Reminder on hypothesis testing

Statistical test and p-value

H. Barbot

BIS2.0 Back2Basic, Mai 2025

Guiding thread

Differential gene expression

Goal: Identify genes which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

You are Mathéo Lode, and you have acces to preclinical data. Tumoral cells are given to mice to see the efficiency of a treatment. Each condition (control and treatment) have 3 observations available of RNAseq data.

Null and alternative hypothesis

We have two hypothesis, which one is the most likely ?

$$\begin{cases} H_0 : \text{the effect can't be observed at population scale} \\ H_1 : \text{the effect is observed at population scale} \end{cases}$$

We test the expression of one gene:

$\begin{cases} H_0 : \text{We don't see a greater difference in expression than expected for this gene} \\ H_1 : \text{We find out a significant difference in expression for this gene} \end{cases}$

Null and alternative hypothesis

We have two hypothesis, which one is the most likely ?

$$\begin{cases} H_0 : \text{the effect can't be observed at population scale} \\ H_1 : \text{the effect is observed at population scale} \end{cases}$$

We test the expression of one gene:

$$\begin{cases} H_0 : \text{We don't see a greater difference in expression than expected for this gene} \\ H_1 : \text{We find out a significant difference in expression for this gene} \end{cases}$$

The **null hypothesis** H_0 is the presumption of innocence. We reject H_0 only if it is obvious that observations are incoherent with it.

The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation [Fisher et al., 1966],
Absence of evidence is not evidence of absence [Altman and Bland, 1995].

Three keys of a statistical test
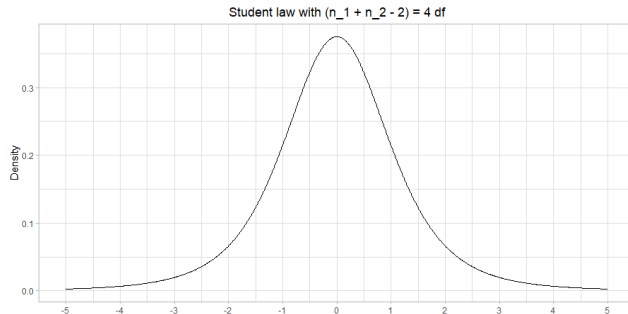
Test of H_0: rule to reject or not H_0 from data.

- **Test statisitic** T, it measure the effect: the more evident the effect, the greater the value of T.
- **Distribution** of T under H_0. If the probability that T being superior than 2 is below 5%, then observing $t_{gene1} = 3$ must lead to reject H_0.
- Test's **p-value**: the probability, under H_0, that the test statistic T is greater than the observed value T_{obs}.

If the **p-value** is smaller than a level α (usually $\alpha = 0.05$), then the effect is significative **at the level** α.

Parametric statistical test   Non-parametric statistical test   Multiple statistical tests   A word on decision tree for statistical test

○○○●○○○○○   ○○   ○○○○○○○○○○   ○○

Reject zone on the distribution

We consider that the expression of the gene of interest have the same variance in the two experimental conditions, with respectively $n_1 = n_2 = 3$ observations.

We test the expression of this gene: $\begin{cases} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A \neq \mu_B \end{cases}$, $\quad T = \dfrac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\dfrac{\hat{\sigma}^2}{n_1} + \dfrac{\hat{\sigma}^2}{n_2}}}$

Student law with (n_1 + n_2 - 2) = 4 df

Parametric statistical test   Non-parametric statistical test   Multiple statistical tests   A word on decision tree for statistical test

○○○●○○○○○   ○○   ○○○○○○○○○○   ○○

Reject zone on the distribution

We consider that the expression of the gene of interest have the same variance in the two experimental conditions, with respectively $n_1 = n_2 = 3$ observations.
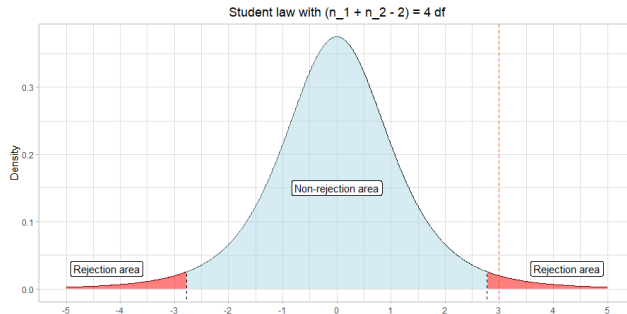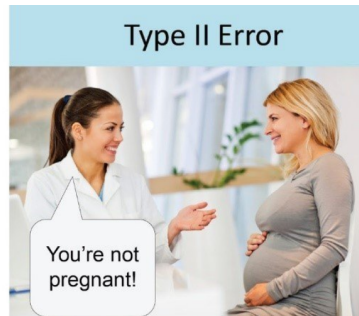
We test the expression of this gene: $\begin{cases} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A \neq \mu_B \end{cases}$, $\quad T = \dfrac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\dfrac{\hat{\sigma}^2}{n_1} + \dfrac{\hat{\sigma}^2}{n_2}}}$


Student law with (n_1 + n_2 - 2) = 4 df

Density

Non-rejection area

Rejection area   Rejection area

Risks in statistical test

		TRUTH	
		There is no effect	There is an effect
		H_0 is true	H_1 is true
DECISION	Fail to reject H_0	Good decision	β (Type II error)
	Reject H_0	α (Type I error)	Good decision (power $1 - \beta$)

Type I Error

You're pregnant!

Type II Error

You're not pregnant!

Parametric statistical test
○○○○○○●○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○○

A word on decision tree for statistical test
○○

Influence of risk α

$\rightarrow$ *a priori* fixed risk: Probability we accept to be wrong when the truth is a absence of effect.
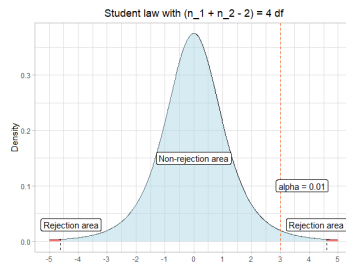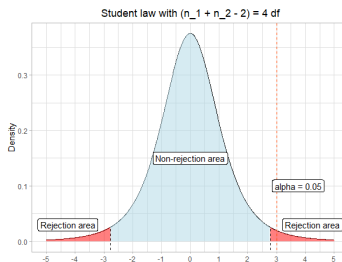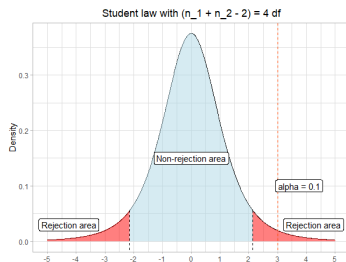
Influence of risk β

$\rightarrow$ **Not a priori fixed risk**: At α fixed, a good test try to minimise β (i.e. maximise the statistical power).

