STATISTICAL MODELS FOR CLINICAL AND HEALTH DATA

V. Monbet

[1] Université de Rennes/UFR Mathématiques

Outline

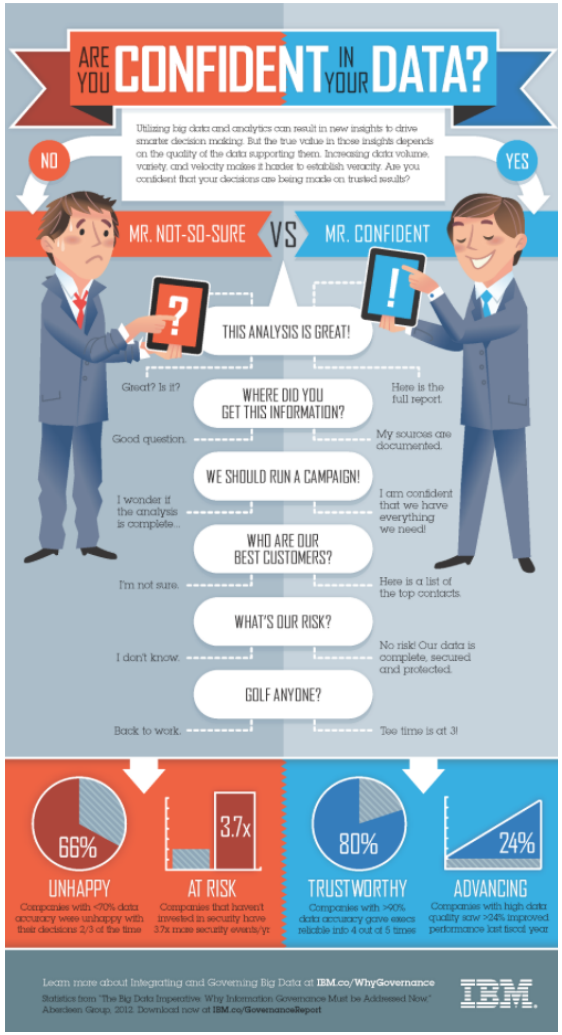
Statistical hypothesis tests

Guilty or innocent ?
True or false ?

How can you be sure you are making a (statistically) valid decision when interpreting data ?

Content

1. Ingredients of of a test : test hypothesis, test statistic, p-value, errors
2. Permutation test
3. Student's test, Fisher's test



Observed data

Samples/populations

A **population** comprises all individuals or objects of interest.
Let's $X(\omega), \omega \in \Omega$ be a random variable, Ω represents the population.

Data is collected from a **sample**,
which is a subset of the population.
$x_1 = X(\omega_1), \cdots, x_n = X(\omega_n)$ are n independant realisation of X, it represents the sample.

To estimate the risk of side effects in the treatment of prostate cancer with radiation therapy, researchers conducted a study on a cohort of more than 300 patients. For 5 years after treatment, they monitored the occurrence of side effects. They found that 35suffered from side effects.

What is the sample size in this study ?
What is a reasonable population to which we could generalize the study's conclusions ?

Statistical inference
Statistical inference is the process of using data from a sample
to obtain information about the population.

Statistical tests, generalities

Statistical tests

A statistical test (or hypothesis test) uses data from a sample to decide between two hypotheses concerning a population.

Examples

- Imagine that we have a cohort and we study the impact of smoking on pregnancy. The question could be : the proportion of pregnancies is the same in the group of smokers as in the group of non-smokers, whereas there are fewer pregnancies among smokers.
- Trouver 2 autres exemples , l'un pour la moyenne et l'autre...
-

by controlling the risk of making a wrong decision.

Effect of caffeine, double-blind experiment

Many people believe they need a cup of coffee or another source of caffeine to start their day. The effects of caffeine on the body have been extensively studied. In one experiment, researchers trained a sample group of students to tap their fingers at a rapid pace. The sample was then randomly divided into two groups of 10 students each. Each student drank the equivalent of two cups of coffee, which contained approximately 200 mg of caffeine for the students in one group but was decaffeinated coffee for the second group. After a period of two hours, each student was tested to measure their finger tapping rate (number of taps per minute). The students did not know whether their drinks contained caffeine, nor did the person measuring the tapping rates.

