Confidence intervals and Statistical tests

Data analysis course

• Hub team - Pascal Campagne and Steven Volant
Overview

2 sessions:

- **Confidence intervals:** How to construct a confidence interval and interpretation
- **Statistical tests:** What is it? How to interpret?
Overview

2 sessions:

- **Confidence intervals:** How to construct a confidence interval and interpretation
- **Statistical tests:** What is it? How to interpret?

R slide: something is running on RStudio

Turn on your computer and open RStudio, it is your turn!

Course slide: listen to me ;-) 

Feel free to ask questions at any time

Theoretical content
Aim

Confidence intervals:
We draw a sample from a huge population and we want to estimate a quantitative characteristic of this population and how precise is the measure.

Some examples:
- Frequency of a gene in the human population
- Average milk production of cows for a given breed
- Number of faulty parts in the cars of the Renault brand

Statistical tests:
We want to compare some results obtained from one or several samples or to a given value.

Examples:
- Is the frequency of a gene different between autism cases and the general population?
- Clinical trials: Is response different between the placebo group and the treatment group
Confidence interval

Data analysis course
Prediction interval

1. Select 10 circles randomly
2. Calculate the mean
Prediction interval

1. Select 10 circles randomly
2. Calculate the mean

\[ PI_{95\%} = [2.37; 4.62] \]

1. Try it yourself!
Confidence intervals

« Les biologistes connaissant la valeur exacte de la moyenne et de la variance du caractère qu’ils étudient dans une population d’organismes vivants sont en effet comme les orangers sur le sol irlandais : on n’en verra jamais. » (Statistiques pour statophobes)

« Biologists knowing the exact value of the mean and variance of the character they are studying in a population of organisms are like the orange trees on Irish soil: we will never see it. »
Sampling

Population

Sampling

Sample

**Principle:**
The population $P$ is often too big, we select a subset $S$ of individuals from the population.

**Sampling:** $P \rightarrow S$

**Inferring:** $S \rightarrow P$
Get information on the population from the observations.
Confidence intervals

Sampling fluctuations

Sample 1 \(\hat{\theta}_1\)

Sample 2 \(\hat{\theta}_2\)

Sample 3 \(\hat{\theta}_3\)

\(\hat{\theta}_1 \neq \hat{\theta}_2 \neq \hat{\theta}_3\)
Confidence intervals

Sampling fluctuations

Questions:
Can we trust the estimation obtained from one sample?
How much?

Confidence intervals
Confidence intervals

Method:
We defined two bounds $B_1$ and $B_2$ such as:

$$P(B_1 \leq \theta \leq B_2) = 1 - \alpha$$

Where $\alpha$ is the confidence interval threshold (e.g. 5%). The $(1 - \alpha)$ IC is then defined by:

$$IC_{1-\alpha}(\theta) = [b_1; b_2]$$

Usual form of a confidence interval: Estimate ± Error margin

the width of the interval provides a measure on the strength of the evidence supporting the hypothesis that the $\mu$ is close to our estimate $x$. 

12 | Hub team | Confidence intervals and Statistical tests | 21-22/01/2019

Institut Pasteur  C3BI  USR 3756
Confidence intervals

Legends about the confidence intervals

- Two different samples provide the same confidence interval
- We can only compute the interval of the mean
- A confidence interval contains the true value of the parameter

Confidence interval **depends on the estimation and its variance**

Confidence interval can be computed on **all the parameters**

Confidence interval represents an area where the true value **probably** is.
Confidence interval of the mean

Reminder on the Central Limit Theorem

Population distribution

<table>
<thead>
<tr>
<th>Normal</th>
<th>Skewed</th>
<th>Uniform</th>
<th>Irregular</th>
</tr>
</thead>
</table>

Sampling distribution of sample mean

<table>
<thead>
<tr>
<th>n = 3</th>
<th>n = 5</th>
<th>n = 10</th>
<th>n = 20</th>
</tr>
</thead>
</table>

Conclusion:
Whatever the population distribution, if the sample size is enough and the variance finite, the distribution of the mean is gaussian

Source: Points of significance, Nature
Confidence interval of the mean

3 different cases

**Huge sample**
\[ n > 30 \]
- The distribution of the mean is gaussian

**Small sample**
\[ n < 30, \text{ gaussian} \]
- The mean is distributed as a student distribution

**Small sample**
\[ n < 30, \text{ non-gaussian} \]
- Need an approximation or determine confidence interval computationally
Confidence interval of the mean

Assumption:
We assume that the sample size is huge.

Given that the mean $\overline{X}$ follows a gaussian distribution, we have:

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

$$\Pr\left\{ u_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq u_{1-\alpha/2} \right\} = 1 - \alpha.$$  

$$\Pr\{ \overline{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \} = 1 - \alpha.$$  

Where $u_\rho$ is the $\rho$-order quantile of a standard gaussian distribution (mean = 0 and sd = 1)
Confidence interval of the mean

Assumption:

We assume that the sample size is huge.

\[
IC_{1-\alpha}(\mu) = \left[ \overline{x} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} ; \overline{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]
\]

Exemple: lungA dataset

Compute the confidence interval of the mean of the pollution
Confidence interval of the mean

```r
## Load the data
data = read.csv("Documents/Cours_stat/lungA.csv")

## Number of observations
n = nrow(data)

## Get the quantile
u = qnorm(0.975)

## Parameter estimation
mean.pollution = mean(data$pollution)
sd.pollution = sd(data$pollution)

## Confidence interval (95%)
ICsup = mean.pollution + u*sd.pollution/sqrt(n)
ICinf = mean.pollution - u*sd.pollution/sqrt(n)

## print the result
cat(paste("95% confidence interval = [",round(ICinf,2),",";","round(ICsup,2),"]"))
```

95.4% between -2\(\sigma\) and 2\(\sigma\)
Confidence interval of the mean

**Assumption:**
We assume that the sample size is huge.

\[ IC_{1-\alpha}(\mu) = \left[ \bar{x} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} ; \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \]

**Exemple:** lungA dataset

Compute the confidence interval of the mean of the pollution

We find:

\[ IC_{95\%}(\mu_{\text{pollution}}) = [67.09; 73.94] \]
Confidence interval of the mean

Your turn

- Dataset
  Description of the dataset: *malaria_longitudinal_data_simul.csv*
Confidence interval of the mean

**Exercises**

- **Dataset**
  Description of the dataset: *malaria_longitudinal_data_simul.csv*

- **temp variable**
  Considering the sick individuals (pv.lm = ‘yes’) of the Control group, compute the 95% ICs of the mean of the temperature for week 1 and 12

- **haem.hb.level variable**
  Compute the 90% IC of the hemoglobin level in the 3 groups at week 1.
**Confidence interval of the mean**

**Answers**

- **temp variable**
  Considering the sick individuals (pv.lm = ‘yes’) of the Control group, compute the 95% ICs of the mean of the temperature for week 1 and 12

  \[
  IC_{95\%}^{(1)} = [38.38; 38.64] \quad IC_{95\%}^{(12)} = [38.45; 38.73]
  \]

- **haem.hb.level variable**
  Compute the 90% IC of the hemoglobin level in the 3 groups at week 1.

  \[
  IC_{90\%}^{\text{control}} = [9.94; 10.68] \quad IC_{90\%}^{\text{art}} = [9.87; 10.57] \quad IC_{90\%}^{\text{artpq}} = [9.54; 10.34]
  \]
Confidence interval of the mean

Assumption:

We assume that the sample size is small and that the distribution is gaussian

Given that X is gaussian, the mean $\overline{X}$ is also gaussian

Problem: when the number of observations is small, we get a bad estimation of the variance

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim t_{n-1}$$

Where $t$ is the student distribution with parameter n-1 corresponding to the degree of freedom

History:
The Student distribution has been studied by William Gosset (aka Student) and then by R.A. Fisher. Hence we use letter $t$ for the Student distribution…
Confidence interval of the mean

**Assumption:** We assume that the sample size is **small** and that the distribution is **gaussian**.

Given that $X$ is gaussian, the mean $\bar{X}$ is also gaussian.