It is easier to conclude a significative difference between two mean if:

- Means are largely different,
- The variability is low in both population
- We have access to a lot of data

Influence of risk β

$\rightarrow$ **Not a priori fixed risk**: At α fixed, a good test try to minimise β (i.e. maximise the statistical power).

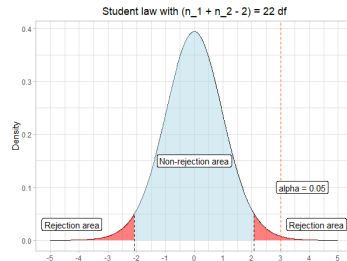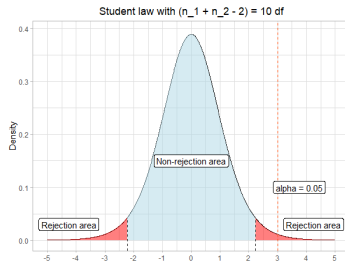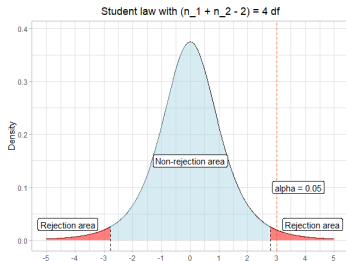It is easier to conclude a significative difference between two mean if:

- Means are largely different,
- The variability is low in both population
- We have access to a lot of data

Definition of p-value

For the distibution of T under H_0, the p-value is the **smallest value** α where the test statistic T is in the reject zone:

$$\text{p-value} = \mathbb{P}(T = t_{gene1}|H_0).$$

Definition of p-value

For the distibution of T under H_0, the p-value is the **smallest value** α where the test statistic T is in the reject zone:

$$\text{p-value} = \mathbb{P}(T = t_{gene1}|H_0).$$

$$T = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$$

Student law with (n_1 + n_2 - 2) = 4 df

Definition of p-value

For the distibution of T under H_0, the p-value is the **smallest value** α where the test statistic T is in the reject zone:

$$\text{p-value} = \mathbb{P}(T = t_{gene1}|H_0).$$

$$T = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$$

$$t_{gene1} = 3$$

Student law with (n_1 + n_2 - 2) = 4 df

Non-rejection area

Rejection area   Rejection area

Definition of p-value

For the distibution of T under H_0, the p-value is the **smallest value** α where the test statistic T is in the reject zone:
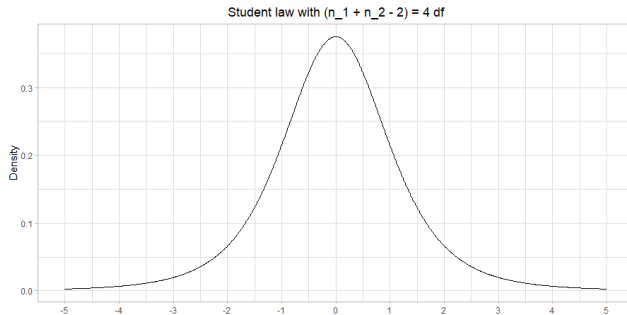
$$\text{p-value} = \mathbb{P}(T = t_{gene1}|H_0).$$

$$T = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$$

$$t_{gene1} = 3$$

Student law with (n_1 + n_2 - 2) = 4 df
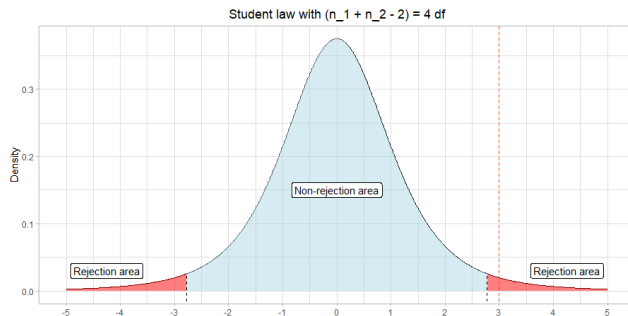
Non-rejection area

Rejection area

Rejection area

Definition of p-value

For the distibution of T under H_0, the p-value is the **smallest value** α where the test statistic T is in the reject zone:

$$\text{p-value} = \mathbb{P}(T = t_{gene1}|H_0).$$

$T = \dfrac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\dfrac{\hat{\sigma}^2}{n_1} + \dfrac{\hat{\sigma}^2}{n_2}}}$

$t_{gene1} = 3$

Student law with (n_1 + n_2 - 2) = 4 df

Non-rejection area

for T = 3,
p-value = 0.0394

Rejection area

Rejection area

A p-value can be interpretted
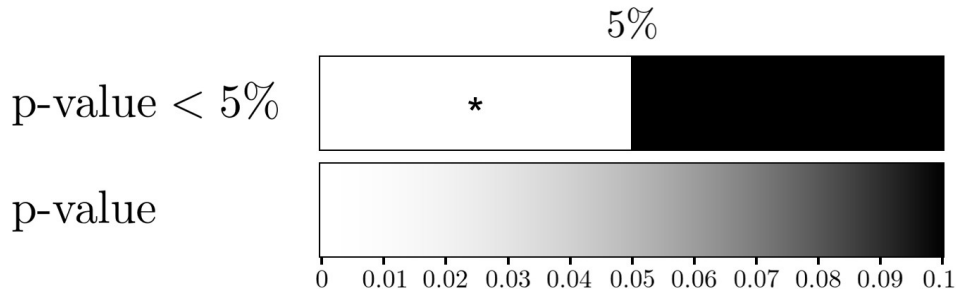
*There is a reason that the speedometer in your car doesn't just read "slow" and "fast".
(F. Harrel, 'warning about the use of cutoffs after logistic regression' in R-help, 2011)*

Parametric statistical test
000000000

Non-parametric statistical test
●○

Multiple statistical tests
0000000000

A word on decision tree for statistical test
○○

Non-parametric statistical tests

→ **Parametric tests** $\implies$ Assumptions on data distribution
(normality, homogeneity of variance, . . .)