Caféine : 245, 246, 246, 248, 248, 248, 250, 250, 250

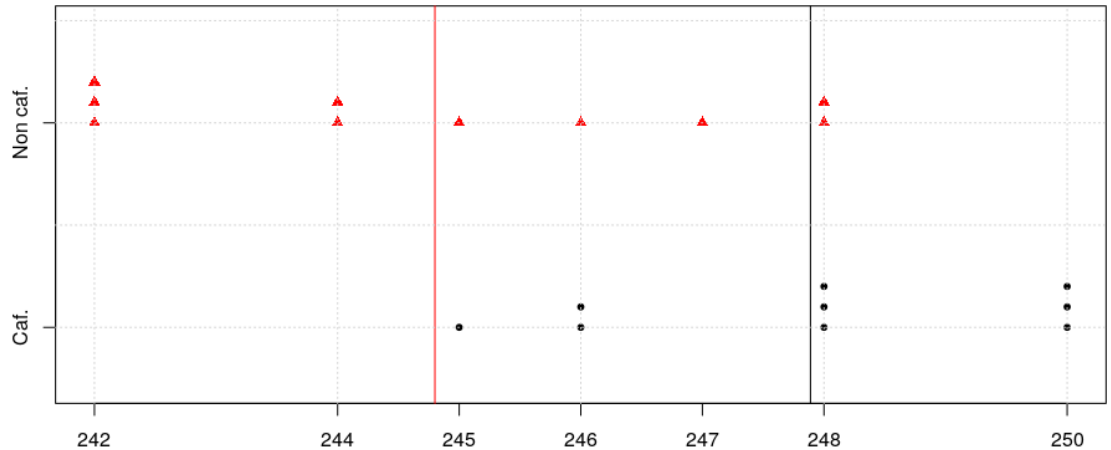
Sans caféine : 242, 242, 242, 244, 244, 245, 246, 247, 248, 248

© Hakan Dere/iStockphoto

Photodisc/Getty Images,Inc.

Propose a (statistical) method to determine whether caffeine significantly increases heart rate.

- Observations
  Caféine : 245, 246, 246, 248, 248, 248, 250, 250, 250
  Sans caféine : 242, 242, 242, 244, 244, 245, 246, 247, 248, 248
- Observations per group



Vertical lines materialize the group means.

Is the gap between means significant ?

Statistical tests (for the mean), null hypothesis

In practice, we begin by studying what happens under the assumption that there is no difference.

We say that the effect of the qualitative variable (caffeine/non-caffeine) is zero and we note

$$H_0 : \mu_C = \mu_{NC}$$

where H_0 stands for *"Null hypothesis"*

μ_C and μ_{NC} are the (theoretical) mean numbers of beats in the caffeine group (C) and the non-caffeine group (NC), respectively.

μ_C and μ_{NC} can not be observed , they are approximated by $\hat{\mu}_C$ and $\hat{\mu}_{NC}$ where

$$\hat{\mu}_C = \frac{1}{n} \sum_{i \in C} X_i^{(C)}$$

$$\hat{\mu}_{NC} = \frac{1}{n} \sum_{i \in NC} X_i^{NC}$$

Alternative hypothesis

In the alternative hypothesis, we indicate the effect of the qualitative variable (with caffeine/without caffeine).

$$H_1 : \mu_C > \mu_{NC}$$

where H_1 stands for *"alternative hypothesis"*

The **problem** consists in decided between hypothesis

$$H_0 : \mu_C = \mu_{NC} \text{ versus } H_1 : \mu_C > \mu_{NC}$$

from the information given by the sample i.e. $\hat{\mu}_C$ and $\hat{\mu}_{NC}$.

Other possible alternatives are $\mu_C < \mu_{NC}$, $\mu_C \neq \mu_{NC}$.

Statistical test, hypothesis

Null and alternative hypothesis

Null hypothesis (H_0) : there is no effect (or no difference).

Alternative hypothesis (H_1) : Claim for which we are seeking significant evidence.