**Problem:** when the number of observations is small, we get a bad estimation of the variance.

\[
\frac{X - \mu}{\sigma / \sqrt{n}} \sim t_{n-1}
\]
Confidence interval of the mean

Assumption:
We assume that the sample size is small and that the distribution is gaussian

\[ IC_{1-\alpha}(\mu) = \left[ \bar{x} - t_{n-1,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} ; \bar{x} + t_{n-1,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \]

Exemple: Viscosity

Data: \( x_i = 78, 84, 91, 76, 79, 71, 83, 84, 75, 90 \);
Confidence interval of the mean

```
# 95% IC of viscosity
vi = c(78, 84, 91, 76, 79, 71, 83, 84, 75, 90)

## Get the quantile
# t = qt(0.975, 9)

## Parameter estimation
mean.vi = mean(vi)
sd.vi = sd(vi)

## Confidence interval (95%)
ICsup = mean.vi + t*sd.vi/sqrt(10)
ICinf = mean.vi - t*sd.vi/sqrt(10)
cat(paste("95% confidence interval = [",round(ICinf,2),",";",round(ICsup,2),"]"))
```
Confidence interval of the mean

Assumption:
We assume that the sample size is small and that the distribution is gaussian

\[ IC_{1-\alpha}(\mu) = \left[ \bar{x} - t_{n-1,1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}}; \bar{x} + t_{n-1,1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}} \right] \]

Exemple: Viscosity

Data: \( x_i = 78, 84, 91, 76, 79, 71, 83, 84, 75, 90 \);
\( \bar{x} = 81.1; \widehat{\sigma} = 6.47 \)
\( t_{9,0.975} = 2.26 \)
\( IC_{95\%}(\mu) = [76.5; 85.7] \)
Confidence interval of the mean

Your turn

- haem hb level variable
Dataset: malaria_longitudinal_data_simul.csv
Considering the individuals of the ART group coming from the UTAMUP village, compute the 95% ICs of the mean of the hemoglobin level for weeks 1 and 12
Confidence interval of the mean

Answers

- **haem.hb.level variable**

Considering the individuals of the ART group coming from the UTAMUP village, compute the 95% ICs of the mean of the hemoglobin level for weeks 1 and 12

\[
IC_{95\%}^{(1)} = [8.42; 11.76] \quad \quad IC_{95\%}^{(12)} = [8.24; 11.74]
\]
Confidence interval of the mean

Assumption:
We assume that the sample size is **small** and that the distribution is **not** gaussian

Best case:
The distribution is « almost » gaussian (at least symmetric), the Student distribution can be used as an approximation.

Always keep in mind that it is an approximation!

Worst case:
The distribution is far from a gaussian (Poisson, Negative binomial …), the Student distribution can not be used.

Redo your experiment hundreds or thousands of times
OR
Use bootstrapping!
Bootstrapping approach

Population

Sample

Sub-sample

\( \hat{\theta}_1 \)

Sub-sample

\( \hat{\theta}_2 \)

Sub-sample

\( \hat{\theta}_B \)

Sampling

Resampling with replacement (\( B \) times)

Estimation
Confidence interval of the mean

Example:

We are interested in the number of patients per day in Necker hospital. Hereafter are the measurement over 20 days:

\[ x_i = 43, 55, 45, 46, 57, 52, 61, 47, 52, 57, 48, 57, 52, 62, 50, 53, 72, 58, 58, 45. \]

By definition, we know that this is Poisson distributed.
Confidence interval of the mean

Example:

We are interested in the number of patients per day in Necker hospital. Hereafter are the measurement over 20 days:
\[ x_i = 43, 55, 45, 46, 57, 52, 61, 47, 52, 57, 48, 57, 52, 62, 50, 53, 72, 58, 58, 45. \]

By definition, we know that this is Poisson distributed.

Non symmetric
Too far from gaussian

Bootstrap
Confidence interval of the mean

```r
## data
x = c(43, 55, 45, 46, 57, 52, 61, 47, 52, 57, 48, 57, 52, 62, 50, 53, 72, 58, 58, 45)
hist(x,breaks=3)

## Number of resampling
B = 5000

## Resampling
estim = c()
for(i in 1:B)
{
    subsamp = sample(x,size = 20,replace = TRUE)
    estim = c(estim, mean(subsamp))
}

## Representation (histogram)
hist(estim,main = "Mean distribution (B=1000)"

## Get the CI
ICinf = quantile(estim,probs = 0.025)
ICsup = quantile(estim,probs = 0.975)
cat(paste("95% confidence interval = [",round(ICinf,2),";",round(ICsup,2),"]"))
```
Confidence interval of the mean

Exemple:

We are interested in the number of patients by day in Necker hospital. Hereafter are the measurement over 20 days:
\[ x_i = 43, 55, 45, 46, 57, 52, 61, 47, 52, 57, 48, 57, 52, 62, 50, 53, 72, 58, 58, 45. \]

The confidence interval is then obtained by the quantiles of the distribution:

\[ IC_{95\%}(\mu) = [50.75; 56.65] \]
Your turn

- **haem.hb.level variable**
  Considering the individuals of the ART group coming from the UTAMUP village, compute the 95% ICs of the mean of the hemoglobin level for week 1 (use bootstrapping)

Remind, with the student distribution: 

\[ IC_{95\%}^{(1)} = [8.37; 11.81] \]
Answer

- **haem.hb.level variable**

  Considering the individuals of the ART group coming from the UTAMUP village, compute the 95% ICs of the mean of the hemoglobin level for week 1 (use bootstrapping)

  Remind, with the student distribution:

  \[
  IC^{(1)}_{95\%} = [8.37; 11.81]
  \]

  Using bootstrapping:

  \[
  IC^{\text{boot}}_{95\%} = [8.62; 11.49]
  \]
Confidence interval of the mean

To summarize

**Huge sample**  
$n > 30$

Use the gaussian distribution

$$IC_{1-\alpha}(\mu) = \left[ \bar{x} \pm \frac{u_{1-\alpha/2}}{\sqrt{n}} \right]$$

**Small sample**  
$n < 30$, gaussian

Use the student distribution

$$IC_{1-\alpha}(\mu) = \left[ \bar{x} \pm \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} \right]$$

**Small sample**  
$n < 30$, non-gaussian

If almost gaussian

$$IC_{1-\alpha}(\mu) = \left[ \bar{x} \pm \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} \right]$$

else

Bootstrap

**Remark**: Bootstrapping can also be used to determine the CI for other parameters (median, …)
Confidence interval of a proportion $p$

Assumption:

We assume that the sample size is huge ($np > 5$ and $n(1-p) > 5$)

\[ \text{Proportion} = \frac{\sum n \text{ Bernoulli variables}}{n} = \frac{\text{Binomial}}{n} \]

For huge samples, the binomial distribution can be approximated by a Normal distribution with mean $np$ and variance $np(1-p)$

\[ \hat{p} \sim_{app} N(p; p(1-p)/n) \]

So,

\[ IC_{1-\alpha}(p) = \left[ \hat{p} - u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] \]
**Confidence interval of a proportion p**

**Assumption:**

We assume that the sample size is huge \((np > 5 \text{ and } n(1-p) > 5)\)

\[
IC_{1-\alpha}(p) = \left[ \hat{p} - u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]
\]

**Example:** 95% confidence interval of the smoker proportion (LungA)
Confidence interval of a proportion $p$

```r
## Get the data
x = data$smoker

## Get the size
n = length(x)

## Get the quantile
u = qnorm(0.975)

## Parameter estimation
prop = sum(x)/n
sd.prop = sqrt(prop*(1-prop)/(n))

## Confidence interval (95%)
ICsup = prop + u*sd.prop
ICinf = prop - u*sd.prop
cat(paste("95% confidence interval = [",round(ICinf,2),",";",round(ICsup,2),"]"))
```
Confidence interval of a proportion $p$

**Assumption:**
We assume that the sample size is **huge** ($np > 5$ and $n(1-p) > 5$)

$$IC_{1-\alpha}(p) = \left[ \hat{p} - u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + u_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

**Example:** 95% confidence interval of the smoker proportion (LungA)

We have: $\hat{p} = \frac{59}{93} \approx 0.634$

Then,

$$IC_{95\%}(p) = [0.54; 0.73]$$
Confidence interval of a proportion $p$

**Your turn**

- **Malaria data**
  Compare the 95% IC for the proportion of patients (pv.lm = ‘yes’) in the ART&PQ group between week 1 and 12.
Confidence interval of a proportion $p$

Your turn

- **Malaria data**
  Compare the 95% IC for the proportion of patients (pv.lm = ‘yes’) in the ART&PQ group between week 1 and 12

\[
IC_{95\%}^{(1)} = [0.43; 0.61] \quad \text{and} \quad IC_{95\%}^{(12)} = [0.22; 0.39]
\]
Confidence interval of a proportion $p$