→ **Non-parametric tests** $\implies$ $\begin{cases} \text{No assumptions of a specific distribution} \\ \\ \text{More reliable when data samples have a small size} \end{cases}$

Non-parametric doesn't mean no assumptions

	parametric test	**non-parametric test**
use	row data	row data, but a lot use ranks ▸ Supp details
assume	identically distributed and independent data, homogenous variance, ...	less needed, at least independent data
power:	optimal if assumptions are respected, drop quickly if not	stable/robust but less than parametric in the best case

Non-parametric doesn't mean no assumptions

	parametric test	**non-parametric test**
use	row data	row data, but a lot use ranks ‣ Supp details
assume	identically distributed and independent data, homogenous variance, . . .	less needed, at least independent data
power:	optimal if assumptions are respected, drop quickly if not	stable/robust but less than parametric in the best case

Non-parametric test still have assumption:

- Wilcoxon-Mann-Whitney assume independent observations.
- Kruskal-Wallis is a non-parametric version of ANOVA, but assume all groups have an identically shaped and scaled distribution [Kruskal and Wallis, 1952] [Wikipedia,].

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

○○○○○○○○○   ○○   ●○○○○○○○○   ○○

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$n_1 = 3$

$n_2 = 3$

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

○○○○○○○○○   ○○   ●○○○○○○○○○   ○○

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$$n_1 = 3$$

$$n_2 = 3$$

$\left.\begin{array}{c} \\ \\ \end{array}\right\}$ 20061 genes

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

○○○○○○○○○   ○○   ●○○○○○○○○○   ○○

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$$
\left.
\begin{array}{l}
n_1 = 3 \\
\\
n_2 = 3
\end{array}
\right\}
\quad 20061 \text{ genes} \quad \xrightarrow[\substack{\text{Perform the} \\ \text{former test}}]{} \quad 20061 \text{ p-values}
$$

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$$\left.\begin{array}{c} n_1 = 3 \\ \\ n_2 = 3 \end{array}\right\} \text{20061 genes} \xrightarrow[\text{former test}]{\text{Perform the}} \text{20061 p-values}$$

$\implies$ How much false positive discovered gene can we expect ?

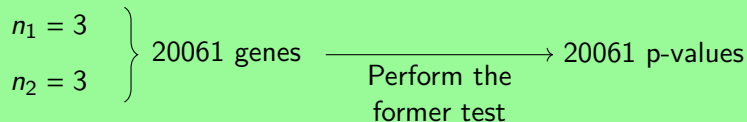# Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$n_1 = 3$
$n_2 = 3$ } 20061 genes $\xrightarrow[\text{former test}]{\text{Perform the}}$ 20061 p-values

$\implies$ How much false positive discovered gene can we expect ?
Since we fixed $\alpha = 0.05$, we expect 5% of the 20061 gene tested to be false positive,
so **1003 genes**

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).
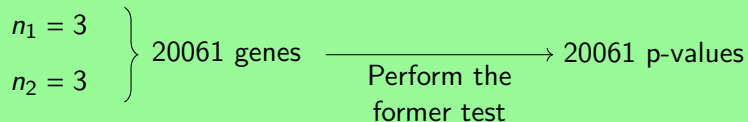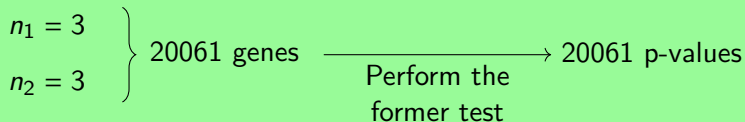
$$
\left.\begin{array}{l} n_1 = 3 \\ \\ n_2 = 3 \end{array}\right\} \;\; 20061 \text{ genes} \;\; \xrightarrow[\substack{\text{Perform the} \\ \text{former test}}]{} \;\; 20061 \text{ p-values} \;\; \implies \;\; \substack{3227 \text{ genes with} \\ \text{p-value} < 5\%}
$$

$\implies$ How much false positive discovered gene can we expect ?
Since we fixed $\alpha = 0.05$, we expect 5% of the 20061 gene tested to be false positive,
so **1003 genes**

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

OOOOOOOOO   OO   ●OOOOOOOOO   OO

Testing all the genes

Differential gene expression

Goal: Identify gene**s** which show a **difference** in expression between two experimental conditions (i.e. greater than expected just due to natural random variation).

$\left.\begin{array}{l} n_1 = 3 \\ \\ n_2 = 3 \end{array}\right\}$ 20061 genes $\xrightarrow[\text{Perform the former test}]{}$ 20061 p-values $\implies$ 3227 genes with p-value < 5%
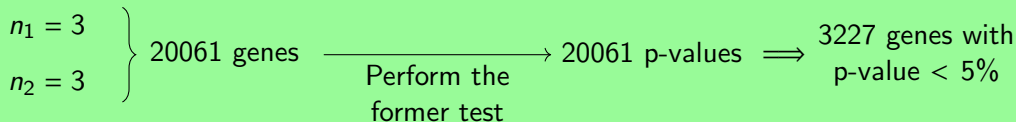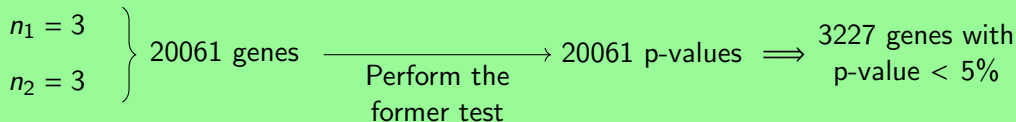
$\implies$ How much false positive discovered gene can we expect ?

Since we fixed $\alpha = 0.05$, we expect 5% of the 20061 gene tested to be false positive, so **1003 genes   or 31.1% of our list of gene differentialy expressed**.