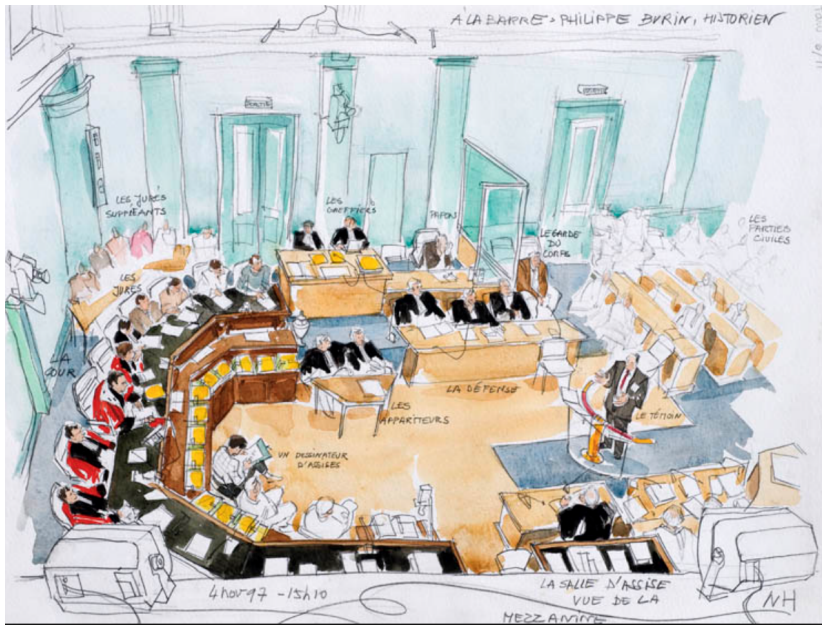
The alternative hypothesis is established by observing evidence (data) that contradicts the null hypothesis and supports the alternative hypothesis.

Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement into statistical testing terms.

“A person is innocent until proven guilty.”

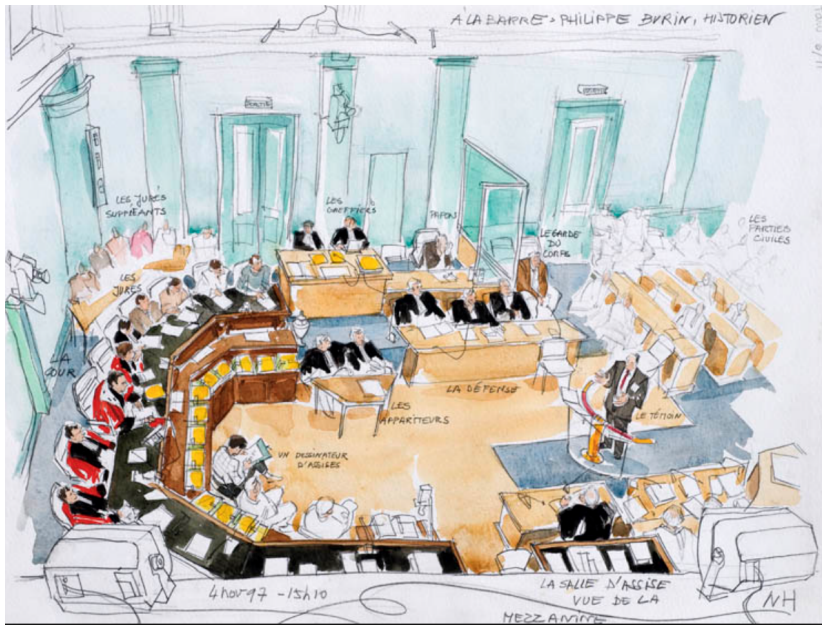


Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement into statistical testing terms.

"A person is innocent until proven guilty."



"Innocent" is the null hypothesis, H_0 i.e. the status quo that we assume to be the situation until we see convincing evidence to the contrary.