**Assumption:**

We assume that the sample size is small ($np < 5$ or $n(1-p) < 5$)

1. Compute all the probabilities from the binomial distribution

2. Use the binomial table (probabilities already calculated)

### Table

<table>
<thead>
<tr>
<th>$p$</th>
<th>$k$</th>
<th>0,05</th>
<th>0,10</th>
<th>0,15</th>
<th>0,20</th>
<th>0,25</th>
<th>0,30</th>
<th>0,35</th>
<th>0,40</th>
<th>0,45</th>
<th>0,50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.5987</td>
<td>0.3487</td>
<td>0.1969</td>
<td>0.1074</td>
<td>0.0563</td>
<td>0.0282</td>
<td>0.0135</td>
<td>0.0060</td>
<td>0.0025</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.9139</td>
<td>0.7361</td>
<td>0.5443</td>
<td>0.3758</td>
<td>0.2440</td>
<td>0.1493</td>
<td>0.0860</td>
<td>0.0464</td>
<td>0.0233</td>
<td>0.0107</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.9885</td>
<td>0.9298</td>
<td>0.8202</td>
<td>0.6778</td>
<td>0.5256</td>
<td>0.3828</td>
<td>0.2616</td>
<td>0.1673</td>
<td>0.0996</td>
<td>0.0547</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.9990</td>
<td>0.9872</td>
<td>0.9500</td>
<td>0.8791</td>
<td>0.7759</td>
<td>0.6496</td>
<td>0.5138</td>
<td>0.3823</td>
<td>0.2660</td>
<td>0.1719</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.9999</td>
<td>0.9984</td>
<td>0.9901</td>
<td>0.9672</td>
<td>0.9219</td>
<td>0.8497</td>
<td>0.7515</td>
<td>0.6331</td>
<td>0.5044</td>
<td>0.3770</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9986</td>
<td>0.9936</td>
<td>0.9803</td>
<td>0.9527</td>
<td>0.9051</td>
<td>0.8338</td>
<td>0.7384</td>
<td>0.6230</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9991</td>
<td>0.9965</td>
<td>0.9894</td>
<td>0.9740</td>
<td>0.9452</td>
<td>0.8980</td>
<td>0.8281</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9996</td>
<td>0.9984</td>
<td>0.9952</td>
<td>0.9877</td>
<td>0.9726</td>
<td>0.9453</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9995</td>
<td>0.9983</td>
<td>0.9955</td>
<td>0.9893</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Confidence interval of a proportion $p$

**Your turn**

- **Proportion of patients**
  Dataset: malaria_longitudinal_data_simul.csv
  Compute the 95% IC for the proportion of patients ($pv.lm = \text{‘yes’}$) in the ART&PQ group for week 12

Reminder (with gaussian approximation):

$$IC^{(12)}_{95\%} = [0.22; 0.39]$$
Answer

- **Proportion of patients**

Dataset: malaria_longitudinal_data_simul.csv

Compute the 95% IC for the proportion of patients (pv.lm = ‘yes’) in the ART&PQ group for week 12

Reminder (with gaussian approximation):

\[ IC_{95\%}^{(12)} = [0.22; 0.39] \]

Result without the gaussian approximation:

\[ IC_{95\%}^{(12)} = [0.22; 0.39] \]

```r
## Confidence interval (95%)
ICsup = qbinom(0.975,n,prop)/n
ICinf = qbinom(0.025,n,prop)/n
```

Remark: Here we find exactly the same interval but sometimes it can differ a bit.
Size effect

Source: rpsychologist.com
How to choose n?

Setting accuracy

We set the desired precision (i.e. the margin error of the confidence interval)

For instance: we want a margin error approximately equal to $M$ for a confidence interval of a proportion

Formula:

$$n = \left( \frac{u_{1-\alpha/2}}{M} \right)^2 \times p(1-p) + 1$$

Example:

We want to estimate the smoker proportion with a 1% margin ($M = 0.01$) with a 95% probability.

Remind: \[ \hat{p} = \frac{59}{93} \approx 0.634 \]  

\[ n \approx 358 \]
Conclusion on confidence intervals

- Confidence intervals provide a precision on the estimations
- They must be calculated for each estimated parameter
- Formulas depend on the sample size and the distribution
Warning on confidence intervals

Don’t be confused by confidence interval …

Compute a confidence interval using the estimation $\hat{\theta}$ in which the true value $\theta$ has a given probability (e.g. 95%) to belong

… and sampling interval

Compute a sampling interval using the true value $\theta$ in which the estimated value $\hat{\theta}$ has a given probability (e.g. 95%) to belong
Exercises (optional)

- **Mean of the haem.hb.level variable**
  1. Compare the 95% IC of the mean of the hemoglobin for the patients (pv.lm = ‘yes’) in the ART group between week 1 and 12
  2. Compute the same intervals only for village = « UTAMUP »

- **Proportions**
  1. Compute the proportion of people from the UTAMUP village among the patients (pv.lm = ‘yes’) of the ART&PQ group in week 12, use gaussian approximation and binomial distribution
Answers

• Mean of the haem.hb.level variable
  1. Compare the 95% IC of the mean of the hemoglobin for the patients (pv.lm = ‘yes’) in the ART group between week 1 and 12

\[ IC_{95\%}^{(1)} = [8.9; 9.81] \quad IC_{95\%}^{(12)} = [9.03; 10.16] \]

  1. Compute the same intervals only for village = « UTAMUP ».

\[ IC_{95\%}^{(1)} = [6.12; 10.8] \quad IC_{95\%}^{(12)} = [6.92; 10.04] \]

• Proportions
  1. Compute the 90% IC of the proportion of people from the UTAMUP village among the patients (pv.lm = ‘yes’) of the ART&PQ group in week 12, use gaussian approximation and binomial distribution

\[ IC_{90\%}^{\text{gauss}} = [0.08; 0.3] \quad IC_{90\%}^{\text{binom}} = [0.08; 0.31] \]
Statistical testing

Data analysis course
Remember Confidence Intervals...

I received a sample of 10 adult mice (females) of a given strain/line. I know that a mass body mass different from 23g may reflect abnormal rearing conditions.

Does the mean mass differs from this standard mass of 23g?
I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

$\mu_{WT} = 7.03$

$\mu_{M} = 5.34$
Context 2/2

I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

**YES:** the two populations of virus are characterised by different means

**NO:** …?
Context 2/2

I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

YES: the two populations of virus are characterised by different means

NO: population means are the same

→ the observed difference $\mu_M - \mu_{WT}$ is due to random fluctuations bound to sampling
I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>WT</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>Mutant</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>Mutant</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>
Context

I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>WT</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>Mutant</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>Mutant</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>Mutant</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>Mutant</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>WT</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>WT</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>
**Context**

I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>WT</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>Mutant</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>Mutant</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>Mutant</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>Mutant</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>WT</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>WT</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>Mutant</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>Mutant</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>WT</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>WT</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>
I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>WT</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>Mutant</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>Mutant</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>WT</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>WT</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>WT</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>Mutant</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>Mutant</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>Mutant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Values</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55</td>
<td>WT</td>
</tr>
<tr>
<td>2</td>
<td>3.14</td>
<td>Mutant</td>
</tr>
<tr>
<td>3</td>
<td>3.73</td>
<td>Mutant</td>
</tr>
<tr>
<td>4</td>
<td>4.32</td>
<td>Mutant</td>
</tr>
<tr>
<td>5</td>
<td>6.91</td>
<td>Mutant</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>WT</td>
</tr>
<tr>
<td>7</td>
<td>7.8</td>
<td>WT</td>
</tr>
<tr>
<td>8</td>
<td>6.68</td>
<td>WT</td>
</tr>
</tbody>
</table>

Calculate the difference $\mu_M - \mu_{WT}$ in all tables...
I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?
Context

I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Intuitively, how could we tackle the problem?

“risk” = 0.131

approx. 13% differences obtained with permutations

> observed difference

[Histogram showing permuted and observed differences]
I want to compare the viral load in two strains of virus (WT and mutant)

Does the mean viral load significantly differ among strains?

