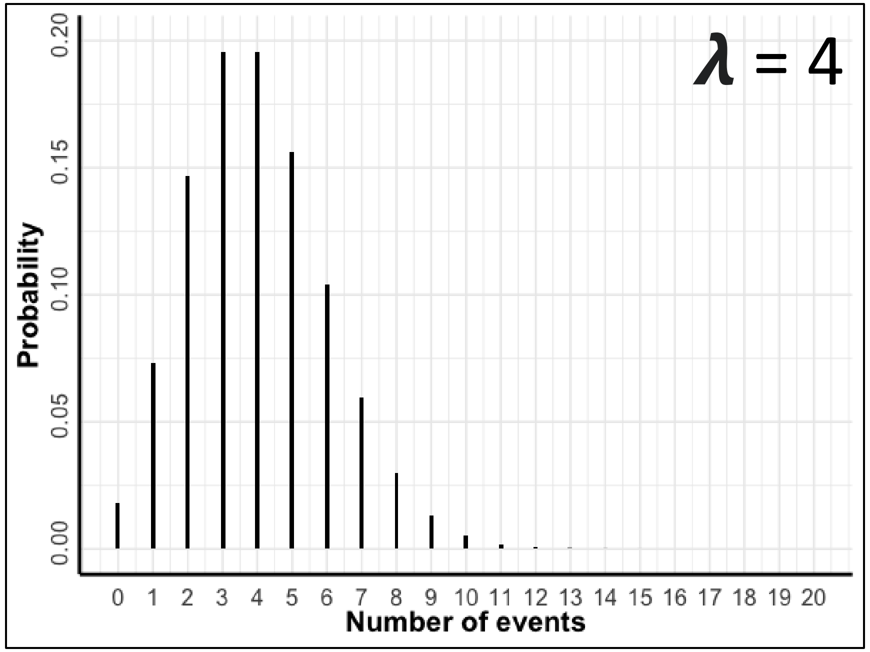
- The outcome of each trial is binary (0/1)

- Each Bernoulli trial is independent (i.e. the outcome of each trial does not depend on the outcome of the other trials)

- The probability of success *p* is constant (i.e. it does not change for each trial)

# Poisson distribution

**Def:** the **Poisson distribution** with parameter $\boldsymbol{\lambda}$ is the discrete probability distribution of the number of events that occur randomly and uniformly in a fixed time interval or in a given area.



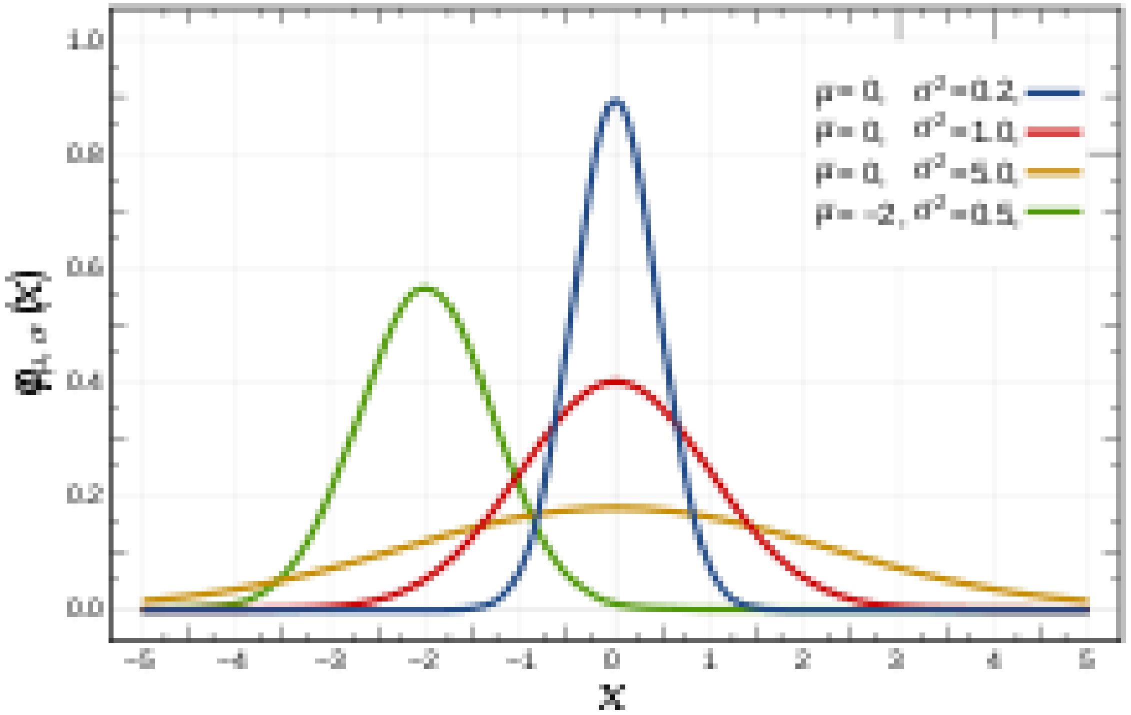
**Assumptions of the Poisson distribution**

- The outcome is a count [0,1,…,k,…]
- Independence of events: the occurrence of one event does not affect the probability that another event will occur
- Two events cannot occur at exactly the same instant in time or at the same point of the given area
- Events occur at a uniform rate over the entire time period or area. $\boldsymbol{\lambda}$ is the expected (mean) number of events per time/area unit

# Normal distribution

**Def:** a **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution. It is determined by two parameters (**μ, σ**).

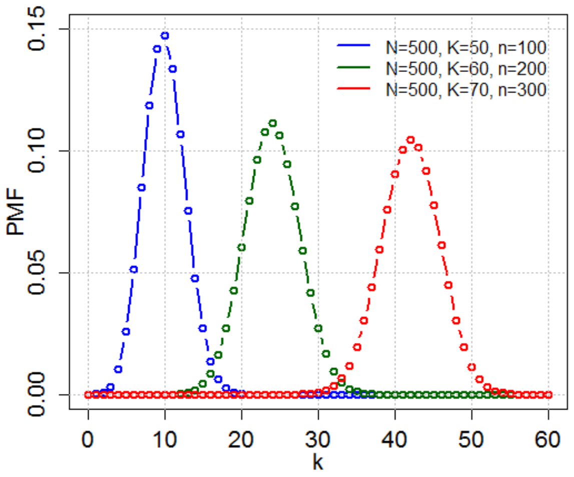
1) the parameter **μ** is the mean or expectation of the distribution (and also its median and mode)
2) the parameter **σ** is its standard deviation



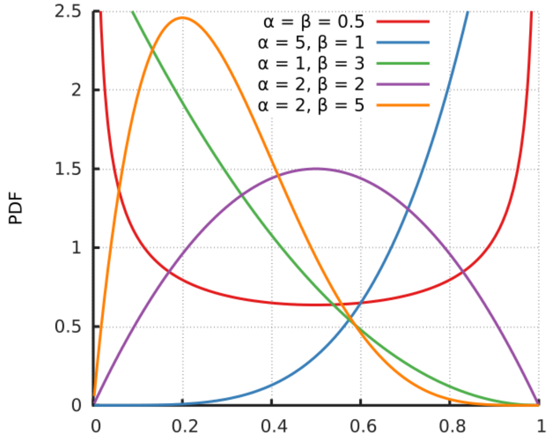
Aidan Lyon, Why are Normal Distributions Normal?, The British Journal for the Philosophy of Science 2014 65:3, 621-649

# Other statistical distributions

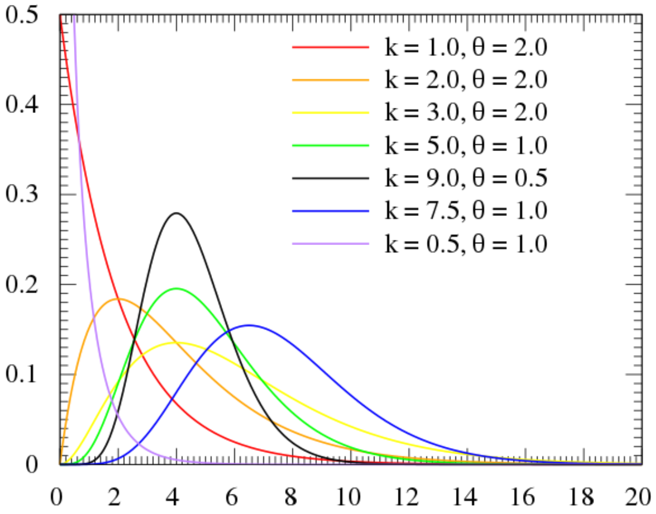
### Hypergeometric distribution



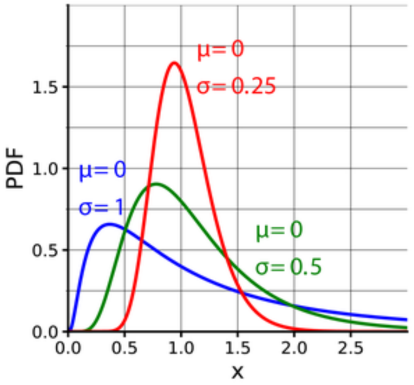
### Beta distribution

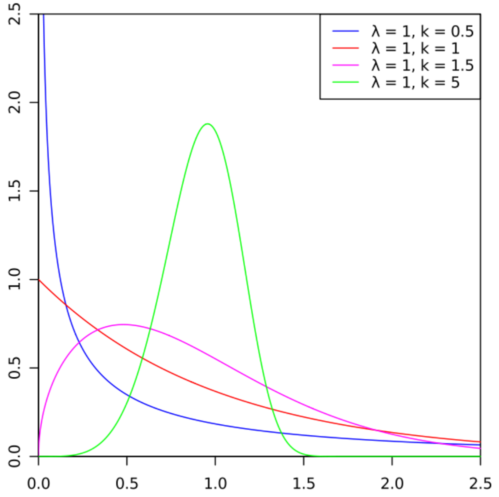


### Gamma distribution



### Log-normal distribution



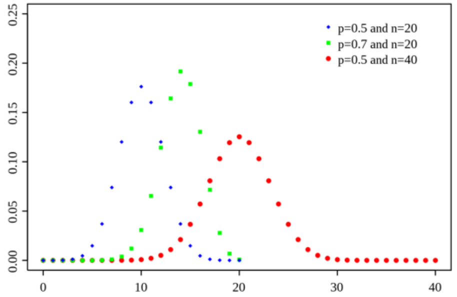


### Weibull distribution

# Data analysis: frequentist inference

Here, it is assumed that data are generated by a statistical distribution with parameters **η,θ,…,ψ**.
The principal aim is to infer information about **η,θ,…,ψ**.