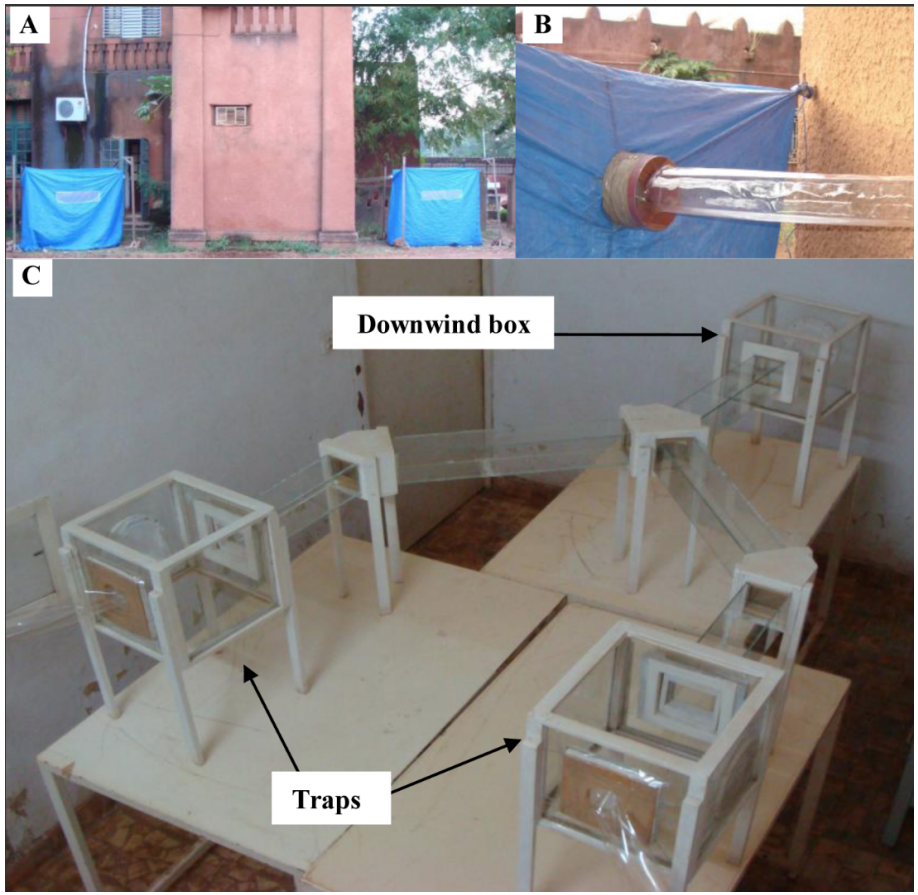
"Guilty" represents i.e. the alternative hypothesis, H_1 the claim that prompts the trial.

Does drinking beer attract mosquitoes ?

In a study [1] conducted in Burkina Faso, Africa, on the spread of malaria, researchers looked at the link between beer consumption and mosquito attraction.

In the experiment, 25 volunteers consumed one liter of beer while 18 others consumed one liter of water. The volunteers were randomly assigned to the two groups. Each volunteer's attractiveness to mosquitoes was tested twice : before drinking beer or water and after. The mosquitoes were captured in traps when they approached the volunteers.

For the beer-drinking group, the total number of mosquitoes caught in the traps before consumption was 434, and the total was 590 after consumption. For the water-drinking group, the total was 337 before and 345 after.



1. Lefevre, T., et al., "Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes," PLoS ONE, 2010 ; 5(3) : e9546.

Exercise : beer and mosquitoes

Do we attract mosquitoes when we drink beer ?
In a study [2] conducted in Burkina Faso, Africa, on the spread of malaria, researchers looked at the link between beer consumption and mosquito attraction.

In the experiment, 25 volunteers consumed one liter of beer while 18 others consumed one liter of water. The volunteers were randomly assigned to the two groups. Each volunteer's attractiveness to mosquitoes was tested twice : before the beer or water and after. Mosquitoes were captured in traps when they approached the volunteers.

For the beer-drinking group, the total number of mosquitoes caught in the traps before consumption was 434, and the total was 590 after consumption. For the water-drinking group, the total was 337 before and 345 after.

- What are the variables in this experiment ?
- State the test hypotheses.
- Why is it important that the volunteers be assigned to groups randomly ?
- Why is it important to conduct two tests, before and after fluid intake ?

2. Lefevre, T., et al., "Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes," PLoS ONE, 2010 ; 5(3) : e9546.

Outline

Permutation test

Let us consider the null hypothesis,

$$H_0 : \mu_C = \mu_{NC}$$

We want to know whether the observed statistic is close to what we would observe if H_0 were true, or whether it is far from it.

However, under H_0, we know that $\mu_C = \mu_{NC}$.