Testing a difference consists in using a standardised mathematical model of the distribution below.
Statistical tests

Population A
- Random sampling
- Sample
- Estimation
  \( \hat{\theta}_A \)

Population B
- Random sampling
- Sample
- Estimation
  \( \hat{\theta}_B \)

\[ =, >, <, \neq \]

?
Statistical tests

Definition:

A statistical test provides a mechanism for **making quantitative decisions at the population level** based on a sample of observations. The intent is to determine whether there is **enough evidence to «reject»** a hypothesis about the process.
Definition:

A statistical test provides a mechanism for making quantitative decisions at the population level based on a sample of observations. The intent is to determine whether there is enough evidence to « reject » a hypothesis about the process.

i.e. a statistical test helps decide if something is « reasonably unlikely » to be true.
Definition:

A statistical test provides a mechanism for making quantitative decisions at the population level based on a sample of observations. The intent is to determine whether there is enough evidence to «reject» a hypothesis about the process.

_i.e._ a statistical test helps decide if something is «reasonably unlikely» to be true.

It is a little bit like innocence presumption, the judge will not reject the innocence unless clear evidence against it.

This course _will not_ provided you:
- an overview of all the statistical tests
Statistical tests

**Exhaustive list** of statistical tests for a biologist

- Chi2
- Student (t-test)
- Fisher
- Mann-Whitney
Statistical tests

**Exhaustive list** of statistical tests for a biologist

- Chi2
- Student (t-test)
- Fisher
- Mann-Whitney

**Non-exhaustive list** of statistical tests for a statistician

- Skillings
- Box
- Friedman
- Bartlett
- Chi2
- Dixon
- Page
- Mood
- Triangle
- Correlation
- ANOVA
- Mann
- Mack
- Shapiro
- Mantel
- Welch
- Student
- Kullback
- Mann-Whitney
- Grubbs
- Monte-Carlo
- Fisher
- Wilcoxon
- Cochran
- Levene
- Kolmogorov-Smirnov
- Kruskall
- Durbin
- McNemar
- Hypergeometric
- Mann-Kendall
- Skillings
- Page
Statistical tests

Definition:

A statistical test provides a mechanism for making quantitative decisions at the population level based on a sample of observations. The intent is to determine whether there is enough evidence to «reject» a hypothesis about the process.

This course will not provide you:
- an overview of all the statistical tests
- a recipe to use the appropriate test
  (depends on your data)
Statistical tests

Exhaustive list of experiments for a statistician
Statistical tests

**Exhaustive list** of experiments for a statistician

Non-exhaustive list of experiments for a biologist

- PCR
- HDX-MS
- DNA-seq
- ChIP-seq
- RNA-seq
- microarray
- Proteomic
Definition:

A statistical test provides a mechanism for **making quantitative decisions at the population level** based on a sample of observations. The intent is to determine whether there is **enough evidence to «reject»** a hypothesis about the process.

This course **will not** provide you:
- an overview of all the statistical tests
- a recipe to use the appropriate test (depends on your data)
- a magic test to get your p-value under 5%
Statistical tests

Definition:
A statistical test provides a mechanism for **making quantitative decisions at the population level** based on a sample of observations. The intent is to determine whether there is **enough evidence to «reject»** a hypothesis about the process.

This course **will not** provided you:
- an overview of all the statistical tests
- a recipe to use the appropriate test (depends on your data)
- a magic test to get your p-value under 5%

This course **will** provided you:
- the general principle of a statistical test
- things you must be careful of
- correct interpretation of the results
General principle

1. Choose the test according to your data
2. Define the hypotheses (H0 & H1)
3. Select the level $\alpha$
4. Compute the test statistic
5. Compare the value to the threshold
6. Conclude
General principle

0. Choose the test according to your data
1. Define the hypotheses (H0 & H1)
2. Select the level $\alpha$

Before data collection
After data collection

3. Compute the test statistic
4. Compare the value to a threshold
5. Conclude
Example

**Aim:**

Compare the mean of the tumor size according to the sex.
Example

Aim:

Compare the mean of the tumor size according to the sex.

Significant difference?
Example

0. Choose the test according to your data

1. Define the hypotheses (H0 & H1)

2. Select the level $\alpha$

3. Compute the test statistic

4. Compare the value to a threshold

5. Conclude
Example: What is the question?

« Is there a difference in true mean tumor size between men and women? »

- **Outcome**: Quantitative, continuous (measurement could range from 0 to any (realistic) positive value)

- **Statistical considerations**: Need a test that can compare the true means of two population through the observed means of the two samples
Example: What is the question?

« Is there a difference in true mean tumor size between men and women? »

- **Outcome**: Quantitative, continuous (measurement could range from 0 to any (realistic) positive value)

- **Statistical considerations**: Need a test that can compare the true means of two population through the observed means of the two samples

Asking yourself these questions helps you define your hypotheses
Example: Statistical hypotheses

A statistical test compares two opposing hypotheses: the null hypothesis (H0) and the alternative hypothesis (H1 or Ha).

Null hypothesis (H0)

This hypothesis states that there is no difference (equality to a value, same distribution…)

Mnemonic trick: H0: 0 difference

Alternative hypothesis (H1)

This hypothesis states that there is a difference (parameter different than the value for H0,… )

Sometimes called Ha or Halt

Note: Hypotheses are made on population-level (unobserved) values, never on sample-level observed values
A statistical test compares two opposing hypotheses: the *null hypothesis* (H0) and the *alternative hypothesis* (H1 or Ha).

**Null hypothesis (H0)**

This hypothesis states that there is no difference (equality to a value, same distribution…)

*Mnemonic trick: H0: 0 difference*

**Alternative hypothesis (H1)**

This hypothesis states that there is a difference (parameter different than the value for H0,… )

*S sometimes called Ha or Halt*

For instance, in our example:

\[
\begin{align*}
H0: & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H1: & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]
A statistical test compares two opposing hypotheses: the **null hypothesis (H0)** and the **alternative hypothesis (H1 or Ha)**.

### Null hypothesis (H0)

This hypothesis states that there is no difference (equality to a value, same distribution…)

**Mnemonic trick:** $H_0$: 0 difference

### Alternative hypothesis (H1)

This hypothesis states that there is a difference (parameter different than the value for H0,…)

*Sometimes called Ha or Halt*

For instance, in our example:

\[
\begin{align*}
H_0 & : \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 & : \mu_{\text{male}} > \mu_{\text{female}}
\end{align*}
\]
Example: Statistical hypotheses

A statistical test compares two opposing hypotheses: the null hypothesis (H0) and the alternative hypothesis (H1 or Ha).

**Null hypothesis (H0)**

This hypothesis states that there is no difference (equality to a value, same distribution…)

Mnemonic trick: H0: 0 difference

**Alternative hypothesis (H1)**

This hypothesis states that there is a difference (parameter different than the value for H0,…)

For instance, in our example:

\[
\begin{align*}
H0 & : \mu_{\text{male}} = \mu_{\text{female}} \\
H1 & : \mu_{\text{male}} \geq \mu_{\text{female}}
\end{align*}
\]
Example: Statistical hypotheses

A statistical test examines two opposing hypotheses: the **null hypothesis (H0)** and the **alternative hypothesis (H1 or Ha)**.

### Null hypothesis (H0)

This hypothesis states that there is no difference (equality to a value, same distribution…)

**Mnemonic trick:** $H_0$: 0 difference

### Why H0 is useful:

If H0 were true, the two samples would come from the same population and observed fluctuations are only the result of randomness.

*In other words:* H0 allows to make predictions (think « fluctuation intervals »)
Choose the test according to your data
Define the hypotheses (H0 & H1)
Select the level $\alpha$
Compute the test statistic
Compare the value to a threshold
Conclude
Example: Statistical hypotheses

Aim:
Compare the mean of the tumor size according to the sex.

Hypothesis:
\[
\begin{align*}
H_0 : & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]

Recall: hypotheses on population-level measures (μ), not on observations (m)
Example

0. Choose the test according to your data
1. Define the hypotheses (H0 & H1)
2. Select the level $\alpha$
3. Compute the test statistic
4. Compare the value to a threshold
5. Conclude
## Type I and type II errors

<table>
<thead>
<tr>
<th>Decision of the test</th>
<th>Null hypothesis (unknown)</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject</td>
<td>Rejecting when it is true</td>
<td>Type I error (Level: $\alpha$)</td>
<td>Correct decision (Power)</td>
</tr>
<tr>
<td>Fail to reject</td>
<td>Correct decision</td>
<td>Fail to reject when it is false</td>
<td>Type II error (Level: $\beta = 1 - \text{Power}$)</td>
</tr>
</tbody>
</table>
Example

Aim:
Compare the mean of the tumor size according to the sex.

Hypothesis:
\[
\begin{align*}
H_0 : & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]

Level:
We set $\alpha = 5\%$
Example

0. Choose the test according to your data
1. Define the hypotheses (H0 & H1)
2. Select the level $\alpha$
3. Compute the test statistic
4. Compare the value to a threshold
5. Conclude
What is a test statistic?