1) Statistical distribution

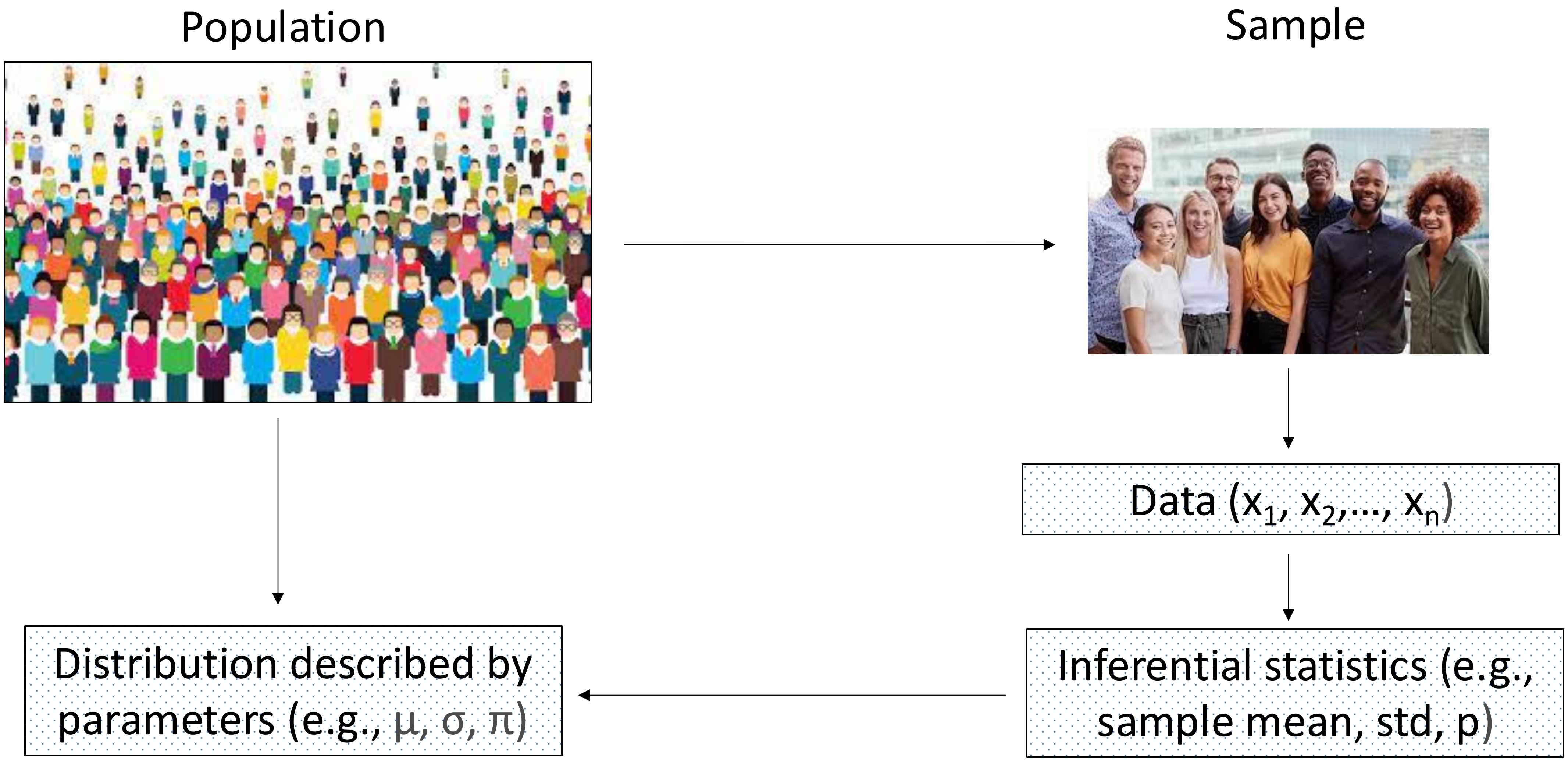


3) Frequentist inference

2) Data

E.L. Lehmann, George Casella, Theory of Point Estimation, Second Edition

# …to be clear about frequentist inference



Population

Sample

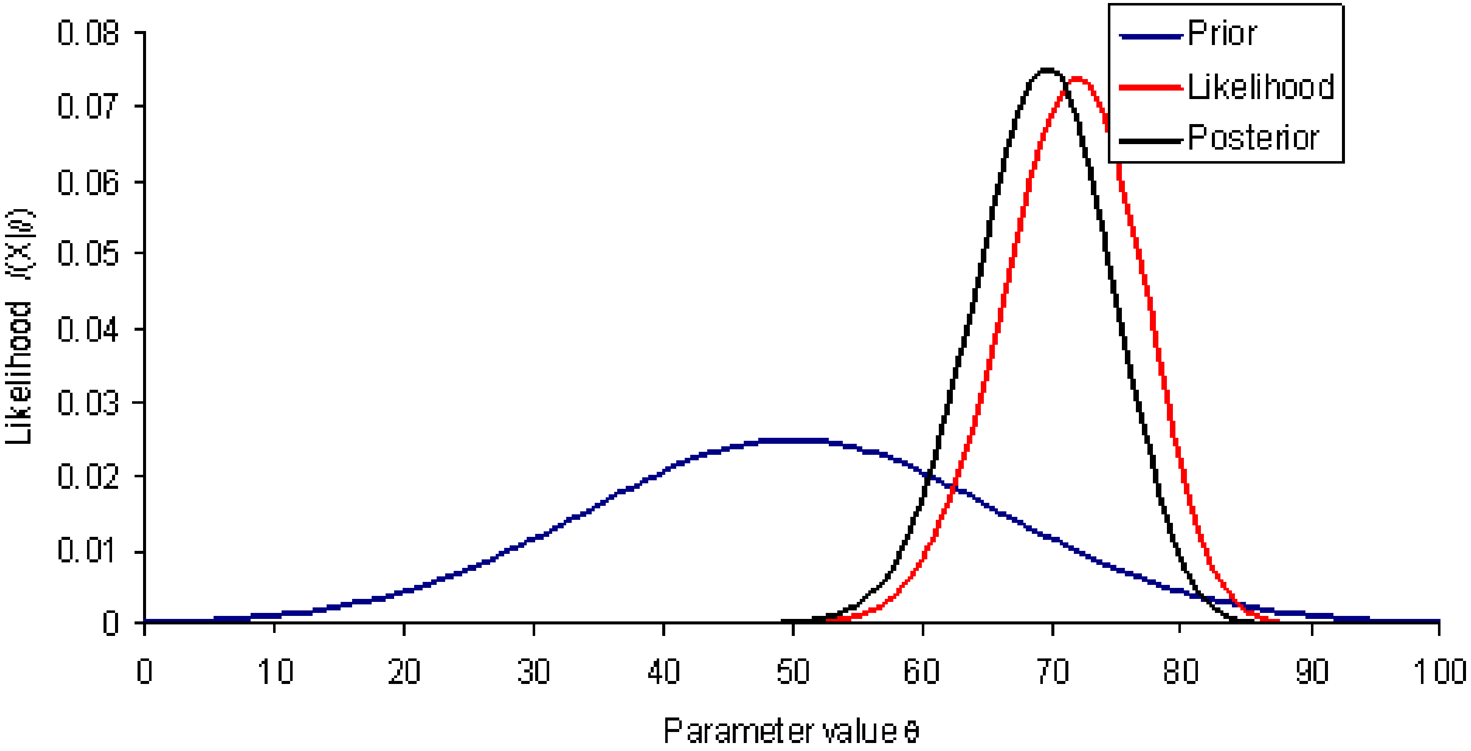
Data ($x_1, x_2, \ldots, x_n$)

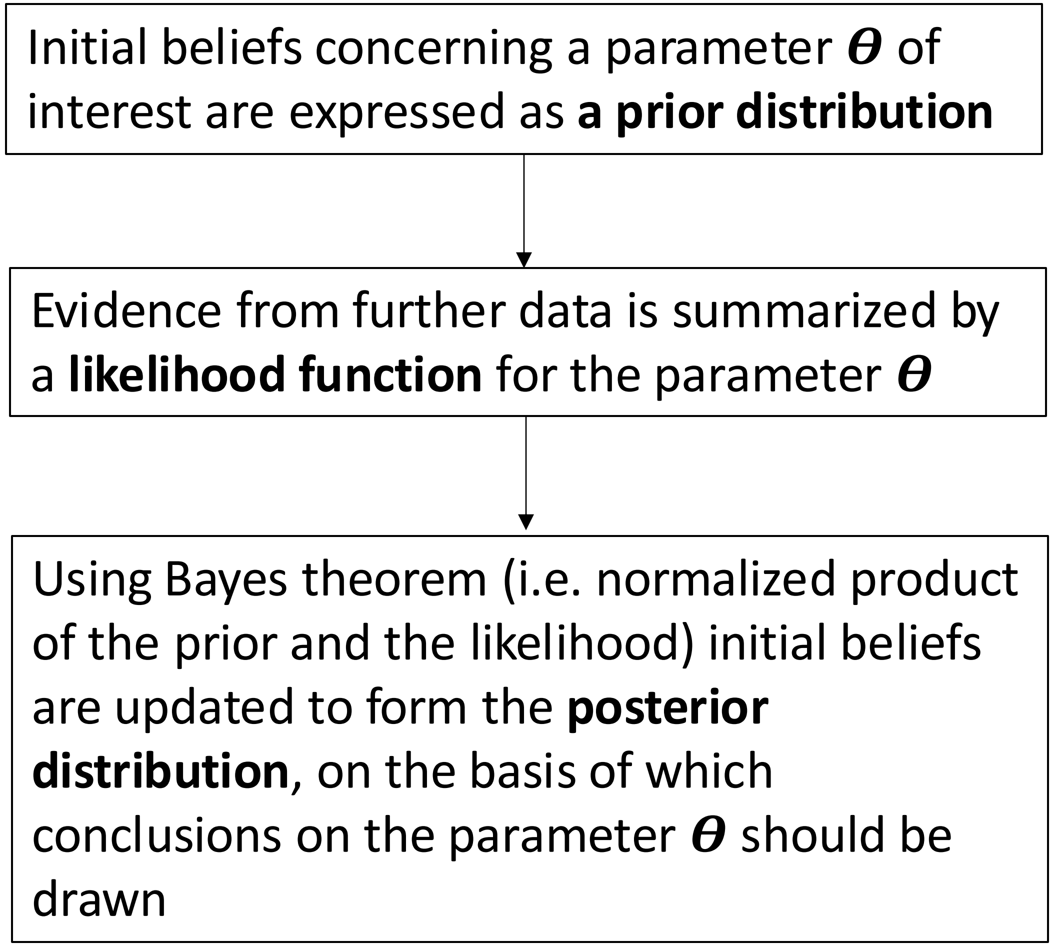
Inferential statistics (e.g., sample mean, std, p)

Distribution described by parameters (e.g., $\mu$, $\sigma$, $\pi$)

# Data analysis: bayesian inference

The basic paradigm of bayesian statistics





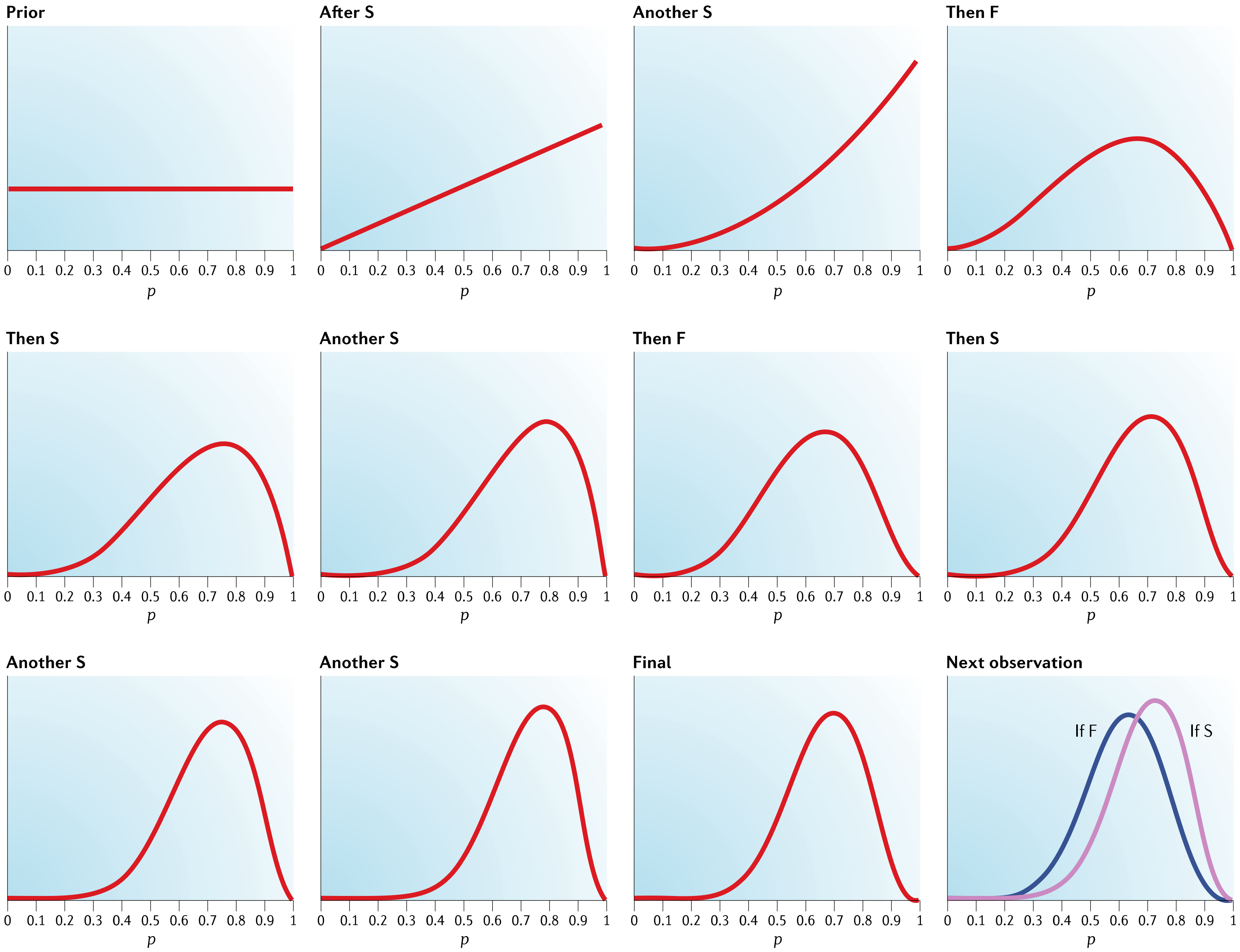
Initial beliefs concerning a parameter $\boldsymbol{\theta}$ of interest are expressed as **a prior distribution**

Evidence from further data is summarized by a **likelihood function** for the parameter $\boldsymbol{\theta}$

Using Bayes theorem (i.e. normalized product of the prior and the likelihood) initial beliefs are updated to form the **posterior distribution**, on the basis of which conclusions on the parameter $\boldsymbol{\theta}$ should be drawn

D.J. Spiegehalter et al., Bayesian Approaches to Randomized Trials, *J.R. Statist. Soc.* A (1994) 157, *Part 3*, *pp.* 357-416

# A practical example of bayesian inference



Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006 Jan;5(1):27-36. doi: 10.1038/nrd1927. PMID: 16485344

# Data analysis: bayesian inference

The recourse to the prior distribution on the parameters of a model is questionable. There is in fact a major step from the notion of an *unknown* parameter to the notion of a *random* parameter.