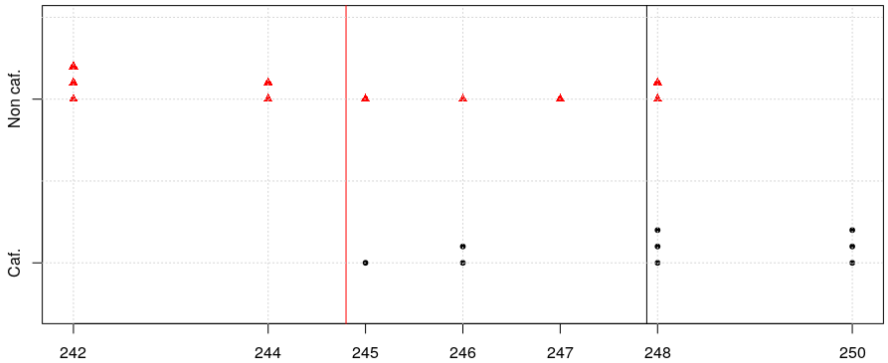
In other words, if H_0 is true, we can
- randomly mix observations by assigning them to a group at random
- and the observed difference in means $\hat{\mu}_C - \hat{\mu}_{NC}$ will be a probable value of the simulated distribution of differences in means.

If H_0 is false, $\hat{\mu}_C - \hat{\mu}_{NC}$ will be an unlikely value of the simulated distribution under H_0.

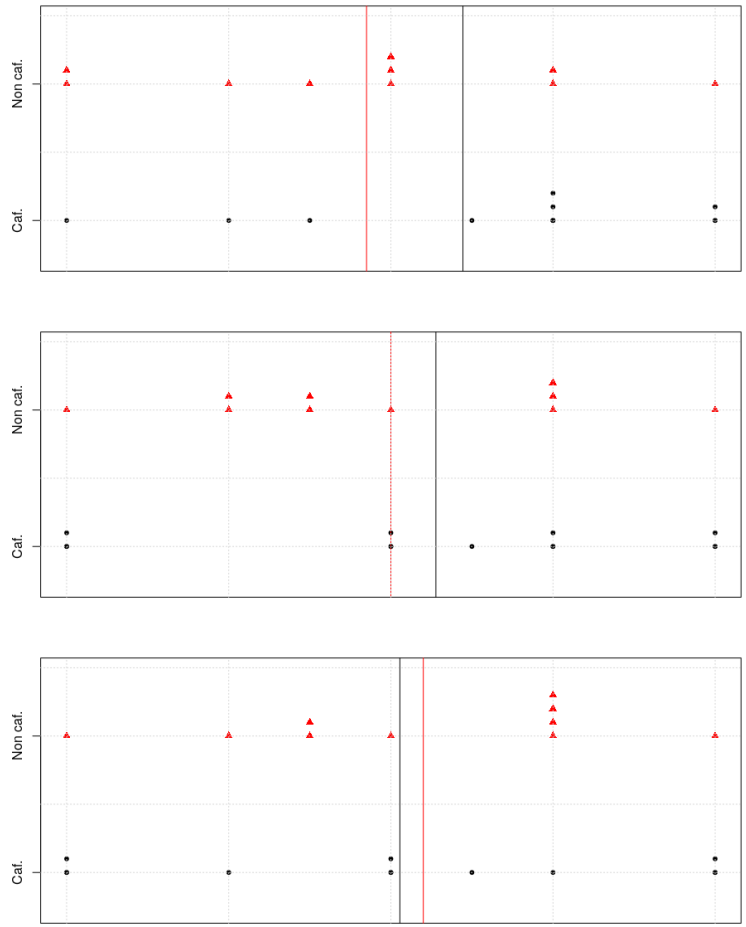
See also :
https://www.jwilber.me/permutationtest/