**Definition:**
A standardized value $S$ calculated from sample which is used to reject or not the null hypothesis. **We know the probability distribution of this variable.**

*Recall: when H0 is true, the two samples come from the same population*
Definition:
A standardized value $S$ calculated from sample which is used to reject or not the null hypothesis. **We know the probability distribution of this variable.**

Formula? Nope!
There is no general formula for the test statistic $S$, it depends on the statistical test (which depends on the hypothesis and on the data).
What is a test statistic?

| Definition: A standardized value $S$ calculated from sample which is used to reject or not the null hypothesis. **We know the probability distribution of this variable.** |

| Formula ? Nope ! There is no general formula for the test statistic $S$, it depends on the statistical test (which depends on the hypothesis and on the data) |

| Example: In our example, we will use the classical t-test to compare the two means. **In this case**, the test statistic $S$ is: $S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\sigma^2_{\text{male}}}{n_{\text{male}}} + \frac{\sigma^2_{\text{female}}}{n_{\text{female}}}}}$ |
Example

**Aim:**
Compare the mean of the tumor size according to the sex.

**Hypothesis:**

\[
\begin{align*}
H_0 &: \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 &: \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]

**Level:**
We set $\alpha = 5\%$

---

```r
## Get the data
xmale = data$\text{tumor\_size}[data$\text{gender}=="\text{male}"]
xfemale = data$\text{tumor\_size}[data$\text{gender}=="\text{female}"]

## Parameter estimation
mu.male = mean(xmale); mu.female = mean(xfemale)
sigma2.male = var(xmale); sigma2.female = var(xfemale)
n.male = length(xmale); n.female = length(xfemale)

## Test statistic
S = (mu.male - mu.female)/sqrt(sigma2.male/(n.male) + sigma2.female/(n.female))
```
Example

Aim:
Compare the mean of the tumor size according to the sex.

Hypothesis:
\[ \begin{align*}
H_0 : & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*} \]

Level:
We set \( \alpha = 5\% \)

Test statistic:
\[
S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\sigma^2_{\text{male}}}{n_{\text{male}}} + \frac{\sigma^2_{\text{female}}}{n_{\text{female}}}}} = \frac{5.11 - 2.57}{\sqrt{\frac{4.80}{54} + \frac{4.46}{39}}} = 5.62
\]
Example

0. Choose the test according to your data
1. Define the hypotheses (H0 & H1)
2. Select the level $\alpha$
3. Compute the test statistic
4. Compare the value to a threshold
5. Conclude
Threshold: a critical value

**Definition:**
Value that is compared to the test statistic to determine whether to reject the null hypothesis.

*Answer the following question: Is the test statistic more extreme than would be expected if the null hypothesis were true.*
Definition:
Value that is compared to the test statistic to determine whether to reject the null hypothesis.

Answer the following question: Is the test statistic is more extreme than would be expected if the null hypothesis were true.

Representation:

- **Bilateral**
  - Reject area: $\alpha/2$
  - Non Reject area: $1 - \alpha$

- **Unilateral**
  - Reject area: $\alpha$
  - Non Reject area: $1 - \alpha$

Distribution of the test statistic $S$, if $H0$ true
Threshold: a critical value

Definition:
Value that is compared to the test statistic to determine whether to reject the null hypothesis.

Answer the following question: Is the test statistic is more extreme than would be expected if the null hypothesis were true.

Our example:

\[ S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\hat{\sigma}^2_{\text{male}}}{n_{\text{male}}} + \frac{\hat{\sigma}^2_{\text{female}}}{n_{\text{female}}}}} \sim \text{app } N(0, 1) \]

\[ H_1 : \mu_{\text{male}} \neq \mu_{\text{female}} \]

Bilateral test

\[ \alpha = 5\% \]
Example

**Aim:**
Compare the mean of the tumor size according to the sex.

**Hypothesis:**
\[
\begin{align*}
H_0 : & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]

**Level:**
We set \( \alpha = 5\% \)

**Test statistic:**
\[
S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\hat{\sigma}^2_{\text{male}}}{n_{\text{male}}} + \frac{\hat{\sigma}^2_{\text{female}}}{n_{\text{female}}}}} = \frac{5.11 - 2.57}{\sqrt{\frac{4.80}{54} + \frac{4.46}{39}}} = 5.62
\]

**Comparison to a threshold:**
\[|S| = 5.62 > 1.96\]
Example

0. Choose the test according to your data
1. Define the hypotheses (H0 & H1)
2. Select the level $\alpha$
3. Compute the test statistic
4. Compare the value to a threshold
5. Conclude
What is a p-value?

Definition:
Assuming the null hypothesis is true, a p-value is defined as the probability of obtaining a result (S) equal or more extreme to what is actually observed. It helps you to determine the significance of your result.
What is a p-value?

Definition:
Assuming the null hypothesis is true, a p-value is defined as the probability of obtaining a result (S) equal or more extreme to what is actually observed. It helps you to determine the significance of your result.

\[
p\text{-value} = P_{H_0}(|S| \geq |s|) \quad \text{Bilateral test}
\]

\[
p\text{-value} = P_{H_0}(S \leq |s|) \quad \text{Unilateral test}
\]
How to interpret a p-value

Reminder:
At the beginning of the test, we set a level $\alpha$ (eg 5%) which corresponds to the probability to reject H0 when H0 is actually true.

Interpretation of the result:

- $p$-value $\leq \alpha$
  - We reject the null hypothesis with a probability $\alpha$
- $p$-value $> \alpha$
  - We do not reject the null hypothesis with a probability $\alpha$

A hypothesis is never accepted
I want to compare the viral load in two strains of virus (WT and mutant) does the mean viral load significantly differ among strains? Intuitively, how could we tackle the problem?

P-value

“risk” = 0.131

approx. 13% differences obtained with permutations

> observed difference
Aim:
Compare the mean of the tumor size according to the sex.

Hypothesis:
\[
\begin{cases}
H_0 : \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : \mu_{\text{male}} \neq \mu_{\text{female}}
\end{cases}
\]

Level:
We set \( \alpha = 5\% \)

Test statistic:
\[
S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\sigma^2_{\text{male}}}{n_{\text{male}}} + \frac{\sigma^2_{\text{female}}}{n_{\text{female}}}}} = \frac{5.11 - 2.57}{\sqrt{\frac{4.80}{54} + \frac{4.46}{39}}} = 5.62
\]

Comparison to a threshold:
\(|S| = 5.62 > 1.96\)

\[\# p-value (Normal approx)\]
\[pv.\text{norm} = 2\cdot\text{pnorm}(-S)\]

\[\# p-value (Welch test)\]
\[pv.\text{welch} = 2\cdot\text{pt}(-S,\text{df}=83.76)\]
Aim:
Compare the mean of the tumor size according to the sex.

Hypothesis:
\[
\begin{align*}
H_0 : & \quad \mu_{\text{male}} = \mu_{\text{female}} \\
H_1 : & \quad \mu_{\text{male}} \neq \mu_{\text{female}}
\end{align*}
\]

Level:
We set \( \alpha = 5\% \)

Test statistic:
\[
S = \frac{\hat{\mu}_{\text{male}} - \hat{\mu}_{\text{female}}}{\sqrt{\frac{\sigma^2_{\text{male}}}{n_{\text{male}}} + \frac{\sigma^2_{\text{female}}}{n_{\text{female}}}}} = \frac{5.11 - 2.57}{\sqrt{\frac{4.80}{54} + \frac{4.46}{39}}} = 5.62
\]

Comparison to a threshold:
\[|S| = 5.62 > 1.96\]

P-value:
p-value = 2.422e-07 < 0.05

Conclusion: We reject the null hypothesis. The difference between the tumor sizes is statistically significant.
R outputs

### Welch Two Sample t-test

data: xmale and xfemale

t = 5.6222, df = 83.76, p-value = 2.422e-07
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
1.638720 3.432505

sample estimates:
mean of x mean of y
5.107407 2.571795

Confidence interval

Estimates from the sample

---

## Confidence intervals and Statistical tests

- **Test statistic and distribution parameter**
- **Confidence interval**
- **Estimates from the sample**

---

### R code examples

```r
## Run t.test
res.test1 = t.test(xmale, xfemale)

## Alternative
res.test2 = t.test(data$tumor_size ~ data$gender)
```
Paired vs unpaired statistical tests

Unpaired tests

Should be used when groups being compared are independent.