Edwards A.W.F., Likelihood (Expanded Edition), 1992

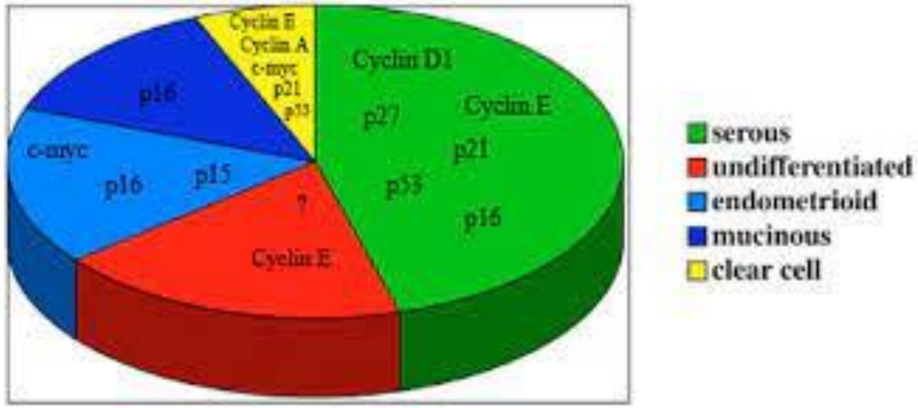
# Descriptive statistics, univariate analysis

**Nominal data**

| **Histotype** | **N** | **%** |
|---|---|---|
| Serous | 48 | 41.7 |
| Undifferentiated | 22 | 19.1 |
| Endometrioid | 21 | 18.3 |
| Mucinous | 16 | 13.9 |
| Clear cell | 8 | 7.0 |

Serous is the mode.
The mode is the category that appears most often in a set of data values

**Pie chart**



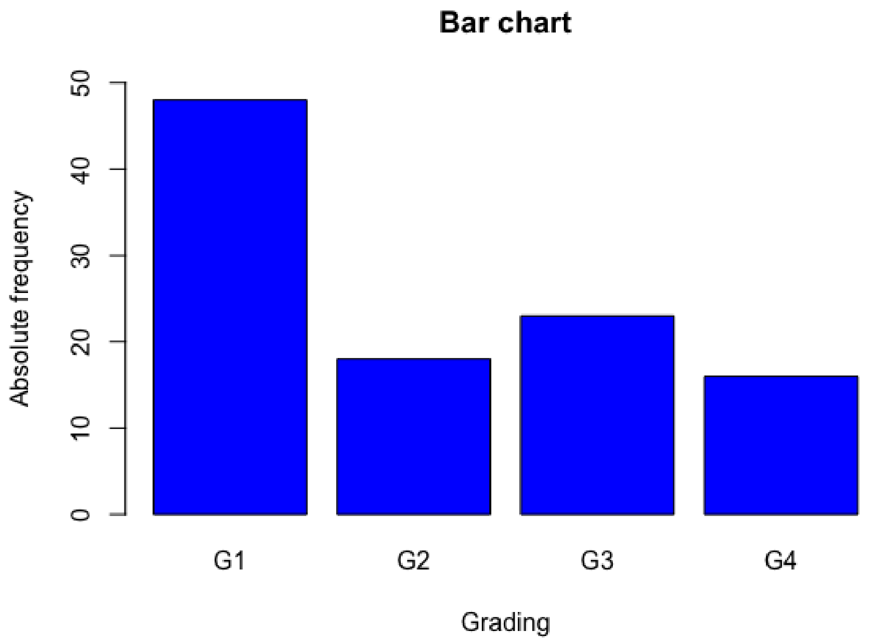
# Descriptive statistics, univariate analysis

**Ordinal data**

| **Grading** | **N** | **%** |
|---|---|---|
| G1 | 48 | 45.7 |
| G2 | 18 | 17.1 |
| G3 | 23 | 21.9 |
| G4 | 16 | 15.2 |

G1 is the mode.

G2 is the median.
The median is the category separating the higher half from the lower half of a data sample



# Descriptive statistics, univariate analysis

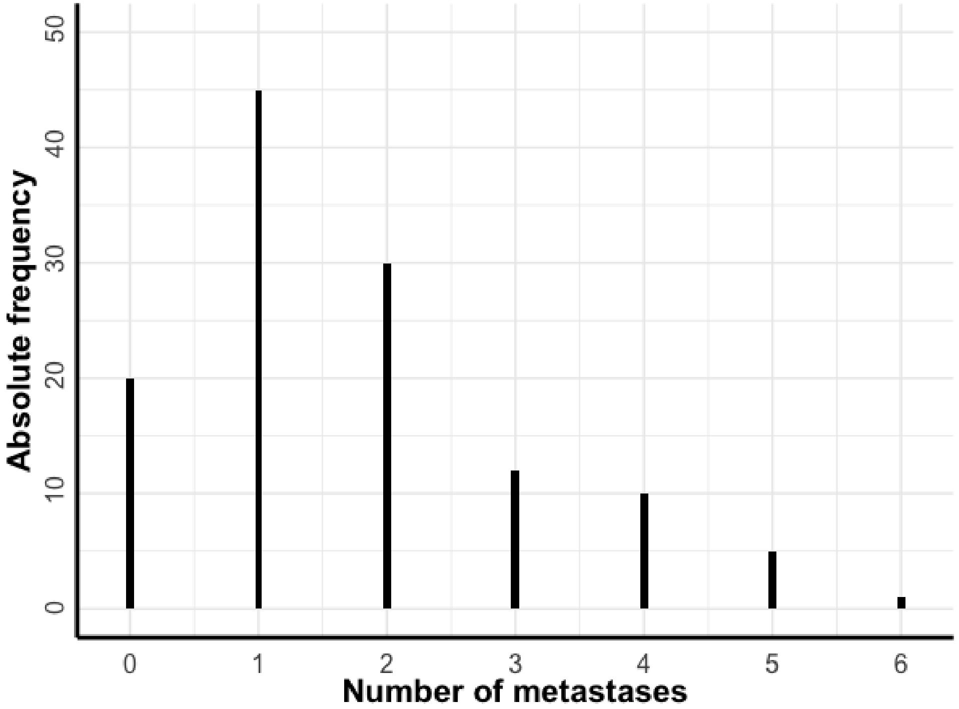
**Discrete data**

| N. of metastases | N | % |
|---|---|---|
| 0 | 20 | 16.3 |
| 1 | 45 | 36.6 |
| 2 | 30 | 24.4 |
| 3 | 12 | 9.8 |
| 4 | 10 | 8.1 |
| 5 | 5 | 4.1 |
| 6 | 1 | 0.8 |

1 metastasis is the mode.

1 metastatis is the median.

1.7 is the mean number of metastases.



# Descriptive statistics, univariate analysis

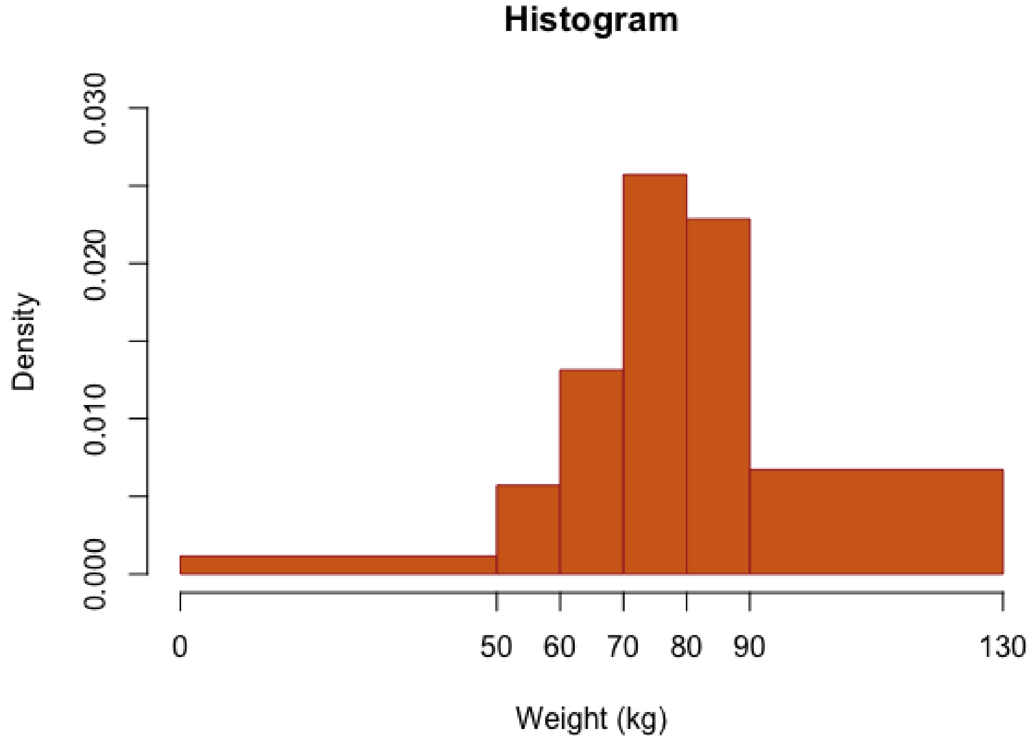
**Continuous data**

| **Weight (kg)** | **N (%)** | **N/10kg** |
|---|---|---|
| 0 -\| 50 | 10 (5.7) | 2 |
| 50 -\| 60 | 10 (5.7) | 10 |
| 60 -\| 70 | 23 (13.1) | 23 |
| 70 -\| 80 | 45 (25.7) | 45 |
| 80 -\| 90 | 40 (22.9) | 40 |
| 90 -\| 130 | 47 (26.9) | 11.8 |

→ 70 -| 80 is the modal interval.

→ 70 -| 80 is the median interval.

→ 81.4 kg is the mean weight of patients.



# Descriptive statistics, univariate analysis

**Quantitative data**

Properties of the mean

- ○ $\sum_i (a_i - \boldsymbol{\mu}) = 0$
- ○ $\sum_i (a_i - \boldsymbol{\mu})^2 < \sum_i (a_i - x)^2, x \neq \boldsymbol{\mu}$
- ○ Linearity
- ○ Associative property

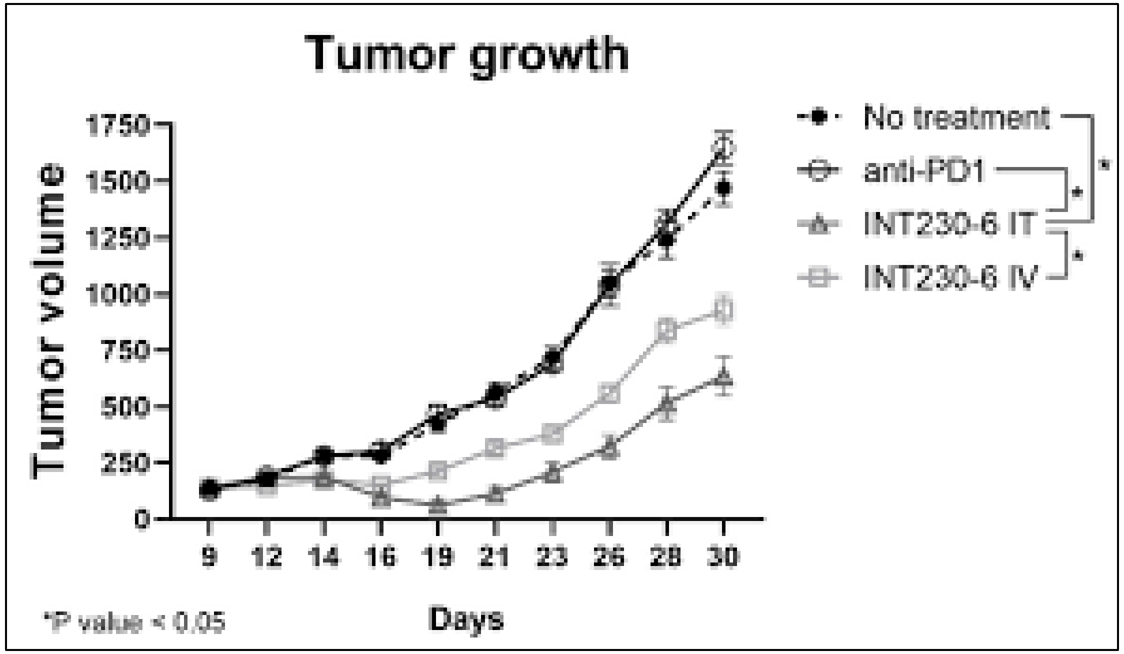
# Descriptive statistics, univariate analysis