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

000000000   00   0●00000000   00

Multiple hypothesis

A null hypothesis collection, $H_0^{(k)}$ $k = 1, \ldots, m$, with m_0 true null hypothesis.

m boxes out of which m_0 are empty

For the kth box, $H_0^{(k)}$: box is empty

Parametric statistical test
000000000

Non-parametric statistical test
00

Multiple statistical tests
00●0000000

A word on decision tree for statistical test
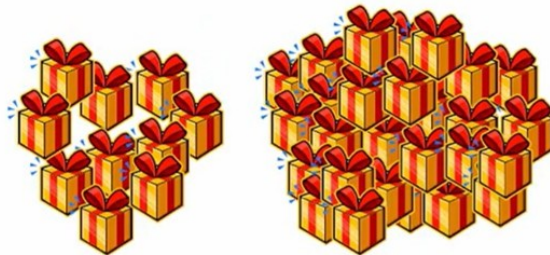00

Multiple p-values

For the kth test, the associated p-value is $p_k = \mathbb{P}_{H_0^{(k)}}\big(\text{reject } H_0^{(k)}\big)$.

I weight each box and evaluate its probability p_k of being empty

Probability of being empty

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

000000000   oo   ooooooooo   oo

Multiple p-values

For the kth test, the associated p-value is $p_k = \mathbb{P}_{H_0^{(k)}}\left(\text{reject } H_0^{(k)}\right)$.

Take these ones   ...   and declare those empty

$p_k \leqslant \alpha$ $p_k > \alpha$

Multiple p-values

For the kth test, the associated p-value is $p_k = \mathbb{P}_{H_0^{(k)}}\big(\text{reject } H_0^{(k)}\big)$.
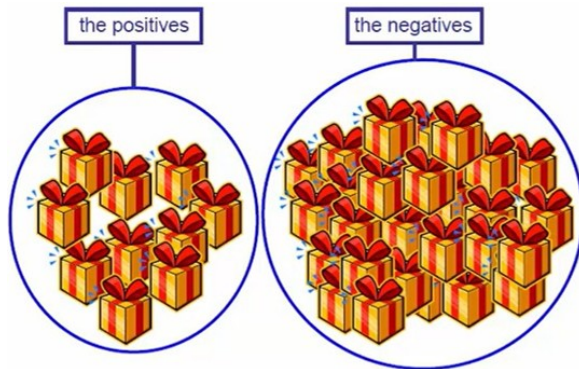

Take these ones   ...   and declare those empty
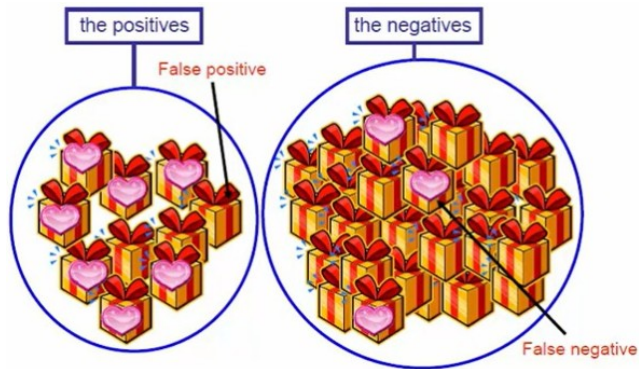
$p_k \leqslant \alpha$ $p_k > \alpha$

Multiple p-values

For the kth test, the associated p-value is $p_k = \mathbb{P}_{H_0^{(k)}}\big(\text{reject } H_0^{(k)}\big)$.

Take these ones   ...   and declare those empty

$p_k \leqslant \alpha$ $p_k > \alpha$

An efficient procedure

What is an **efficient procedure** ?

- Controlling the risk of false positive; as few disappointing discovery as possible.
- Great proportion of true positives within the positives; as much discovery as possible.
- No cofounding effects; knowing what we discover.

An efficient procedure

What is an **efficient procedure** ?

- Controlling the risk of false positive; as few disappointing discovery as possible.

- Great proportion of true positives within the positives; as much discovery as possible.

- No cofounding effects; knowing what we discover.

How to build an **efficient procedure** ?

1. A powerfull Design of Experiment; if the gift is much heavier than the box, it's easier.

2. A good choice for the threshold α.

While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment—no results or incorrect results. (K.M. Kerr, 'Experimental design to make the most of microarray studies', 2003)

Parametric statistical test   Non-parametric statistical test   **Multiple statistical tests**   A word on decision tree for statistical test

○○○○○○○○○   ○○   ○○○○●○○○○○   ○○

1. Design of Experiment (DoE)

For example, we study our 2 conditions (**Control**/**Treatment**) with 8 samples each:

Day 1   ☐☐☐☐☐☐☐☐   8 **Control** samples

Day 2   ■■■■■■■■   8 **Treatment** samples

The global mean of samples for our 2 conditions are significatively differents,

1. Design of Experiment (DoE)

For example, we study our 2 conditions (**Control**/**Treatment**) with 8 samples each:

Day 1 ⬜⬜⬜⬜⬜⬜⬜⬜ 8 **Control** samples

Day 2 ⬛⬛⬛⬛⬛⬛⬛⬛ 8 **Treatment** samples

The global mean of samples for our 2 conditions are significatively differents,
but there are **unidentifiability** between effect of the treatment 🧬 and the day ⚠️.

Parametric statistical test
○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○●○○○○○

A word on decision tree for statistical test
○○

1. Design of Experiment (DoE)

For example, we study our 2 conditions (**Control**/**Treatment**) with 8 samples each:

Day 1   8 **Control** samples

Day 2   8 **Treatment** samples

The global mean of samples for our 2 conditions are significatively differents,
but there are **unidentifiability** between effect of the treatment and the day .

Day 1   8 **Control** samples

Day 2   8 **Treatment** samples

<u>Randomisation</u>

1. Design of Experiment (DoE): Fractional plan

Especially if money is an issue, you can test multiple factors at the same time with a minimum of combination [Husson,].

2. A good choice for the threshold α

We want to have a decision rule on the m p-values p_k, $1, \ldots, m$:

- At which threshold α do we reject $H_0^{(k)}$?

- If p-values are ordered, at which rank $\hat{k}$ p-values $p_{(1)}$ to $p_{(\hat{k})}$ lead to reject the null hypothesis, while p-values $p_{(\hat{k}+1)}$ to $p_{(m)}$ doesn't ?