We built a standardized quantity (S) with known distribution under the null. This indicator includes a variance term…

\[ \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) + 2 \times \text{Cov}(X, Y) \]
Paired vs unpaired statistical tests

Unpaired tests

Should be used when groups being compared are independent.

We built a standardized quantity (S) with known distribution under the null. This indicator includes a variance term…

$$Var(X - Y) = Var(X) + Var(Y) + 2 \times Cov(X, Y)$$

acija Independence => Corr = Cov = 0
but not the other way around

Estimates of variance are different if samples are not independent
Paired vs unpaired statistical tests

Unpaired tests

Should be used when groups being compared are independent.

We built a standardized quantity (S) with known distribution under the null. This indicator includes a variance term…

\[ \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) + 2 \times \text{Cov}(X, Y) \]

\( \triangledown \) Independence \( \Rightarrow \) Corr = Cov = 0 but not the other way around

Estimates of variance are different if samples are not independent

Paired tests

Should be used when groups being compared are not independent.

Observations = Within-groups variations + between-groups variations

Provides a way to adjust for possible confounding effects (e.g. within-groups)

Think « repeated measurements »

Measure at time \( t = 0 \)

vs.

Measure at time \( t = 100 \)
Paired vs unpaired statistical tests

**Unpaired tests**

Should be used when groups being compared are independent.

We built a standardized quantity (S) with known distribution under the null. This indicator includes a variance term...

\[
Var(X - Y) = Var(X) + Var(Y) + 2 \times Cov(X, Y)
\]

\[\text{Independence} \Rightarrow Corr = Cov = 0\]

but not the other way around

Estimates of variance are different if samples are not independent

**Paired tests**

Should be used when groups being compared are not independent.

Observations = Within-groups variations + between-groups variations

Provides a way to adjust for possible confounding effects (e.g. within-groups)

Think « repeated measurements »

Measure at time \( t = 0 \)

vs.

Measure at time \( t = 100 \)

**How ?**

1) Look at documentation !

2) `t.test(vect1, vect2, paired = TRUE)`
## Paired vs unpaired statistical tests

<table>
<thead>
<tr>
<th>Type de test à mettre en évidence</th>
<th>Qualitative nominale (2 groupes)</th>
<th>Qualitative nominale (plus de 2 groupes)</th>
<th>Qualitative ordinaire</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indépendants</td>
<td>Z de comparaison de proportions *</td>
<td>Chi² (χ²)</td>
<td>Test de Cochran-Armitage *</td>
<td>Test de Mann-Whitney.</td>
</tr>
<tr>
<td>Appariés</td>
<td>Chi² (χ²)</td>
<td></td>
<td></td>
<td>t de Student.</td>
</tr>
<tr>
<td>Appariés</td>
<td>Test de McNemar.</td>
<td>Q de Cochran *</td>
<td>Tests des signes *</td>
<td>Test de Wilcoxon. *</td>
</tr>
<tr>
<td>Appariés</td>
<td>Test exact de Fisher.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitatif</td>
<td>Q de Cochran *</td>
<td>Q de Cochran *</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Régression logistique *</td>
<td>Régression logistique multinomiale *</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(source: biostaTGV)
Statistical test

Your turn

- **LungA data**
The mean weight of the female smokers is greater than 62 (level=5%)?

- **LungA data**
The proportion of smokers is statistically different from 50% (level=5%)?
Statistical test

ANSWERS

• LungA data
  The mean weight of the female smokers is greater than 62 (level=5%)?

We **do not** reject the null hypothesis

• LungA data
  The proportion of smokers is statistically different from 50% (level=5%)?

We reject the null hypothesis with 5% risk
Studying two qualitative variables

Is there a significant difference between drug A and B?
Contingency table

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Drug B</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>
Hypothesis

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Drug B</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Define the hypothesis

- **H0**: The distribution of sick and healed patients is the same for drug A and B
- **H1**: The distribution of sick and healed patients is not the same for drug A and B

What happens under the null hypothesis?
Define the hypothesis

- **H0**: The distribution of sick and healed patients is the same for drug A and B
- **H1**: The distribution of sick and healed patients is not the same for drug A and B

What happens under the null hypothesis?

The distribution is the same for each drug

→ we can compute **expected values** (E) by considering that the repartition is the same and compare them to the **observed ones** (O)
Expected value

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Drug B</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>26</td>
</tr>
</tbody>
</table>

What happens under the null hypothesis?

The distribution is the same for each drug

→ we can compute expected values (E) by considering that the repartition is the same and compare them to the observed ones (O)

\[
E_{\text{sick,A}} = \frac{(18 \times 21)}{44} = 8.6
\]

If the distribution are the same, the expected number of sick people for drug A is 8.6
Test statistic

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Cured</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>7 (8.6)</td>
<td>14 (12.4)</td>
<td>21</td>
</tr>
<tr>
<td>Drug B</td>
<td>11 (9.4)</td>
<td>12 (13.6)</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>26</td>
<td>44</td>
</tr>
</tbody>
</table>

Test statistic

\[ T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 (I-1) \times (J-1) \]

with I and J and the number of modalities for each variables
## Test statistic

<table>
<thead>
<tr>
<th></th>
<th>Sick</th>
<th>Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>7 (8.6)</td>
<td>14 (12.4)</td>
</tr>
<tr>
<td>Drug B</td>
<td>11 (9.4)</td>
<td>12 (13.6)</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>26</td>
</tr>
</tbody>
</table>

### Test statistic

\[
T = \frac{(7 - 8.6)^2}{8.6} + \frac{(14 - 12.4)^2}{12.4} + \frac{(11 - 9.4)^2}{9.4} + \frac{(12 - 13.6)^2}{13.6} = 0.965
\]

Compare this value to a theoretical threshold from the Khi-2 distribution.
Compare to a threshold

\[ T = \frac{(7 - 8.6)^2}{8.6} + \frac{(14 - 12.4)^2}{12.4} + \frac{(11 - 9.4)^2}{9.4} + \frac{(12 - 13.6)^2}{13.6} = 0.965 \]

Conclusion

0.965 < 3.84 → we don’t reject the null hypothesis

Chi-2 test

Note: 3.84 = 1.96^2
Chi-2 test - assumption

One assumption must be verified to use a Chi-2 test:

- All expected values ≥ 5

If the assumption is not verified...

The Khi-2 distribution can not be use anymore, use a sampling approach (can be time consuming if some observed values are huge)

Yates continuity

It aims at correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared).

→ Not suitable for small samples
Chi-2 test on R

```r
tab = matrix(c(7,14,11,12),2,2,byrow=T)

## Usual test (with default options)
chisq.test(tab)

## Remove Yates Correction
chisq.test(tab, correct = FALSE)

## If the assumption is not verified
chisq.test(tab, simulate.p.value = FALSE)
```
Chi-2 test

Your turn

Using LungA dataset…

● Are smoker and sex are two independent variables?

● Is there a significant difference between the proportion of male and female in the sample?

● Is the sample representative of the population in terms of smokers/non smokers distribution? (The proportion of smokers in the population is 32%)
Using LungA dataset…

```r
## Are smoker and sex are two independent variables ?
chisq.test(table(data$gender, data$smoker))

## Is parity between male and female respected ?
chisq.test(table(data$gender))

## Is the sample representative of the population in terms of smokers/non smokers distribution ?
(The proportion of smokers in the population is 32%)
chisq.test(table(data$smoker), p = c(0.68, 0.32))
```
Parametric VS Non-parametric tests

- **Parametric tests**
  - Information about population parameter is required

- **Non-parametric tests**
  - No idea regarding the population parameter
## Parametric VS Non-parametric tests

<table>
<thead>
<tr>
<th></th>
<th>Parametric</th>
<th>Non-parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>Information about population is completely known</td>
<td>No information about the population</td>
</tr>
<tr>
<td><strong>Assumptions</strong></td>
<td>Several assumptions are made regarding the population</td>
<td>No assumptions are made regarding the population</td>
</tr>
<tr>
<td><strong>H0</strong></td>
<td>Null hypothesis is made on the parameters of the population</td>
<td>Null hypothesis is free from parameters</td>
</tr>
<tr>
<td><strong>Test statistic</strong></td>
<td>Based on the distribution</td>
<td>Arbitrary</td>
</tr>
</tbody>
</table>

Can be very useful when the distribution is very far from a normal distribution (e.g. compare if a microbiota species is more abundant in the samples A versus the samples B with a rank test).
Mann Whitney test

X1 are observations from population A (size $n_1$)
X2 are observations from population B (size $n_2$)

Samples don’t have to be normally distributed.