**Quantitative data**

Measures of variability

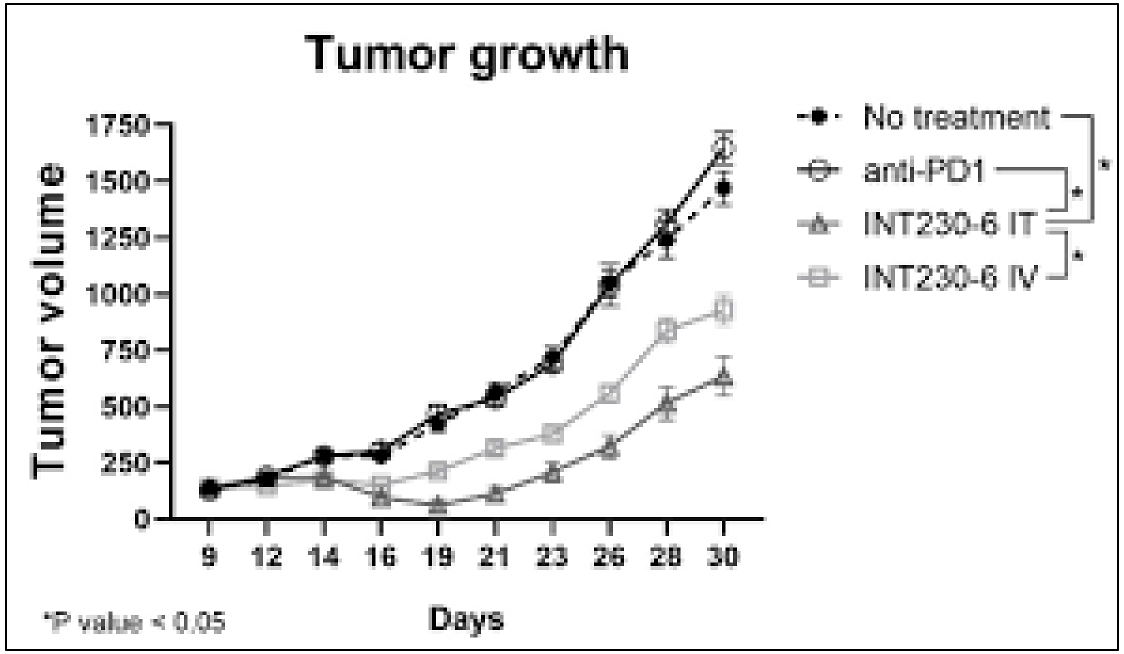
- ○ $(\sum_i |a_i - M|)/n$, where M is a measure of central tendency
- ○ $\sqrt{}[(\sum_i |a_i - M|^2)/n]$, if $M = \boldsymbol{\mu}$, it is called standard deviation ($\boldsymbol{\sigma}$)
- ○ Interquartile range (IQR)

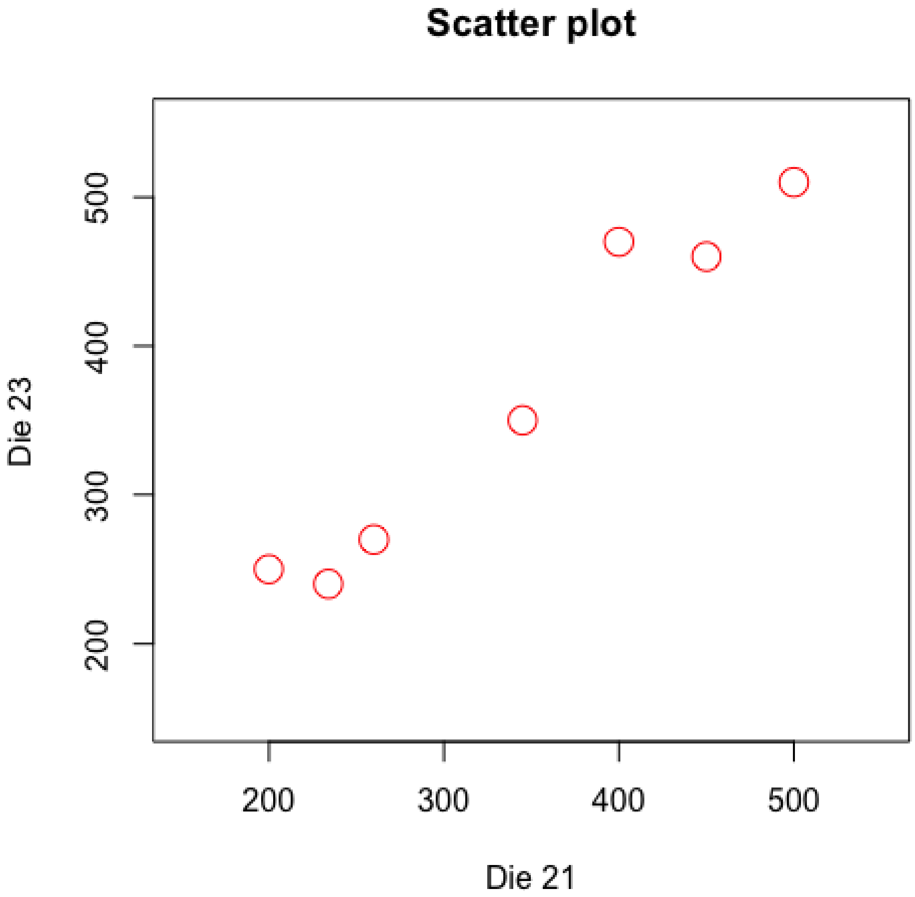
# Descriptive statistics, bivariate analysis



| ID mouse | Day 21, mm $^3$ | Day 23, mm $^3$ |
|---|---|---|
| M101 | 260 | 270 |
| M102 | 234 | 240 |
| M103 | 400 | 470 |
| M104 | 345 | 350 |
| M105 | 450 | 460 |
| M106 | 200 | 250 |
| M107 | 500 | 510 |

# Descriptive statistics, bivariate analysis





**Take home message:**

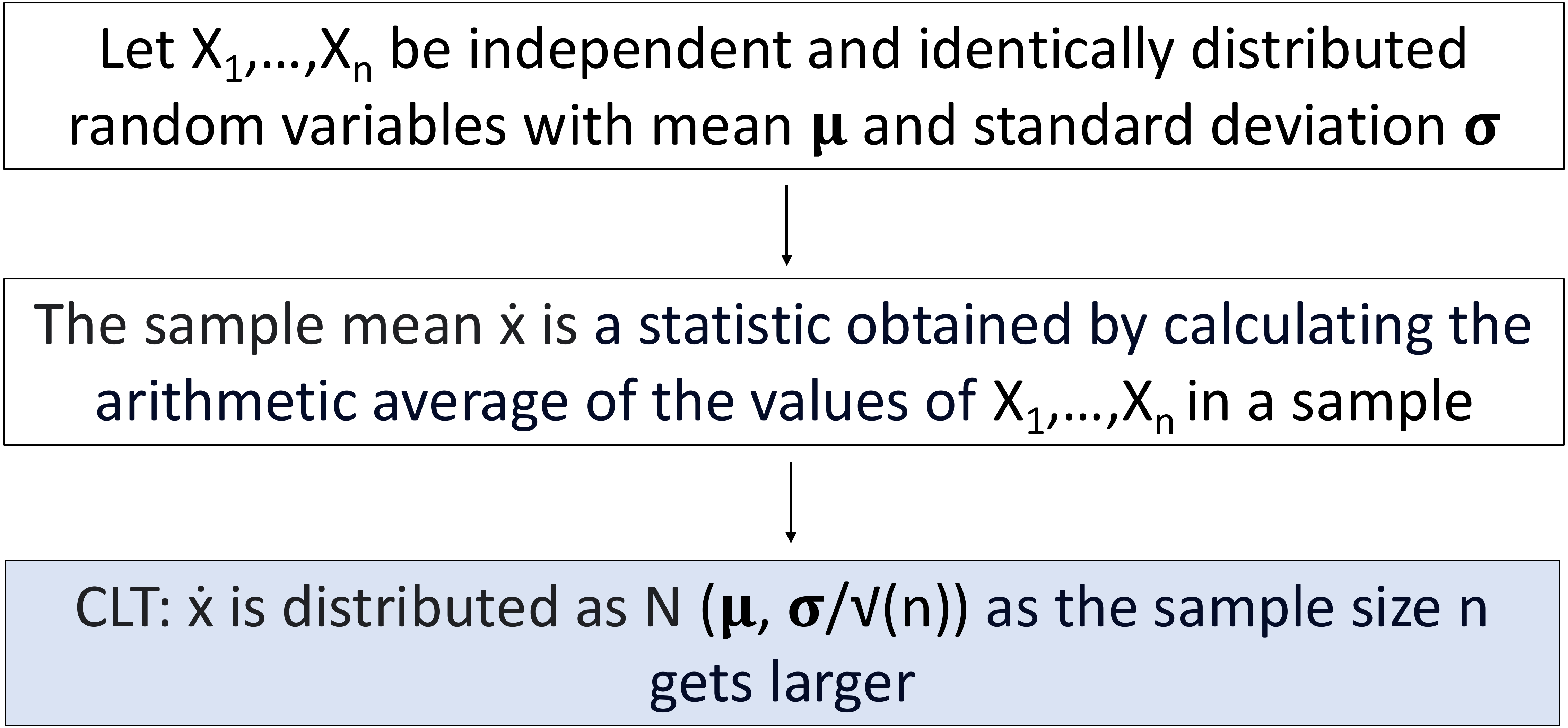
Paired data are not independent. They correlate.

# Online quiz

Exercises

http://bioinformatics-core-shared-training.github.io/IntroductionToStats

# Central limit theorem (CLT)



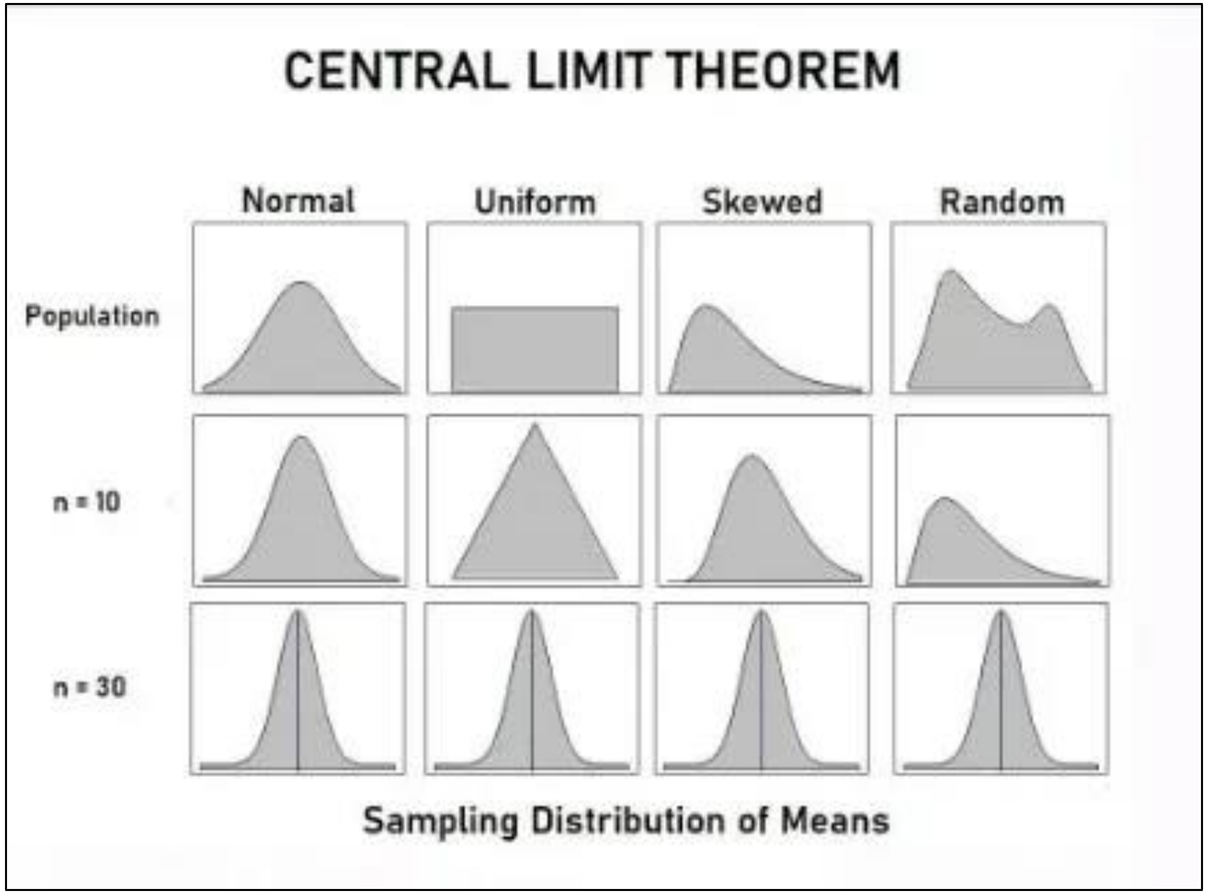
Let $X_1,\ldots,X_n$ be independent and identically distributed random variables with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$