For a threshold α:

- P_α: numbers of gene (gift box) which reject the null hypothesis (known).

- FP_α: numbers of gene (gift box) which wrongly reject the null hypothesis (unknown).
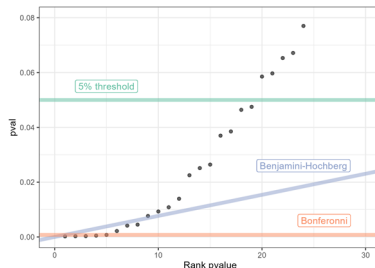
Family Wise error rate (FWER)

To define the new threshold α_{FWER}, we set:

$$\text{FWER}_{\alpha_{FWER}} \leqslant \mathbb{P}\big(\text{FP}_{\alpha_{FWER}} > 0\big) = \alpha, \quad \text{equivalently } \mathbb{P}\big(\text{FP}_{\alpha_{FWER}} = 0\big) \geq 1 - \alpha$$

Bonferroni procedure

If the number of expected false positive discovery is $\text{FP}_\alpha = m \times \alpha$ we only have to take

$$\alpha_{FWER} = \frac{\alpha}{m}.$$

The same result is obtain if we **adjust p-values**: $p_k \leqslant \frac{\alpha}{m} \Leftrightarrow \tilde{p}_k = m p_k \leqslant \alpha$. ▸ Supp details

---

$\Rightarrow$ FWER procedures tends to be very restrictive with few false positive.

In our case, $\alpha_{FWER} = \frac{0.05}{20061} = 2.492 \times 10^{-6}$ which lead to 38 positive genes.

False discovery rate (FDR)

To define the new threshold α_{FDR}, we set:

$$\text{FDR}_{\alpha_{FDR}} = \mathbb{E}\left[\frac{FP_{\alpha_{FDR}}}{P_{\alpha_{FDR}}}\right] \leqslant \alpha$$

Benjamini-Hochberg procedure

The number of false positive is unknown, so we control the number of positive and estimate the esperance of fale positive. For that, we order p-values and for the kth p-value we have

$$\alpha_{FDR} = k\frac{\alpha}{m}.$$

Again we can rather look at **adjust p-values**: $\tilde{p}_k = \frac{mp_k}{k} \leqslant \alpha$. ▸ Supp details

$\Rightarrow$ FDR procedures are less demanding than FWER, and provides much more positives.

In our case, $\alpha_{FDR} = k\frac{0.05}{20061}$ which lead to 1958 positive genes ($\alpha_{FDR} = 0.00488$).

Parametric statistical test
○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○●

A word on decision tree for statistical test
○○

Ranked p-values

Parametric statistical test
○○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○●

A word on decision tree for statistical test
○○

Ranked p-values

Parametric statistical test
○○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○●

A word on decision tree for statistical test
○○

Ranked p-values

Parametric statistical test
○○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○○

A word on decision tree for statistical test
●○

Decision tree have gaps

Parametric statistical test
○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○○

A word on decision tree for statistical test
●○

Decision tree have gaps

Oversimplified or ambiguous assumptions: Only normal ? same variance too ? Independant measure ?

Type of data?

- Numerical → How numerical data are organized
  - Two groups → Data distribution assumptions?
    - Parametric → t-test
    - Non-parametric → Mann-Whitney U Test
  - Paired data → Data distribution assumptions?
    - Parametric → Paired t-test
    - Non-parametric → Wilcoxon signed-rank test
  - More than two groups → Data distribution assumptions?
    - Parametric → ANOVA
    - Non-parametric → Kruskal-Wallis test
  - Relationship between multiple variables → Data distribution assumptions?
    - Parametric → Pearson correlation Regression analysis
    - Non-parametric → Spearman correlation
- Categorical → Target task?
  - Compare two variables → Chi-Square test of independence
  - Test proportions → Z-test for proportions Binomial test

Parametric statistical test
○○○○○○○○○

Non-parametric statistical test
○○

Multiple statistical tests
○○○○○○○○○○

A word on decision tree for statistical test
●○

Decision tree have gaps

Oversimplified or ambiguous assumptions: Only normal ? same variance too ? Independant measure ?

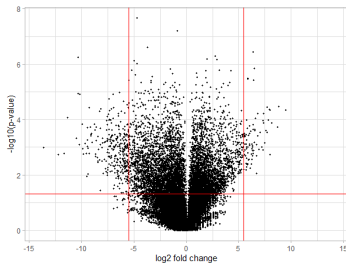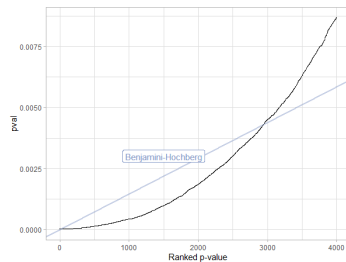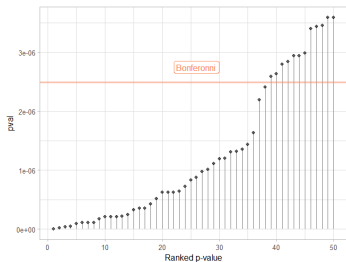Doesn't prevent (▸ p-hacking): Which test for homoscedasticity or normality to keep?

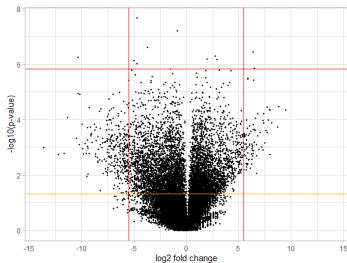Parametric statistical test
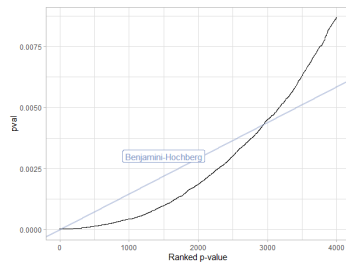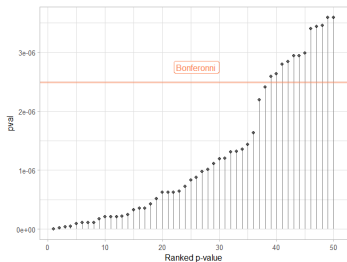ooooooooo