We will test the following null hypothesis $H_0$ : the distributions of both populations are equal
Mann Whitney test

X1 are observations from population A (size $n_1$)
X2 are observations from population B (size $n_2$)
Samples don’t have to be normally distributed.

We will test the following null hypothesis H0 : the distributions of both populations are equal

1. Sort data and assign ranks to all observations

<table>
<thead>
<tr>
<th>Values</th>
<th>3.27</th>
<th>3.78</th>
<th>3.85</th>
<th>3.93</th>
<th>3.94</th>
<th>4.12</th>
<th>4.58</th>
<th>4.77</th>
<th>4.79</th>
<th>5.21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Group</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Mann Whitney test

X1 are observations from population A (size $n_1$)
X2 are observations from population B (size $n_2$)
Samples don’t have to be normally distributed.

We will test the following null hypothesis $H_0$: the distributions of both populations are equal

1. Sort data and assign ranks to all observations
2. Add up the ranks and calculate $U_1$ and $U_2$ (NB: the sum of all ranks = $N$ x $(N - 1)/2$)

<table>
<thead>
<tr>
<th>Values</th>
<th>3.27</th>
<th>3.78</th>
<th>3.85</th>
<th>3.93</th>
<th>3.94</th>
<th>4.12</th>
<th>4.58</th>
<th>4.77</th>
<th>4.79</th>
<th>5.21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Group</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

$R_1 = 19$
$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 9$

$R_2 = 36$
$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} = 26$
Mann Whitney test

X1 are observations from population A (size $n_1$)
X2 are observations from population B (size $n_2$)
Samples don’t have to be normally distributed.

We will test the following null hypothesis H0: the distributions of both populations are equal

1. Sort data and assign ranks to all observations
2. Add up the ranks and calculate $U_1$ and $U_2$ (NB: the sum of all ranks = $N \times (N - 1)/2$)
3. Take the minimum value between $U_1$ and $U_2$
4. Compute the p-value (not developed here, but it is based on the number of possibilities to obtain values lower than the test statistic by rearranging the data at random)

$R_1 = 19$

$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 9$

p-value = 0.095

$X_1$ are observations from population A (size $n_1$)
$X_2$ are observations from population B (size $n_2$)
# Parametric VS Non-parametric tests

<table>
<thead>
<tr>
<th></th>
<th>Parametric</th>
<th>Non-parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>Information about population is completely known</td>
<td>No information about the population</td>
</tr>
<tr>
<td><strong>Assumptions</strong></td>
<td>Several assumptions are made regarding the population</td>
<td>No assumptions are made regarding the population</td>
</tr>
<tr>
<td><strong>H0</strong></td>
<td>Null hypothesis is made on the parameters of the population</td>
<td>Null hypothesis is free from parameters</td>
</tr>
<tr>
<td><strong>Test statistic</strong></td>
<td>Based on the distribution</td>
<td>Arbitrary</td>
</tr>
</tbody>
</table>

Are non-parametric tests magic?
Statistical power (of a test)

Data analysis course
Introductory example: mouse weight

Consider two statistical populations (female mice - of a given strain):
- fed with a specific high fat diet
- fed with a standard control diet

Based on previous experiments we know that diet induces differences in weight of adults:

Means: $\mu_{co} = 23.91$ (control) and $\mu_{hf} = 26.37$ (high-fat diet)

Standard deviation: $\sigma_{co} = \sigma_{hf} = 11.42$
Introductory example: mouse weight

Suppose we want to test this difference in weight (in g):
- based on a sample of size $n = 10$ (5 mice per group)
- using a Student test

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>high-fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.89</td>
<td>27.50</td>
<td></td>
</tr>
<tr>
<td>22.72</td>
<td>24.15</td>
<td></td>
</tr>
<tr>
<td>22.01</td>
<td>25.62</td>
<td></td>
</tr>
<tr>
<td>23.92</td>
<td>22.13</td>
<td></td>
</tr>
<tr>
<td>22.77</td>
<td>27.85</td>
<td></td>
</tr>
</tbody>
</table>

Average:
- control: 23.26
- high-fat: 25.45

P-value = 0.12
Introductory example: mouse weight

The p-value is not significant ($P$-value $> 0.05$).

Is there something wrong? May the diet have no actual effect?
Introductory example: mouse weight

The p-value is not significant ($P\text{-value} > 0.05$).

Is there something wrong? May the diet have no actual effect?

We could not reject the null hypothesis but this does not mean there is no effect.

In fact, we did not have enough resolution / power to show such an effect!
Definition:
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = P(\text{Reject } H_0 | H_1 \text{ true}) \]
**Power**

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

$$1 - \beta = \mathbb{P}({\text{Reject } H_0|H_1 \text{ true}})$$

![Magnification x10](image1.png)

- **Do not reject**

![Magnification x100](image2.png)

- **Reject**
Definition:
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = P(\text{Reject } H_0 | H_1 \text{ true}) \]
Power

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true}) \]
**Power**

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true}) \]
Power

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true}) \]
**Power**

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

$$1 - \beta = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true})$$

![Graph showing the comparison between t-test and Mann-Whitney tests for statistical power.](image-url)
**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ true}) \]
Power

**Definition:**
The power of a statistical test is the probability that the test correctly rejects the null hypothesis when the alternative is true.

\[ 1 - \beta = P(\text{Reject } H_0 | H_1 \text{ true}) \]
Statistical power — the basics

Under $H_0$, the test statistic $S_{H_0}$ follows a theoretical distribution (e.g., Student) and we can compute the critical value $C_\alpha$.

$H_0$ is rejected if the test statistic ($S$) is higher than the critical value, i.e., if $S_{\text{obs}} > C_\alpha$, with a risk $\alpha = 0.05$ of being wrong.
Statistical power — the basics

The power of a statistical test is the probability of rejecting the null hypothesis when we should — that is, when the alternative hypothesis is actually true.

\[ \text{power} = 1 - P(\text{type II error}) = 1 - \beta \]

The power is a probability and is very often expressed as a percentage.
Statistical power  — the basics

The power of a statistical test is the probability of rejecting the null hypothesis when we should — that is, when the alternative hypothesis is actually true.

\[ \text{power} = 1 - P(\text{type II error}) = 1 - \beta \]

The power is a probability and is very often expressed as a percentage.

The power depends on the overlap between the distributions under \( H_0 \) and \( H_1 \).
The power increases with sample size

Increasing n decreases the spread of the distribution of sample averages in proportion to $1/\sqrt{n}$

Krzywinski & Naomi Altman - Nat. Meth. 2013
The power increases with effect size $d$

\[
d = \frac{(\mu_1 - \mu_0)}{(\sigma)}
\]

Power increases with $d$, making it easier to detect larger effects

Krzywinski & Naomi Altman - Nat. Meth. 2013
The power increases with effect size $d$

$d = (\mu_1 - \mu_0)/(\sigma)$

Power increases with $d$, making it easier to detect larger effects

*Krzywinski & Naomi Altman - Nat. Meth. 2013*
Relationships among the 4 parameters

Input (3 out of 4)

\[ d = \frac{\mu_1 - \mu_0}{\sigma} \]
\[ 1 - \beta \]
\[ n \]

Output (the 4\(^{th}\))

\[ \alpha \]
\[ d \]
\[ 1 - \beta \]
\[ n \]

Insightful for experiment design!
Given some technical constraints, I want to use 15 Petri dishes to compare bacteria abundance after 24h, under two different treatments.

What is the size of the experimental effect I may be able to capture ($\alpha = 0.05$) with reasonable confidence (power = 0.8)?

NB: you need the R package ‘pwr’

- small: $d = 0.2$
- medium: $d = 0.5$
- large: $d = 0.8$
Given some technical constraints, I want to use 15 Petri dishes to compare bacteria abundance after 24h, under two different treatments.

What is the difference (in average abundance) I may be able to show ($\alpha = 0.05$) with reasonable confidence (power = 0.8)?

NB: in a paper, I read that the variance of log-abundance was 1.2 under similar conditions of growth.
Back to an earlier example:
we know that diet induces differences in weight of adults

Means: $\mu_{co} = 23.91$ (control) and $\mu_{hf} = 26.37$ (high-fat diet)

Standard deviation: $\sigma_{co} = \sigma_{hf} = 11.42$

How many individuals per group shall I use to be able to show a significant difference in 80% of cases (i.e., power = 0.8)?
Multiple testing

Data analysis course
Multiple comparisons

Willis S1, Villalobos VM2, Gevaert Q3, Abramovitz M1, Williams C1, Sikic BI3, Leyland-Jones B1.