The sample mean $\dot{x}$ is a statistic obtained by calculating the arithmetic average of the values of $X_1,\ldots,X_n$ in a sample

CLT: $\dot{x}$ is distributed as N ($\boldsymbol{\mu}$, $\boldsymbol{\sigma}/\sqrt{(n)}$) as the sample size n gets larger

# Central limit theorem (CLT)

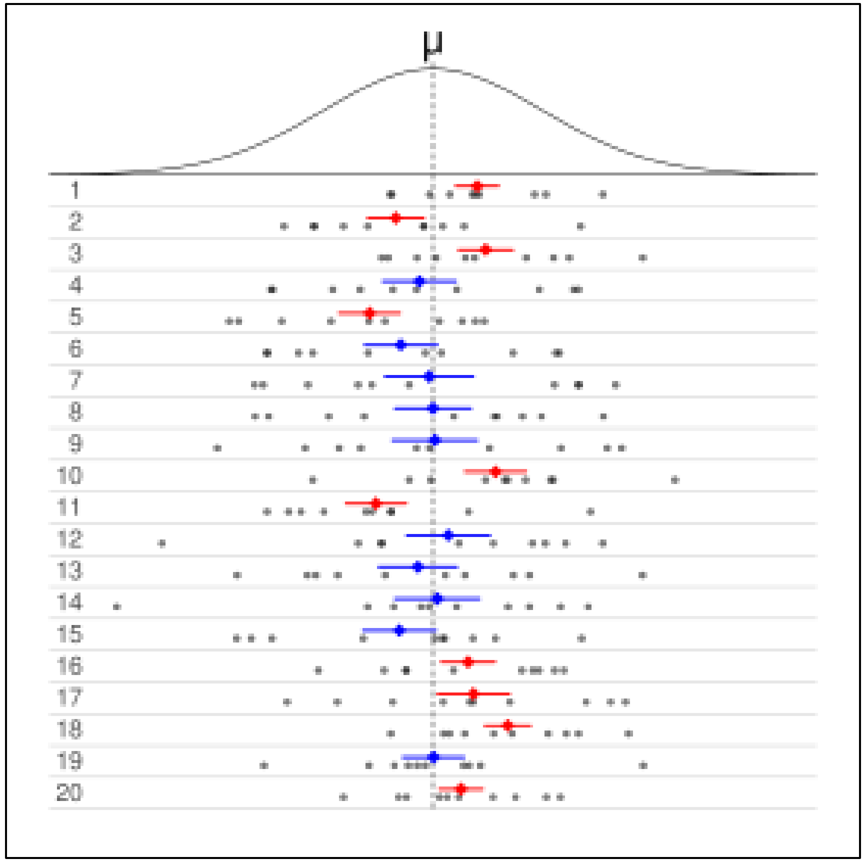
Central limit theorem



**The usefulness of the CLT is that the distribution of sample means approaches normality regardless of the distribution of the population**

# Confidence intervals

In frequentist inference, a confidence interval (CI) is a **range of estimates** for an unknown parameter Θ.



It is computed at a designated confidence level (e.g., 95% CI). The confidence level represents the long-run proportion of CIs that theoretically contain the true value of the parameter Θ.

For example, out of all intervals computed at the 95% level, 95% of them should contain the parameter's true value.

https://en.wikipedia.org/wiki/Confidence_interval, 15th August 2024

# Confidence intervals for the normal distribution

Normal data, **σ** known: one sample z-confidence interval

Sample mean $\dot{x}$ is <u>exactly</u> distributed according to N (**μ**, **σ**/√(n))

95% CI = $\dot{x} \pm z_{0.975} \cdot \boldsymbol{\sigma}/\sqrt{(n)}$, where $z_{0.975} \simeq 1.96$

If you do not know **σ**

# Student’s *t*-distribution

Let $x_1,…,x_n$ be independent and identically distributed observations from a normal distribution with mean $\boldsymbol{\mu}$ and std $\boldsymbol{\sigma}$.

↓

The sample mean and unbiased sample standard deviation are given by:

$\dot{x} = (x_1+…+x_n)/n$ [biological signal collected in the sample]

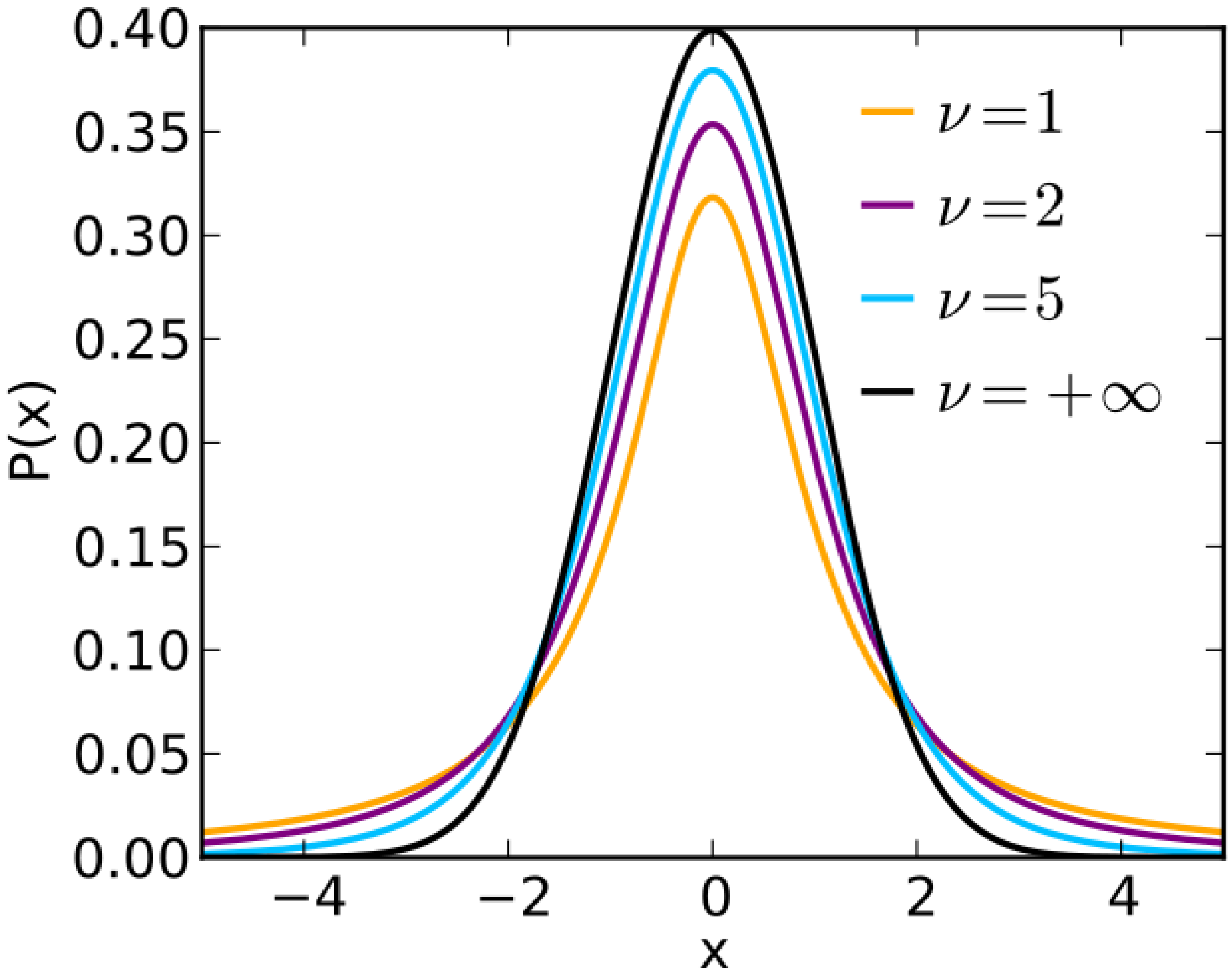
$std^2 = (1/(n-1))\ \boldsymbol{\Sigma}_i\ (x_i - x_m)^2$ [noise collected in the sample]

↓

$(\dot{x} - \mu) / (std / \surd n) \sim t_{n-1}$ is distributed according to a Student’s *t*-distribution with n-1 degrees of freedom

The *t*-statistic has a probability distribution that not depends on the unknown $\boldsymbol{\sigma}$

# Student's *t*-distribution



# Confidence intervals for the normal distribution

Normal data, $\boldsymbol{\sigma}$ unknown: one sample *t*-confidence interval

Sample mean $\dot{x} - \boldsymbol{\mu}$ is <u>exactly</u> distributed according to $[\mathrm{std}/\sqrt{(n)}] \cdot t_{n-1}$

95% CI = $\dot{x} \pm t_{n-1,\ 0.975} \cdot \mathrm{std}/\sqrt{(n)}$. We use the *t*-tables to obtain these "critical" values

If data are not normally distributed…

# Consequence of CLT

*t*-distribution methods are robust when the sample size is large (n ≥ 30). The data should not have extreme outliers or evidence of severe skewness.

For small samples it is risky to use *t*-confidence intervals. Only use if you are sure the population is roughly normally distributed and the sample has no outliers and very little skew. Otherwise, other methods (e.g. bootstrap) should be used.

# Simulations

Exercises

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html

Shiny web application

https://bioinformatics.cruk.cam.ac.uk/apps/stats/central-limit-theorem

# Hypothesis Testing

- A hypothesis is a statement about the population(s).

  Example n.1: Carboplatin induced response in at least 70% of NSCLC patients

  Example n.2: The mean pressure is the same in C57BL/6J and DBA/2J mice

  Example n.3: The two populations A and B have the same height distribution

- The goal of a hypothesis test is to decide, **based on data collected**, which of two complementary hypotheses is true.

  Example n.1: $H_0$: RR < 0.70; $H_1$: RR $\geq$ 0.70

  Example n.2: $H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$; $H_1$: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}$

  Example n.3: $H_0$: $D_A = D_B$; $H_1$: $D_A \neq D_B$

# Hypothesis Testing

$H_0$: null hypothesis
$H_1$: alternative hypothesis

There is no symmetry between $H_0$ and $H_1$:

Procedure

1st step: We assume $H_0$ to be true

2st step: The strength of evidence provided by the data **against** $H_0$ is measured

3st step: If a contradiction is found, $H_1$ is accepted.
If a contradiction is not found, the method of proof fails and the hypothesis $H_0$ could be either true or false

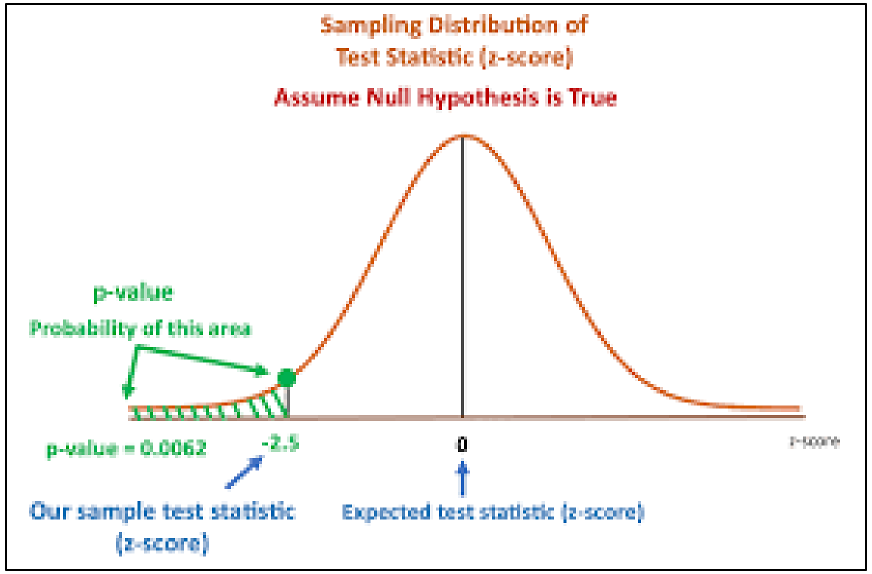
# Strength of evidence provided by the data

Procedure

Data: $x_1,\ldots,x_n$

Test statistic: $t_s = f(x_1,\ldots,x_n)$

Distribution of the test statistic under $H_0$:



The p-value is the statistical index used to measure the strength of evidence against $H_0$.