Non-parametric statistical test
oo

Multiple statistical tests
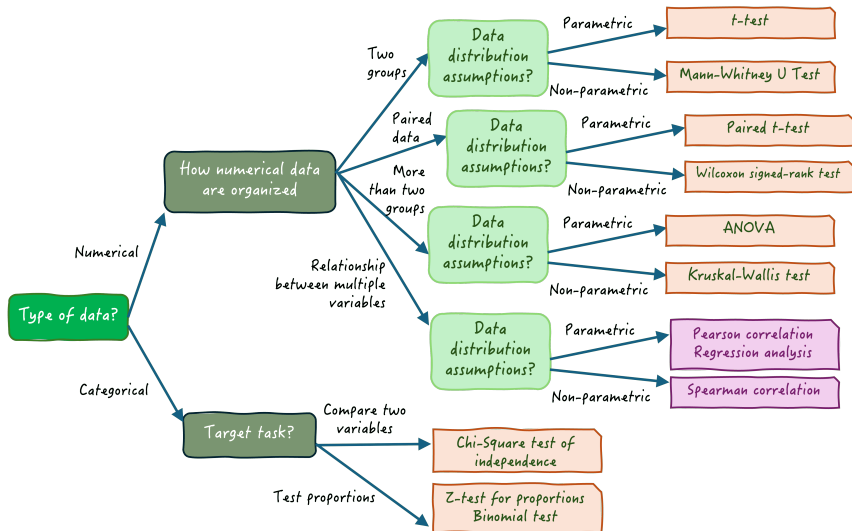oooooooooo

A word on decision tree for statistical test
●o

Decision tree have gaps

Oversimplified or ambiguous assumptions: Only normal ? same variance too ? Independant measure ?

Doesn't prevent ▶ p-hacking : Which test for homoscedasticity or normality to keep?

Some knowledge require to use ▶ paired data or modern method (mixture models, permutations, . . .)

Type of data?

Numerical

Categorical

How numerical data are organized

Two groups

Paired data

More than two groups

Relationship between multiple variables

Data distribution assumptions?

Data distribution assumptions?

Data distribution assumptions?

Data distribution assumptions?

Parametric — t-test

Non-parametric — Mann-Whitney U Test

Parametric — Paired t-test

Non-parametric — Wilcoxon signed-rank test

Parametric — ANOVA

Non-parametric — Kruskal-Wallis test

Parametric — Pearson correlation Regression analysis

Non-parametric — Spearman correlation

Target task?

Compare two variables — Chi-Square test of independence

Test proportions — Z-test for proportions Binomial test

Parametric statistical test
○○○○○○○○○

Non-parametric statistical test
○○

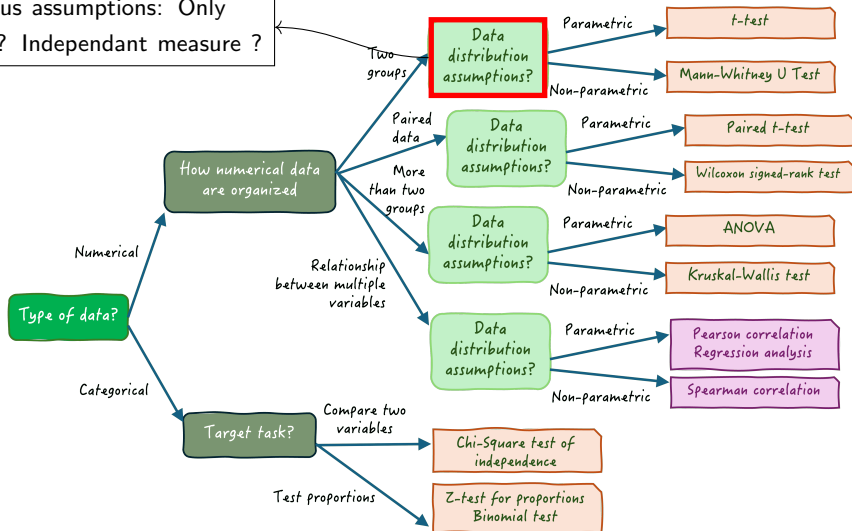Multiple statistical tests
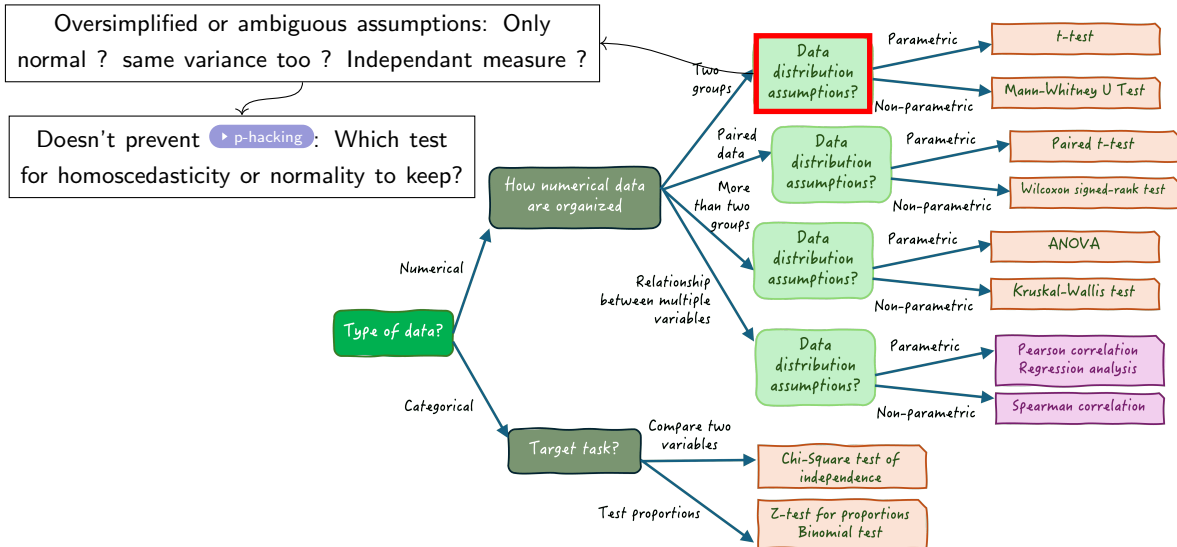○○○○○○○○○○○