Author information

Abstract
PURPOSE: To discover novel prognostic biomarkers in ovarian serous carcinomas.

METHODS: A meta-analysis of all single genes probes in the TCGA and HAS ovarian cohorts was performed to identify possible biomarkers using Cox regression as a continuous variable for overall survival. Genes were ranked by p-value using Stouffer's method and selected for statistical significance with a false discovery rate (FDR) < .05 using the Benjamini-Hochberg method.

Promoting Early, Safe Return to Work in Injured Employees: A Randomized Trial of a Supervisor Training Intervention in a Healthcare Setting.
Spector JT1,2, Reul NK3.

Author information

Abstract
Purpose Supervisors in the healthcare sector have the potential to contribute to disability prevention in injured employees. Published data on the evaluation of return to work (RTW) interventions aimed at direct supervisors are scarce. We sought to determine the effect of a brief audiovisual supervisor training module on supervisor RTW attitudes and knowledge. Methods A parallel-group study, using equal randomization, comparing the training module intervention to usual practice in healthcare supervisors at a quaternary care hospital was conducted. Differences between groups in changes in RTW attitude and knowledge survey question scores between baseline and 3 months were assessed using the Mann-Whitney U test. The Benjamini-Hochberg-Yekutieli procedure was used to control for false discovery rate and generate adjusted p values. Results Forty supervisors were allocated to the intervention group and 41 to the usual practice group. Attitude and knowledge scores for most questions improved between baseline and immediately after intervention administration. Comparing intervention (n = 33) and usual practice groups (n = 37), there was a trend toward
Multiple comparisons

Genome-wide association study identifies novel candidate genes for aggressiveness, deoxynivalenol production and azole sensitivity in natural field populations of Fusarium graminearum.

Talas E1, Kahin R2, Medaner T3, McDonald BA4.

Author Information

Abstract
Genome wide association studies (GWAS) can identify novel genomic regions and genes that affect quantitative traits. Fusarium head blight is a destructive disease caused by Fusarium graminearum that exhibits several quantitative traits, including aggressiveness, mycotoxin production and fungicide resistance. Restriction site associated DNA sequencing (RADseq) was performed for 220 isolates of F. graminearum. 119 isolates were phenotyped for aggressiveness and deoxynivalenol (DON) production under natural field conditions across four environments. EC50 values for propiconazole resistance were calculated for 220 strains in vitro. Approximately 29,000 single nucleotide polymorphism markers were associated to each trait, resulting in 50, 29, and 74 quantitative trait nucleotides (QTNs) that were significantly associated to aggressiveness, DON production, and propiconazole sensitivity, respectively. Approximately 41% of these QTNs caused non-synonymous substitutions in predicted exons while the remainder were synonymous substitutions or located in intergenic regions. Three QTNs associated with propiconazole sensitivity were significant after Bonferroni correction. These QTNs were located in genes not previously associated with azole sensitivity. The majority of the detected QTNs were

Inappropriate suppression of thyrotropin concentrations in young patients with thyroid nodules including thyroid cancer: the Fukushima Health Management survey.

Suzuki S1, Nakamura J2, Suzuki S3, Ohkouchi C4, Mizunuma H5, Midoikawa S6, Fukushima T7, Ito Y8, Shimura H9, Ohira T10, Matsuzaka T11, Ohtsuni A12, Abe M13, Yamashita S14, Suzuki S15,16.

Author Information

Abstract
BACKGROUND: Serum thyroid hormone concentration is regulated through the hypothalamic-pituitary-thyroid axis. We aimed at clarifying the relationships between thyroid hormone regulation and ultrasonographic findings in subjects with thyroid nodules detected during thyroid ultrasound examination for the Fukushima Health Management Survey.

METHODS: As of October 31, 2014, a total of 296,253 subjects, who had been living in Fukushima Prefecture at the time of the Fukushima nuclear power plant accident and were aged 18 years or younger on March 11, 2011, participated in two concurrent screening programs. In the primary screening, thyroid nodules were detected in a total of 2,241 subjects. A secondary confirmatory thyroid ultrasound examination and blood sampling for thyroid function tests were performed on 2,004 subjects. The subjects were re-assessed and classified into disease-free subjects (Group 1), subjects with cysts only (Group 2), subjects with nodules (Group 3), and subjects with malignancy or suspected malignancy (Group 4). Serum concentrations of FT3, FT4, thyrotropin (TSH), and thyroglobulin, as well as the FT3/FT4 ratio were classified according to the diagnoses.

RESULTS: Inverse relationships between age and log TSH values (Spearman's correlation $r = -0.311$, $P = 0.015$), serum FT3 concentration ($r = -0.685$, $P < 0.001$), and the FT3/FT4 ratio ($r = -0.520$, $P < 0.001$) were observed in Group 1. When ANCOVA with Bonferroni post hoc comparisons was used in the four groups, the log TSH values were significantly lower in either Group 3 or Group 4 compared to either Group 1 or Group 2 after
When, why and how?

When:
If you do more than one test simultaneously and you want to control your Type I error ($\alpha$) globally.
When, why and how?

**When:**
If you do more than one test simultaneously and you want to control your Type I error ($\alpha$) globally.

**Why:**
When several tests are done simultaneously, you accumulate the error.

For example:

- 1 test: $\mathbb{P}$(not making error) = $1 - \alpha$
- 2 tests: $\mathbb{P}$(not making error for test 1 & 2) = $(1 - \alpha) \times (1 - \alpha) = (1 - \alpha)^2$
- $k$ tests: $\mathbb{P}$(not making error for $k$ tests) = $(1 - \alpha)^k$
When, why and how?

When:
If you do more than one test simultaneously and you want to control your Type I error (\( \alpha \)) globally.

Why:
When several tests are done simultaneously, you accumulate the error.

For example: with \( \alpha = 5\% \), \( k = 30 \)

- 1 test: \( \mathbb{P}(\text{not making error}) = 0.95 \)
- 2 tests: \( \mathbb{P}(\text{not making error for test 1 & 2}) = 0.9 \)
- \( k \) tests: \( \mathbb{P}(\text{not making error for } k \text{ tests}) = 0.21 \)

This principle is well illustrated by xkcd
When, why and how?

How:
Re-calcualting the p-values to get adjusted p-values.

Incidence of false negatives

- Multiple tests without correction
- FDR
  - Benjamini
  - Hochberg
- FWER
  - Bonferroni

Incidence of false positives
R commands

Adjust P-values for Multiple Comparisons

Description
Given a set of p-values, returns p-values adjusted using one of several methods.

Usage

p.adjust(p, method = p.adjust.methods, n = length(p))

method
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#    "fdr", "none")

Arguments

p numeric vector of p-values (possibly with NAs). Any other R is coerced by as.numeric.

method correction method. Can be abbreviated.

n number of comparisons, must be at least length(p); only set this (to non-default) when you know what you are doing!

Example: differential gene expression
Link with confidence intervals

Comparison of 2 parameters

Overlap

\[ \hat{\theta}_1 \]

\[ \hat{\theta}_2 \]

We don’t reject the null hypothesis

No overlap

\[ \hat{\theta}_2 \]

\[ \hat{\theta}_1 \]

We reject the null hypothesis

\[ H_0 : \theta_1 = \theta_2 \]
Link with confidence intervals

One parameter vs theoretical value (m)

\[ m \in IC(\theta_1) \]

We don’t reject the null hypothesis

\[ m \notin IC(\theta_1) \]

We reject the null hypothesis

\[ H_0 : \theta_1 = m \]
Conclusion

- In this course, the following concepts were discussed:
  - ✓ Confidence intervals
  - ✓ Bootstrapping
  - ✓ Bayesian inference
  - ✓ Statistical tests
  - ✓ Multiple comparisons

Thank you for your attention
Hub team
Confidence intervals and Statistical tests
21-22/01/2019
Which treatment is the best?

Large stones

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>73%</td>
<td>69%</td>
</tr>
</tbody>
</table>

Small stones

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>93%</td>
<td>87%</td>
</tr>
</tbody>
</table>
Simpson’s paradox

Large stones

A

73%
192/263

B

69%
55/80

Small stones

A

93%
81/87

B

87%
234/270

78%
273/350

83%
289/350

D’OH!