# Evidence provided by the data

$H_0$: θ = 0, θ ∈ {0, 1, 2}
$H_1$: θ = 1,2

Distribution of the test statistic under $H_0$:

| **$t_s$** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Prob ($t_s$ \| $H_0$)** | 0.980 | 0.005 | 0.005 | 0.010 |
| **P-value** | 1.00 | 0.01 | 0.01 | 0.020 |

An **α significance level** (e.g. 0.05) is simply a decision rule as to which p-values will cause one to reject the null hypothesis. In other words, it is merely a decision point as to how weird the data must be before rejecting the null model. If the p-value is less than or equal to α, the null is rejected. Implicitly, an α level determines what data would cause one to reject $H_0$ and what data will not cause rejection. The α level rejection region is defined as the set of all data points that have a p-value less than or equal to α.

# The two types of errors in hypothesis testing

| | | **Decision** | |
|---|---|---|---|
| | | Accept $H_0$ | Reject $H_0$ |
| **Truth** | $H_0$ | Correct decision | Type I error ($\boldsymbol{\alpha}$) |
| | $H_1$ | Type II error ($\boldsymbol{\beta}$) | Correct decision |

1. If the hypothesis test incorrectly decides to reject $H_0$, then the test has made a Type I error  (i.e. false positive decision)

2. If the hypothesis test incorrectly decides to not reject $H_0$, then the test has made a Type II error (i.e. false negative decision)

# Statistical power

The power (1-$\boldsymbol{\beta}$) of a hypothesis test is the probability to reject the null hypothesis ($H_0$) if $H_1$ is true. It is a function of alternative simple hypotheses.

## Distribution of the test statistic under $H_0$:

| **$t_s$** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Prob ($t_s$ \| $H_0$)** | 0.980 | 0.005 | 0.005 | 0.010 |
| **P-value** | 1.00 | 0.010 | 0.010 | 0.020 |

Rejection region of the test with **α significance level= 0.05**

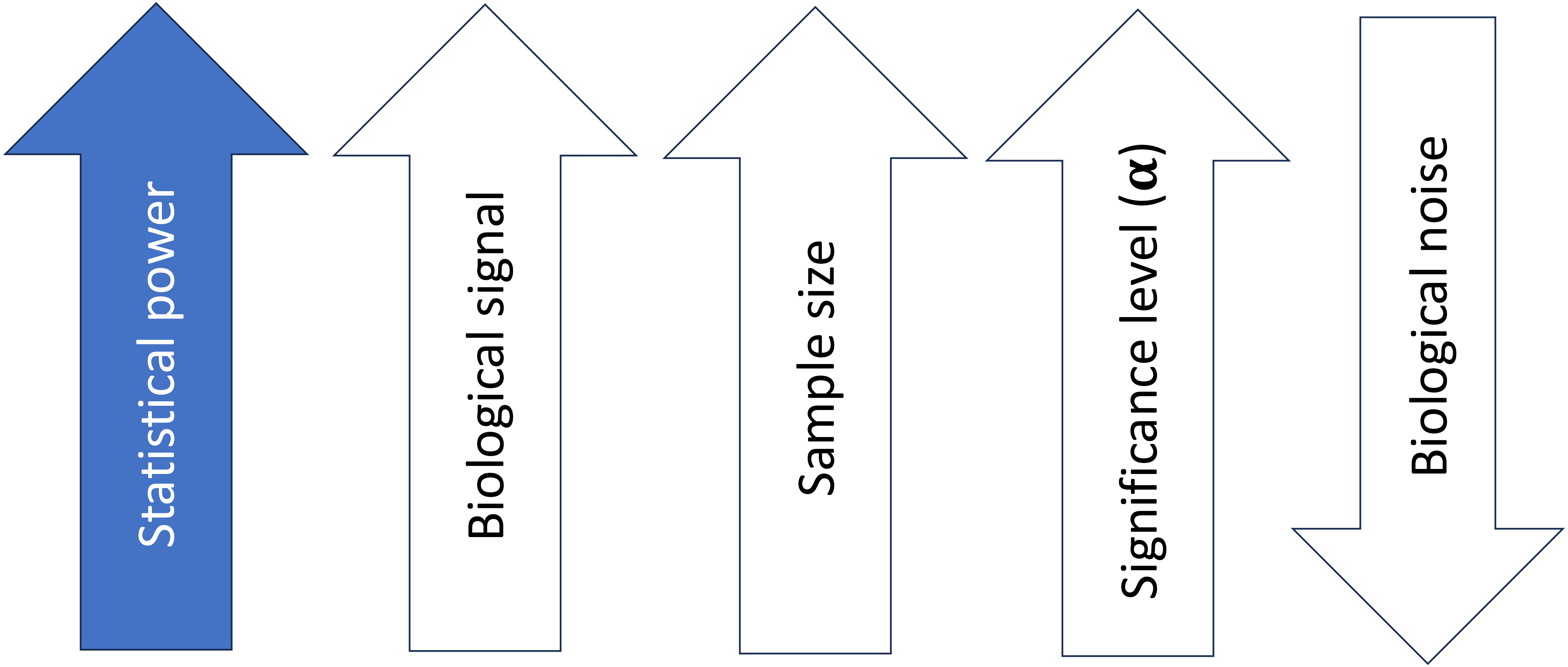
## Distribution of the test statistic under $H_1$:

| **$t_s$** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Prob ($t_s$ \| θ = 1)** | 0.100 | 0.200 | 0.200 | 0.500 |
| **Prob ($t_s$ \| θ = 2)** | 0.098 | 0.001 | 0.001 | 0.900 |

→ (1-$\boldsymbol{\beta}$ | **θ = 1**) = 0.900

→ (1-$\boldsymbol{\beta}$ | **θ = 2**) = 0.902

# Statistical power



# Distribution-free tests

A distribution-free test is one which makes **no assumptions** about the precise form of the sampled population or the assumptions are never so elaborate as to imply a population whose distribution is **completely specified**.

| **Distribution-free tests** | **Distribution-dependent tests** |
|---|---|
| Sign test | One-sample Student's *t*-test |
| Wilcoxon signed-rank test | Two-sample Student's *t*-test |
| Wilcoxon rank-sum test | Unequal variance *t*-test (i.e. Welch's *t*-test) |

- Bradley, J.V. (1968) Distribution-Free Statistical Tests. Prentice-Hall, Englewood Cliffs, NJ
- Kendall, M.G. and R.M.Sundrum, Distribution-Free Methods and Order Properties, Review of the International Statistical Institute, 3 (1953), 124-134

# Multiple testing

Let $m > 1$ the number of null hypotheses $H_1, \ldots, H_m$ to be tested.

| $H_i$ | Not rejected | Rejected | Total |
|---|---|---|---|
| True | U | V | $m_0$ |
| False | T | S | $m-m_0$ |
| Total | W | R | m |

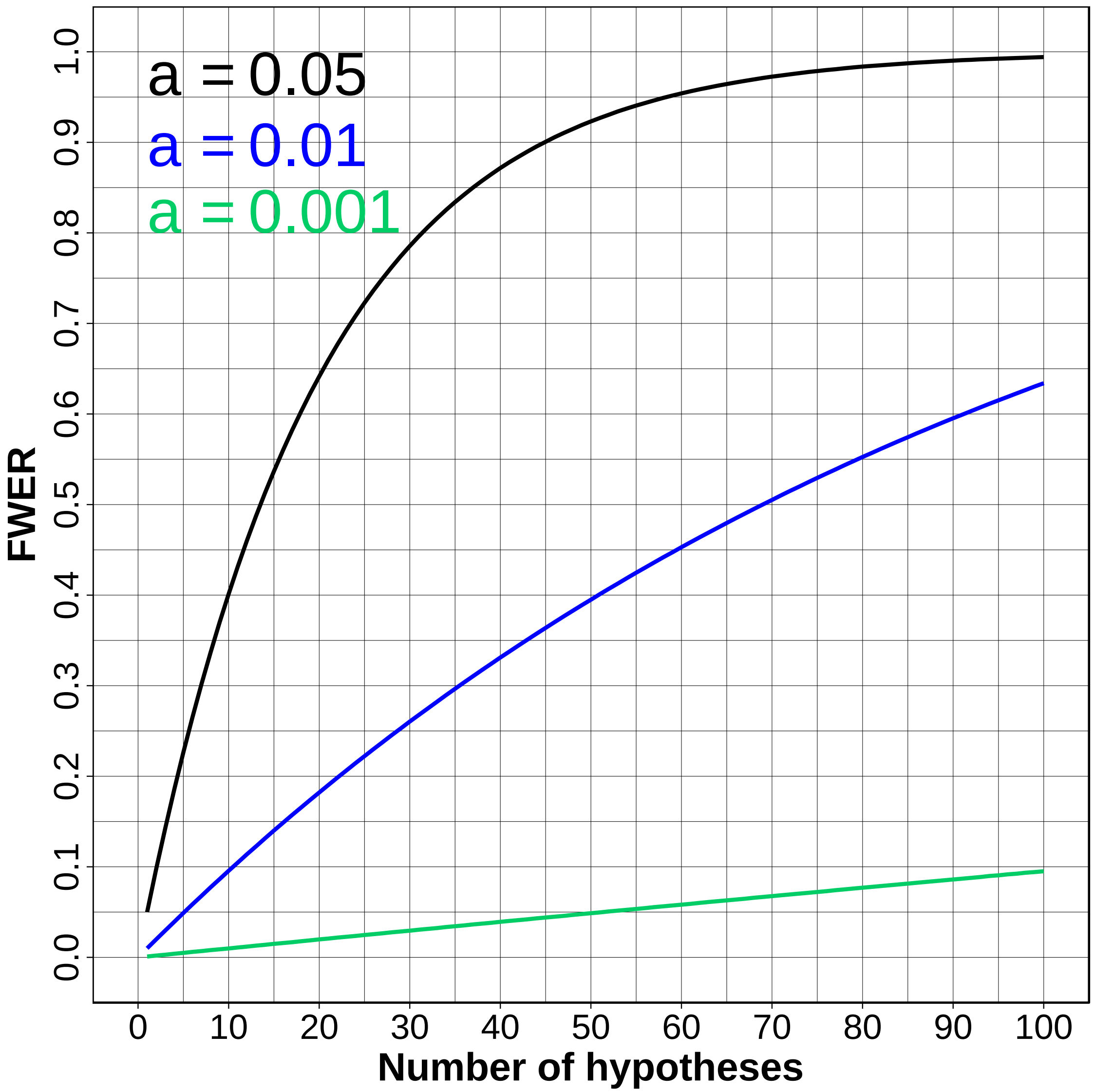
## Familywise error rate (FWER)

P(V>0), which is the probability of committing al least one Type I error.

Notes:

- If all hypothesis tests are conducted with a significance level of $\alpha$, the FWER $\geq \alpha$
- the FWER reduces to the common Type I error rate for m=1.

# Multiple testing



Under the assumption of independent statistical hypotheses.

# Multiple testing

**Simultaneously testing statistical hypotheses**

| | | | |
|---|---|---|---|
| $H_1$ | $H_2$ | .... | $H_m$ |

## Controlling Familywise error rate (FWER)

- **Bonferroni method**

  - Let m > 1 the number of null hypotheses $H_1,....,H_m$ to be tested.
  - Each p-value ($p_i$, i=1,…m) is compared to the threshold $\alpha$ / m. Then FWER $\leq \alpha$.

  ✓ No assumption is requested.

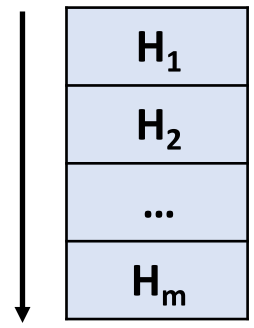
- **Sidak method**

  - Let m > 1 the number of null hypotheses $H_1,....,H_m$ to be tested.
  - Each p-value ($p_i$, i=1,…m) is compared to the threshold $\alpha_{SIDAK} = 1 - (1- \alpha)^{1/m}$. Then FWER $\leq \alpha$.

  ✓ Very strong assumption: hypothesis tests are independent.

# Multiple testing

Sequentially testing statistical hypotheses



## Controlling Familywise error rate (FWER)

- **Sequential Bonferroni method (Holm's method)**

  - Let m > 1 the number of null hypotheses $H_1,\ldots,H_m$ to be tested.
  - Sort p-values ($p_i$, i=1,…m) $p_1 \leq p_2 \leq \ldots \leq p_m$.
  - Each p-value ($p_i$, i=1,…m) is compared to the threshold $\alpha / (m-i+1)$. Reject the corresponding $H_i$ hypothesis until the p-value is no longer significant. Then FWER $\leq \alpha$.

  ✓ No assumption is requested. Holm's method more powerful than Bonferroni.

- **Sequential Sidak method**

  The only difference respect to the sequential Bonferroni method are the thresholds $\alpha_{SIDAK,i} = 1 - (1- \alpha)^{1/(m-i+1)}$. Then FWER $\leq \alpha$.

  ✓ Very strong assumption: hypothesis tests are independent.