A word on decision tree for statistical test
●○

Decision tree have gaps

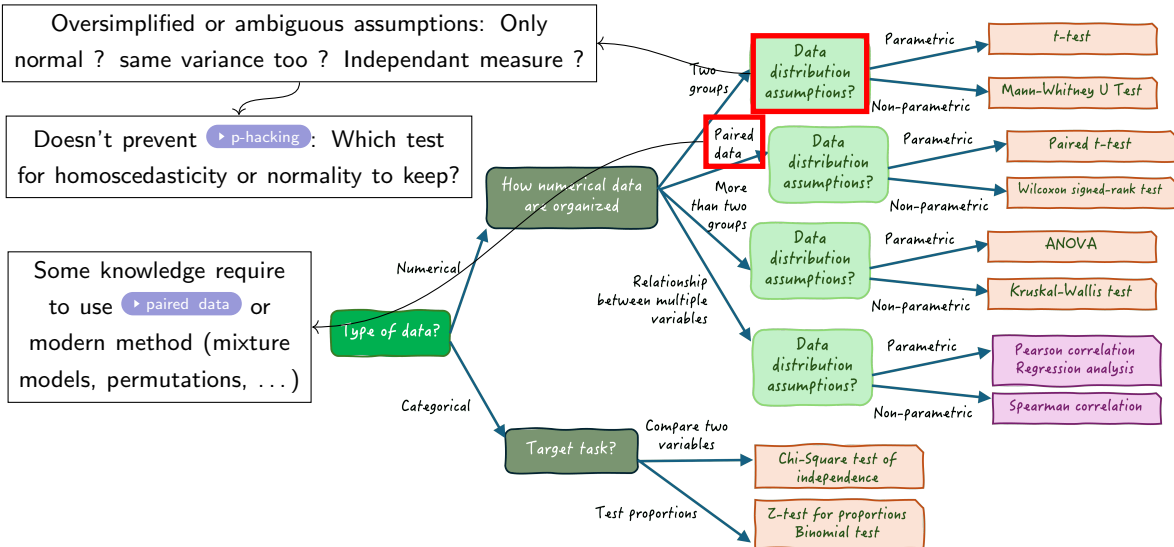Oversimplified or ambiguous assumptions: Only normal ? same variance too ? Independant measure ?
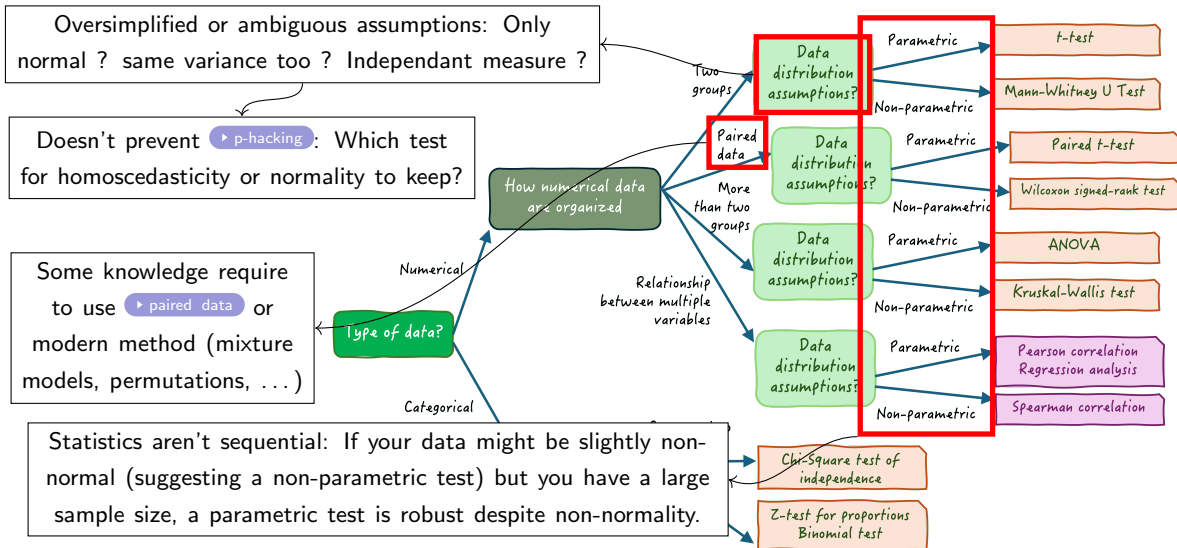
Doesn't prevent ▸ p-hacking : Which test for homoscedasticity or normality to keep?

Some knowledge require to use ▸ paired data or modern method (mixture models, permutations, . . .)

Statistics aren't sequential: If your data might be slightly non-normal (suggesting a non-parametric test) but you have a large sample size, a parametric test is robust despite non-normality.

How numerical data are organized

Two groups

Paired data

More than two groups

Relationship between multiple variables

Type of data?

Numerical

Categorical

Data distribution assumptions?

Data distribution assumptions?

Data distribution assumptions?

Data distribution assumptions?

Parametric

Non-parametric

Parametric

Non-parametric

Parametric

Non-parametric

Parametric

Non-parametric

t-test

Mann-Whitney U Test

Paired t-test

Wilcoxon signed-rank test

ANOVA

Kruskal-Wallis test

Pearson correlation Regression analysis

Spearman correlation

Chi-Square test of independence

Z-test for proportions Binomial test

Decision trees are good reminders (like this presentation)

You knew 90% of all tests used, but now you better understand what you do with them: how they work and what they assume. With these new insights, you precise your limits of understanding and when you reach it.

Take home message

→ Use decision trees as a starting point rather than a definitive guide [MacFarland et al., 2016], consult statistical literature for complex cases.

→ With great power comes great responsability [Stan Lee,], you make arbitrary choices (and it's ok). Always understand why each decision point matters.

References I

Altman, D. G. and Bland, J. M. (1995).
Statistics notes: Absence of evidence is not evidence of absence.
Bmj, 311(7003):485.

Causeur, D.
Analyse de données génomiques.
[Online (https://www.youtube.com/watch?v=ASFvJIW1064); accessed 18-February-2025].

Fisher, R. A., Fisher, R. A., Genetiker, S., Fisher, R. A., Genetician, S., Britain, G., Fisher, R. A., and Généticien, S. (1966).
The design of experiments, volume 21.
Springer.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016).
Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations.
European journal of epidemiology, 31(4):337–350.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015).
The extent and consequences of p-hacking in science.
PLoS biology, 13(3):e1002106.