# One-sample location tests

# One-sample Student’s *t*-test

**Assumptions:**

1. the data are continuous
2. sample data have been randomly sampled from a population
3. independent observations $x_i$, $i=1,\dots,n$
4. the population is normally distributed

**Hypotheses to test:**

$H_0$: mean of the population distribution $\boldsymbol{\mu} = \boldsymbol{\mu}_0$

$H_1$: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

**Test statistic:**

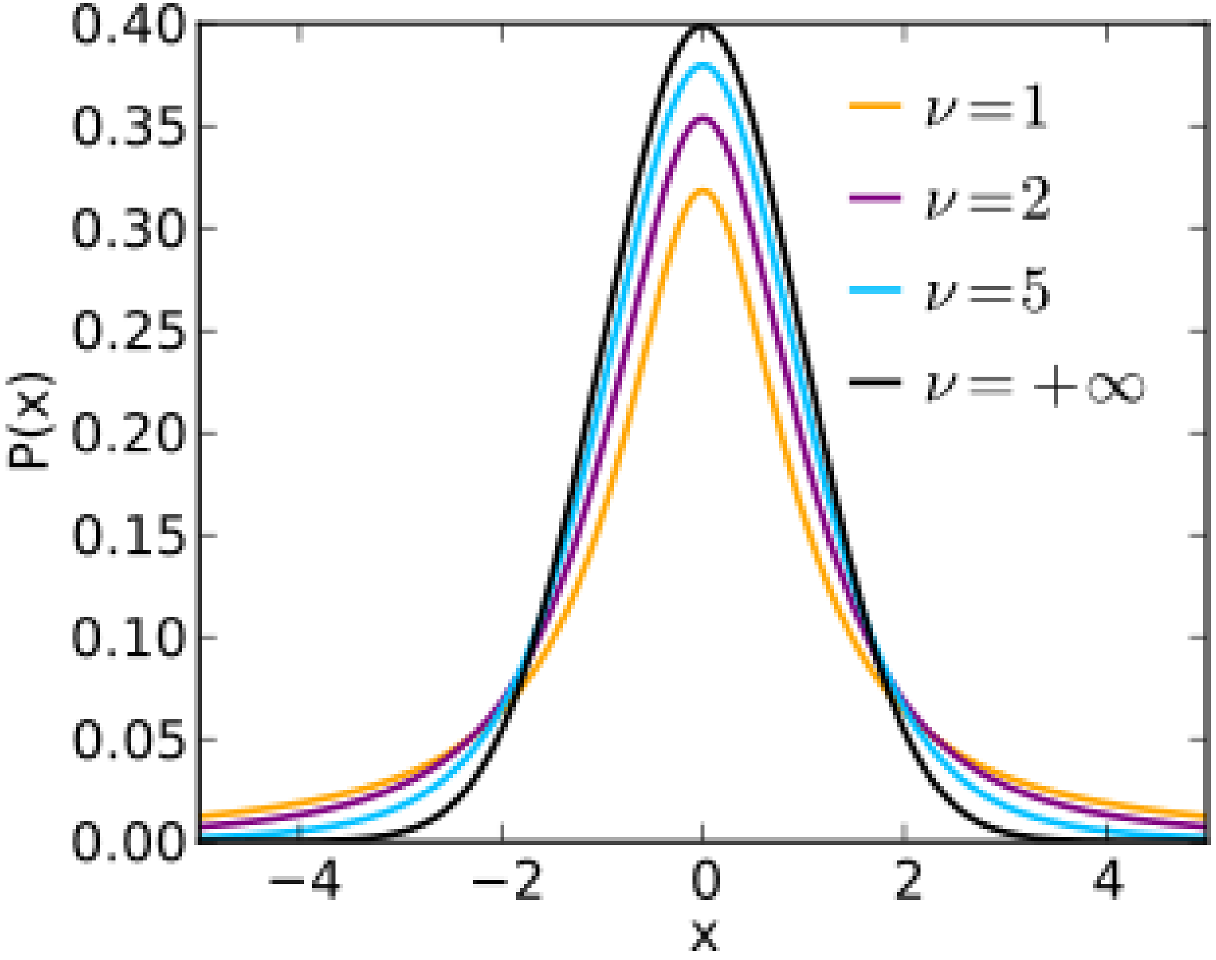
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

ẋ = sample mean

s = sample standard deviation

# One-sample Student's *t*-test

**Distribution of the test statistic:** *t*-distribution with n-1 degrees of freedom



# Sign test

**Assumptions:**

1. the data are continuous
2. sample data have been randomly sampled from a population
3. independent observations $x_i$, i=1,…,n

**Hypotheses to test:**

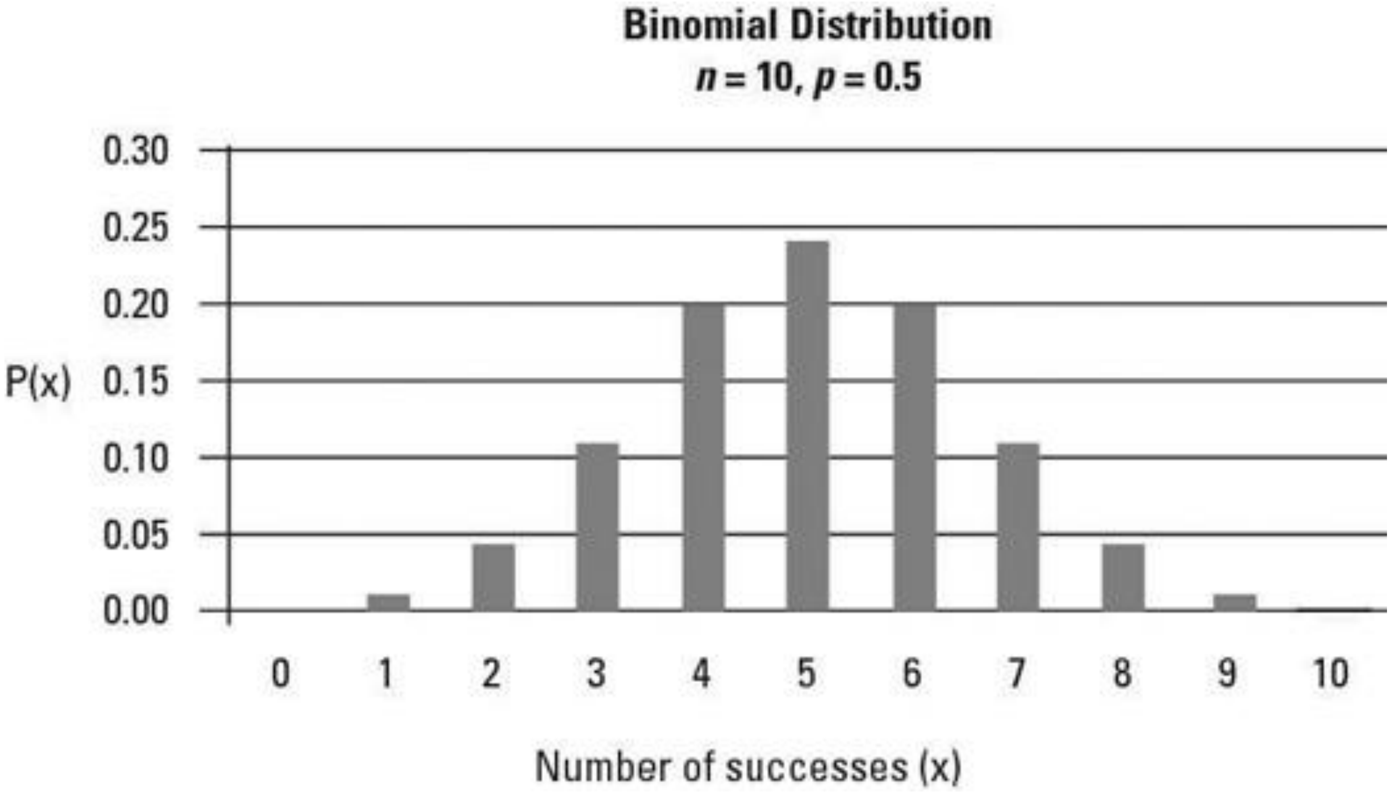
$H_0$: median of the population distribution $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$H_1$: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

**Test statistic:** number of values above (or below) $\boldsymbol{\theta}_0$.

# Sign test

**Distribution of the test statistic:** binomial distribution, X ~ Bin (n, 0.5)



In case of values equal to $\boldsymbol{\theta}_0$, discarde these values and apply the sign test only to the values above or below $\boldsymbol{\theta}_0$.

# Wilcoxon signed-rank test

**Assumptions:**

1. the data are continuous
2. sample data have been randomly sampled from a population
3. independent observations $x_i$, $i=1,\ldots,n$
4. the population distribution is symmetric

**Hypotheses to test:**

$H_0$: median/mean of the population distribution $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$H_1$: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

**Test statistic:** sum of the positive signed ranks.

# Wilcoxon signed-rank test

| | |
|---|---|
| **n:** | 3 |
| **Raw data:** | 67, -12, 55 |
| **$\boldsymbol{\theta}_0$:** | 50 |
| **Absolute differences:** | 5, 17, 62 |
| **Signed ranks:** | +1, +2, -3 |
| **Test statistic:** | +3 |
| **Distribution of the test statistic:** | P(+1,+2,+3) = P(+1,+2,-3) = P(+1,-2,+3) = P(+1,-2,-3) = P(-1,+2,+3) = P(-1,+2,-3) = P(-1,-2,+3) = (-1,-2,-3)= 1/8, hence… |

# Wilcoxon signed-rank test

**Distribution of the test statistic:**

| **Sum of signed ranks** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| **Probability** | 1/8 | 1/8 | 1/8 | 2/8 | 1/8 | 1/8 | 1/8 |

Two-sided p-value: 1.0

One-sided p-value: 5/8=0.625

At the significance level of 0.05, we can't reject the null hypothesis ($\boldsymbol{\theta}$=50).

# Take home message

The Wilcoxon signed-rank test is more powerful than the sign test because it makes use of the magnitudes of the differences rather than just their sign.

It should be the preferred method, but it makes a stronger assumption: the distribution of the differences is symmetric.

In case this assumption is doubtful, the sign test should be used.
Graphical display is *recommended.*

# Take home message

The one-sample location tests could be used for paired data samples.

Each paired data is summarized by the difference and the one-sample location tests are applied to the differences.

| Experimental unit | Paired data | Difference |
|---|---|---|
| 1 | 23-55 | -32 |
| … | … | … |
| k | 107-100 | 7 |

# Exercises

Exercises

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html

Shiny web application

https://bioinformatics.cruk.cam.ac.uk/stats/shinystats/

# Two-sample location tests

# Two-sample Student's *t*-test

**Assumptions:**

1. data are continuous
2. random sampling from the two populations
3. independent observations $x_i$, $i=1,\ldots,n_1$ and $y_j$, $j=1,\ldots,n_2$
4. the two population distributions are normal
5. equal variances $s_1^2$ and $s_2^2$

**Hypotheses to test:**

$H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

$H_1$: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

**Test statistic:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

The statistic t has a Student's *t* distribution with $n_1+n_2-2$ degrees of freedom.

# Unequal variance *t*-test (i.e. Welch's *t*-test)

**Assumptions:**

1. data are continuous
2. random sampling from the two populations
3. independent observations $x_i$, $i=1,\ldots,n_1$ and $y_j$, $j=1,\ldots,n_2$
4. the two population distributions are normal

**Hypotheses to test:**

$H_0$: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

$H_1$: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

**Test statistic:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The statistic t has a Student's *t* distribution with degrees of freedom:

$$\nu \quad \approx \quad \frac{\left( \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2 \nu_1} + \frac{s_2^4}{N_2^2 \nu_2}}.$$

where $\nu_i = n_i - 1$, $i=1,2$

# Student’s *t*-test and Welch’s *t*-test

| $n_1$ | $n_2$ | $s_1$ | $s_2$ | ***t*-test** * | **Unequal** * |
|---|---|---|---|---|---|
| 11 | 11 | 1 | 1 | 0.052 | 0.051 |
| 11 | 11 | 4 | 1 | 0.064 | 0.054 |
| 11 | 21 | 1 | 1 | 0.052 | 0.051 |
| 11 | 21 | 4 | 1 | 0.155 | 0.051 |
| 11 | 21 | 1 | 4 | 0.012 | 0.046 |
| 25 | 25 | 1 | 1 | 0.049 | 0.049 |
| 25 | 25 | 4 | 1 | 0.052 | 0.048 |

* Type I error rate for the *t*-test and unequal variance *t*-test with nominal type I error of 0.05

When **sample sizes are unequal**, the Type I error probabilities of the Student’s *t*-test is decidedly influenced by unequal variances. Similar results have been found for type II error probabilities and statistical power

# Take home message

- Student's *t*-test is robust under violation of homogeneity of variance provided sample sizes are equal.

- When sample size are unequal the type I error, type II error and statistical power of the Student’s *t*-test are decidedly influenced by unequal variances.