Husson, F.
Cours de planification expérimentale: Les plans fractionnaires.
[Online (https://hugobarbot.github.io/teaching.html); accessed 29-April-2025].

Husson, F. and Mahmoud, R.
Démarche statistique.
[Online (https://demarche-stat-lesson.netlify.app/); accessed 18-February-2025].

References II

📄 Ioannidis, J. P. (2005).
Why most published research findings are false.
PLoS medicine, 2(8):e124.

📄 Kruskal, W. H. and Wallis, W. A. (1952).
Use of ranks in one-criterion variance analysis.
Journal of the American statistical Association, 47(260):583–621.

📄 MacFarland, T. W., Yates, J. M., et al. (2016).
Introduction to nonparametric statistics for the biological sciences using R.
Springer.

📄 Razali, N. M., Wah, Y. B., et al. (2011).
Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests.
Journal of statistical modeling and analytics, 2(1):21–33.

📄 Stan Lee, S. D.
Amazing fantasy 15.
[Online; accessed 18-February-2025].

📄 Sterne, J. A. and Smith, G. D. (2001).
Sifting the evidence—what's wrong with significance tests?
Physical therapy, 81(8):1464–1469.

📄 Thompson, B. (1987).
The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice.

References III

Wikipedia.
Kruskal–wallis test — Wikipedia, the free encyclopedia.
[Online (https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_test); accessed 18-February-2025].

Common non-parametric tests ◂ non-parametric tests assumptions

Test	Data Type Used	Note
Rank-based tests		
Mann–Whitney U (Wilcoxon rank-sum)	Ranks	Compares ranks between two independent groups.
Wilcoxon signed-rank test	Ranks of differences	Based on the ranks of paired differences.
Kruskal–Wallis	Ranks	Extension of Mann–Whitney to k independent groups.
Friedman	Ranks within blocks	For k repeated measures or blocks, ranks computed within each block.
Spearman's ρ	Ranks	Monotonic rank correlation.
Kendall's τ	Pair orderings	Measures concordance/discordance.
Tests using raw data (or categories)		
Sign test	Signs ($\pm$)	Uses only the sign of differences, not their magnitude or rank.
McNemar's test	Binary categories	Counts discordances in a 2×2 before/after contingency table.
Chi-squared test of independence	Counts	Works on cell frequencies in a contingency table.
Kolmogorov–Smirnov test	Raw values	Compares empirical distribution functions.
Lilliefors test	Raw values	Adaptation of KS for normality testing without fixed parameters.
Log-rank test	Survival times	Compares survival curves based on exact event times.

Bonferroni procedure ◂ FWER

$$
\begin{aligned}
\text{FWER} \; &= \; \mathbb{P}\big([p_1 \leqslant \alpha_{FWER}] \text{ or } \ldots \text{ or } [p_m \leqslant \alpha_{FWER}]\big) \\
&\leqslant \; \mathbb{P}\big([p_1 \leqslant \alpha_{FWER}]\big) \; + \; \ldots \; + \; \mathbb{P}\big([p_m \leqslant \alpha_{FWER}]\big) \\
&\leqslant \; m_0 \times \alpha_{FWER}
\end{aligned}
$$

If $\alpha_{FWER} = \frac{\alpha}{m}$, then $\text{FWER} = \frac{m_0}{m}\alpha \leqslant \alpha$.

Benjamini-Hochberg procedure ◂FDR

$$
\begin{aligned}
\text{FDR} &= \mathbb{E}\left[\frac{FP_{\alpha_{FDR}}}{P_{\alpha_{FDR}}}\right] = \frac{\mathbb{E}\left[FP_{\alpha_{FDR}}\right]}{P_{\alpha_{FDR}}} \\
&= \frac{m_0 \ \alpha_{FDR}}{P_{\alpha_{FDR}}} = \frac{m_0}{m}\frac{m \ \alpha_{FDR}}{P_{\alpha_{FDR}}} \le \frac{m \ \alpha_{FDR}}{P_{\alpha_{FDR}}}.
\end{aligned}
$$

If, for the kth p-value, $\alpha_{FDR} = k\frac{\alpha}{m}$, then FDR $\le \alpha$.

p-hacking ◂ Statistical tree

Decision tree doesn't prevent p-hacking [Head et al., 2015]: Which normal test will you keep, qq-plot, χ^2 adequation, Kolmogorov-Smirnov, Lilliefors, Anderson-Darling, Shapiro-Wilks,... Same question for homoscedasticity.
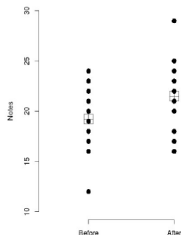
A comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling test concluded that Shapiro-Wilk have the best power for the same significance level, followed by Anderson-Darling [Razali et al., 2011].

Example of paired data ◂ Statistical tree

Before	After
18	22
21	25
16	17
22	24
19	18
24	29
17	20
20	23
23	21
12	16

	Before	After
• Louise	18	22
• Léo	21	25
• Emma	16	17
• Gabriel	22	24
• Chloé	19	18
• Adam	24	29
• Lola	17	20
• Timéo	20	23
• Inès	23	21
• Raphaël	12	16


```
R> wilcox.test(x, y, paired = FALSE)
Wilcoxon rank sum test with continuity
correction

W = 35, p-value = 0.2716
alternative hypothesis: true location shift
is not equal to 0


R> t.test(x, y, paired = FALSE)
Two Sample t-test

t = -1.3529, df = 18, p-value = 0.1928
alternative hypothesis: true difference in
means is not equal to 0
```
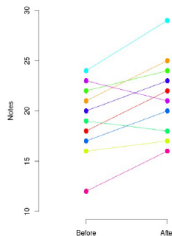
```
R> wilcox.test(x, y, paired = TRUE)
Wilcoxon rank sum test with continuity
correction

V = 5, p-value = 0.02428
alternative hypothesis: true location shift
is not equal to 0


R> t.test(x, y, paired = TRUE)
Two Sample t-test

t = -3.1461, df = 9, p-value = 0.01181
alternative hypothesis: true difference in
means is not equal to 0
```