- Even when the variances are identical, the Welch’s *t*-test performs well in terms of type I error, type II error and statistical power.

# Take home message

- Unless an argument based on logical, physical, or biological grounds can be made as to why the variances are very likely to be identical for the two populations, the Welch’s *t*-test should be applied.

- It is *not recommended* to pre-test for equal variances and then choose between Student's *t*-test or Welch's *t*-test [*] .
  Graphical display is *recommended* to qualitatively evaluate the difference between sample variances.

[*] Zimmerman DW. A note on preliminary tests of equality of variances. Br J Math Stat Psychol. 2004 May;57(Pt 1):173-81. doi: 10.1348/000711004849222. PMID: 15171807

# If the assumption of normality of the underlying populations is violated?

# Wilcoxon rank-sum test

**Assumptions:** 1. data are ordinal or continuous
2. random sampling from the two populations
3. independent observations $x_i$, $i=1,\ldots,n_1$ and $y_j$, $j=1,\ldots,n_2$

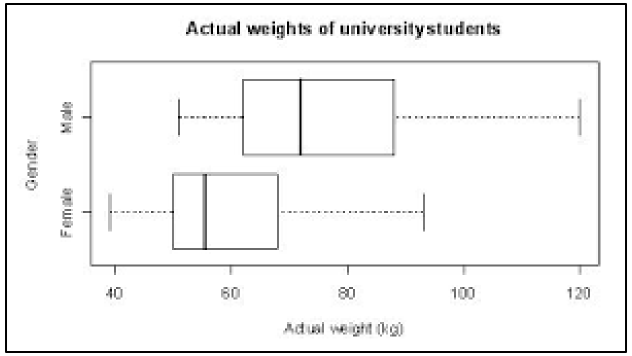
**Hypotheses to test:**

$H_0$: the population distributions are the same ($G=F$).

$H_1$: $G \neq F$ (two-sided $H_1$) or $G < F$ ⋆ (one-sided $H_1$) or $G > F$ ° (one-sided $H_1$).

⋆ G is shifted to the left of F

° G is shifted to the right of F



**Test statistic:** sum of the ranks from one of the two groups.

# Calculation of the test statistic

| ID mouse | Group | Outcome | Rank | Sum rank | Average rank | Sum of ranks | |
|---|---|---|---|---|---|---|---|
| 1 | A | 0 | 1 | 10 | 2.5 | Group A: | 162.5 |
| 5 | A | 0 | 2 | | | Group B: | 302.5 |
| 8 | A | 0 | 3 | | | | |
| 14 | A | 0 | 4 | | | | |
| 6 | A | 1 | 5 | 45 | 7.5 | | |
| 9 | A | 1 | 6 | | | | |
| 11 | A | 1 | 7 | | | | |
| 12 | A | 1 | 8 | | | | |
| 15 | A | 1 | 9 | | | | |
| 21 | B | 1 | 10 | | | | |
| 2 | A | 2 | 11 | 50 | 12.5 | | |
| 7 | A | 2 | 12 | | | | |
| 24 | B | 2 | 13 | | | | |
| 25 | B | 2 | 14 | | | | |
| 3 | A | 3 | 15 | 126 | 18 | | |
| 13 | A | 3 | 16 | | | | |
| 16 | B | 3 | 17 | | | | |
| 20 | B | 3 | 18 | | | | |
| 23 | B | 3 | 19 | | | | |
| 26 | B | 3 | 20 | | | | |
| 29 | B | 3 | 21 | | | | |
| 4 | A | 4 | 22 | 120 | 24 | | |
| 17 | B | 4 | 23 | | | | |
| 18 | B | 4 | 24 | | | | |
| 19 | B | 4 | 25 | | | | |
| 28 | B | 4 | 26 | | | | |
| 22 | B | 5 | 27 | 55 | 27.5 | | |
| 30 | B | 5 | 28 | | | | |
| 27 | B | 7 | 29 | 29 | 29 | | |
| 10 | A | 8 | 30 | 30 | 30 | | |

# Distribution of the test statistic

| **Group 1, ranks** | **3,4,5** | **2,4,5** | **1,4,5** | **2,3,5** | **1,3,5** |
|---|---|---|---|---|---|
| Test statistic | 12 | 11 | 10 | 10 | 9 |
| Probability under $H_0$ | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |

| **Group 1, ranks** | **2,3,4** | **1,3,4** | **1,2,4** | **1,2,3** | **1,2,5** |
|---|---|---|---|---|---|
| Test statistic | 9 | 8 | 7 | 6 | 8 |
| Probability under $H_0$ | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |

$n_1 = 3$

$n_2 = 2$

- **Simulation**: rank j as the same probability to be assigned to one group or the other.

- For large samples, a **normal approximation** with known mean and variance can be applied.

# Distribution-free tests vs *t*-tests

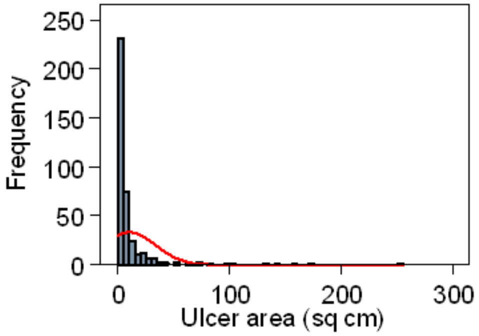
Situations which may suggest the use of distribution-free tests:

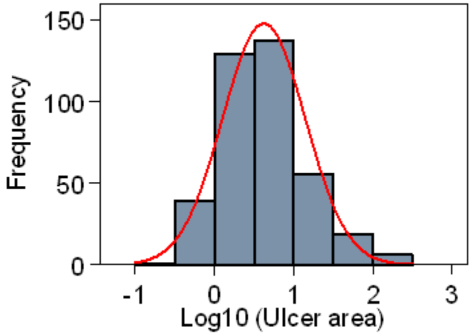
1. When one outcome has a **distribution other than normal**.
2. When the data are **ordered** with many ties or are rank ordered.
3. When the data has **notable outliers**.
4. When there is a **small sample size**.

# Data transformations

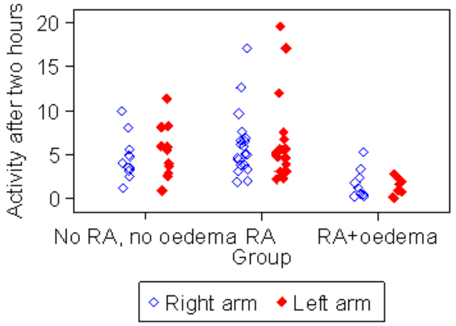
We can transform the data mathematically…

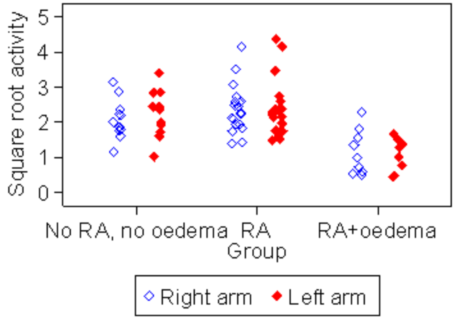
- to make them fit the normality more closely



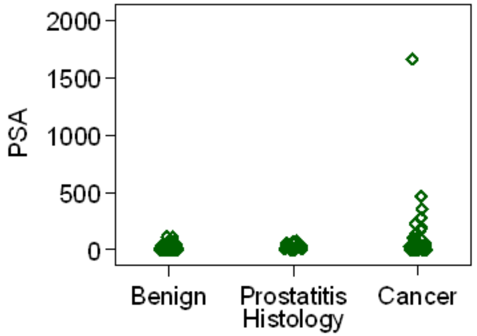


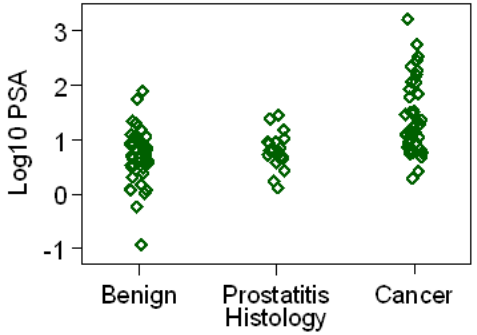
- to obtain more similar variances





- to handle outliers





# The most common used transformations

We can transform the data mathematically into…

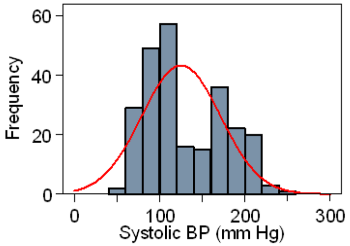
1. the logarithm ($x_i > 0$, i=1,…n)
2. the square root ($x_i \geq 0$, i=1,…n)
3. the reciprocal ($x_i > 0$, i=1,…n)

**Take home message:**

- These transformations could be useful to obtain normality, similar variance and handling outliers

- The best choice depends on the relationship between variability and mean. Graphical display of data is useful to choose the best transformation

- Not all data can be transformed successfully



# Hypothesis to be tested after data transformations

**Assumptions:** 1. Student’s *t*-test assumptions or
2. Welch’s *t*-test assumptions

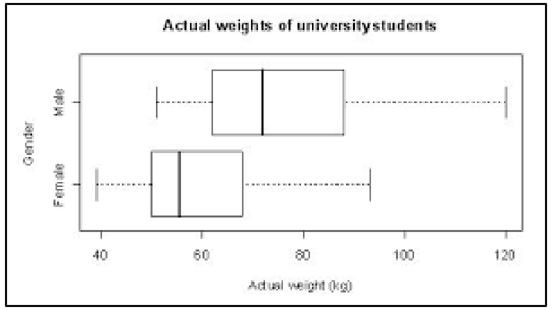
**Hypotheses to test:**

$H_0$: The population distributions are the same (G=F) ⋆⋆

$H_1$: G ≠ F (two-sided $H_1$) or G < F ⋆ (one-sided $H_1$) or G > F ° (one-sided $H_1$)

⋆ G is shifted to the left of F

° G is shifted to the right of F

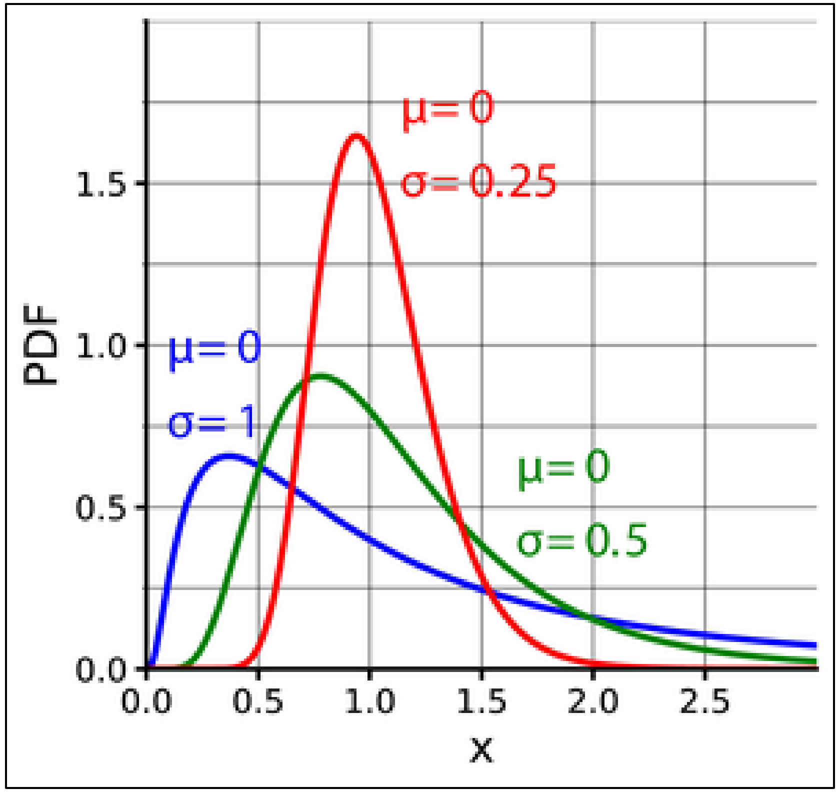


⋆⋆ Previous data transformations are monotonic. Hence,
G=F on the natural scale if and only if G=F on the transformed scale

**Test statistic:** Student’s test statistic or Welch’s test statistic

# Hypothesis to be tested after data transformations

**Log-normal distribution**



Properties of the log-normal distribution

- Mean log-normal: exp($\boldsymbol{\mu} + \boldsymbol{\sigma}^2 / 2$)
- Median log-normal: exp($\boldsymbol{\mu}$)

Consequences

- If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ then Median$_1$ = Median$_2$
- If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}_1 \neq \boldsymbol{\sigma}_2$ then Mean$_1$ ≠ Mean$_2$

# Exercises

Exercises

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html

Shiny web application

https://bioinformatics.cruk.cam.ac.uk/stats/shinystats/