Permutations

Observations

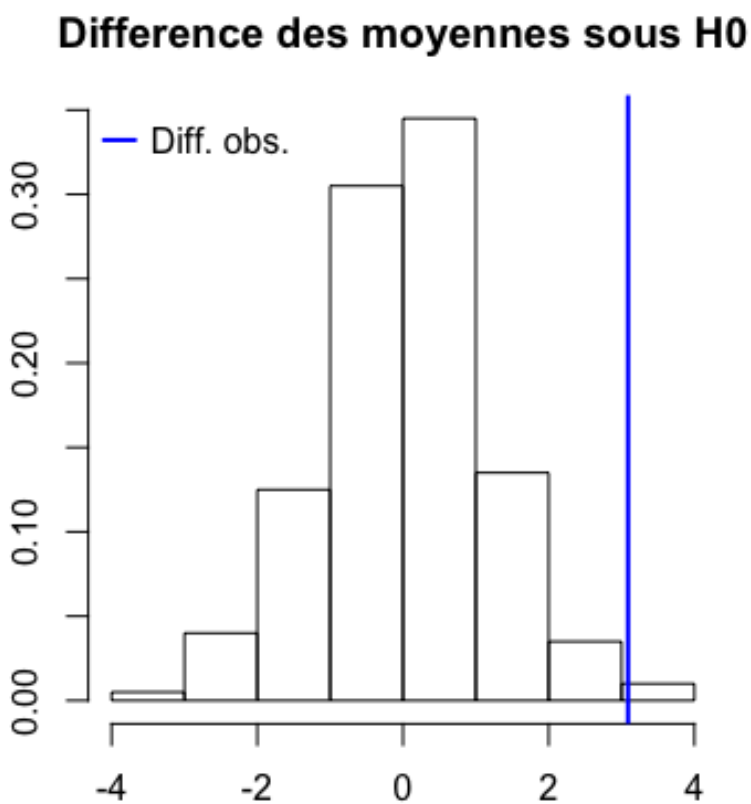


Permutations



Difference between means under H_0

We start again, but this time performing 200 permutations. We then plot the histogram of the differences in proportions simulated by permutation.
The histogram represents the distribution of differences in the average number of beats under the null hypothesis.

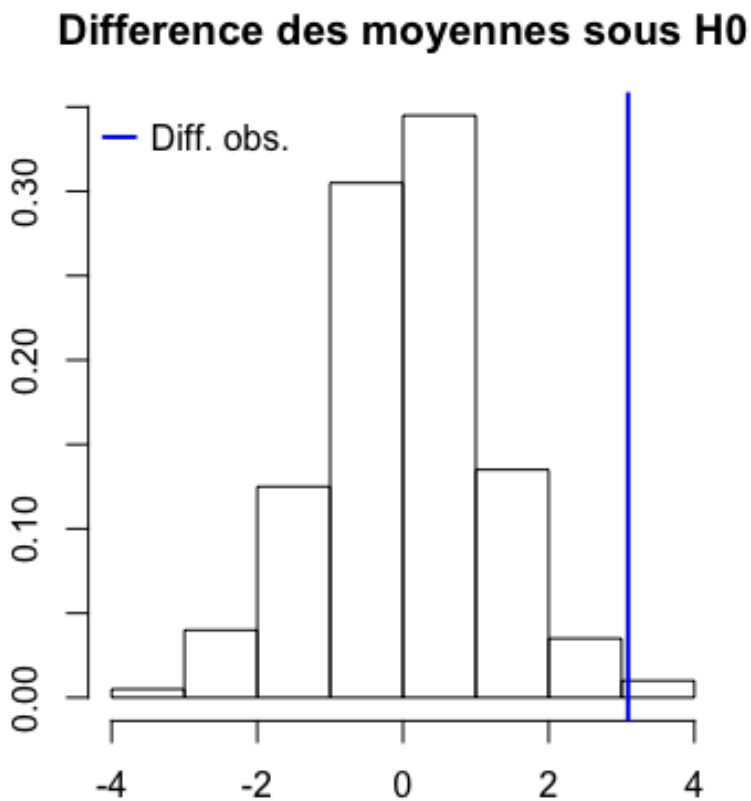


The value of the difference observed in the sample of data is shown in blue.

There are very few values above the observed mean difference. Does this support the null hypothesis ?

Difference between means under H_0

We start again, but this time performing 200 permutations. We then plot the histogram of the differences in proportions simulated by permutation.
The histogram represents the distribution of differences in the average number of beats under the null hypothesis.



The value of the difference observed in the sample of data is shown in blue.

According to the plot, the observed value has a low probability.
We can conclude that H_0 is wrong.

What is the error risk associated to this decision ?

Error risk

Concept of error An error is made when the wrong conclusion is drawn.

There are two ways to make a mistake :

	Rejection of H_0	Non-rejection of H_0
H_0 true	**Type 1 error**	no error
H_0 false	no error	**Type 2 error**

It is generally difficult to measure Type 2 error because the alternative is too general (H_0 false : $\mu_C - \mu_{NC} > 0$).

However, **we know how to estimate the probability of being wrong when we reject** H_0 (i.e., when we decide that caffeine has an effect).

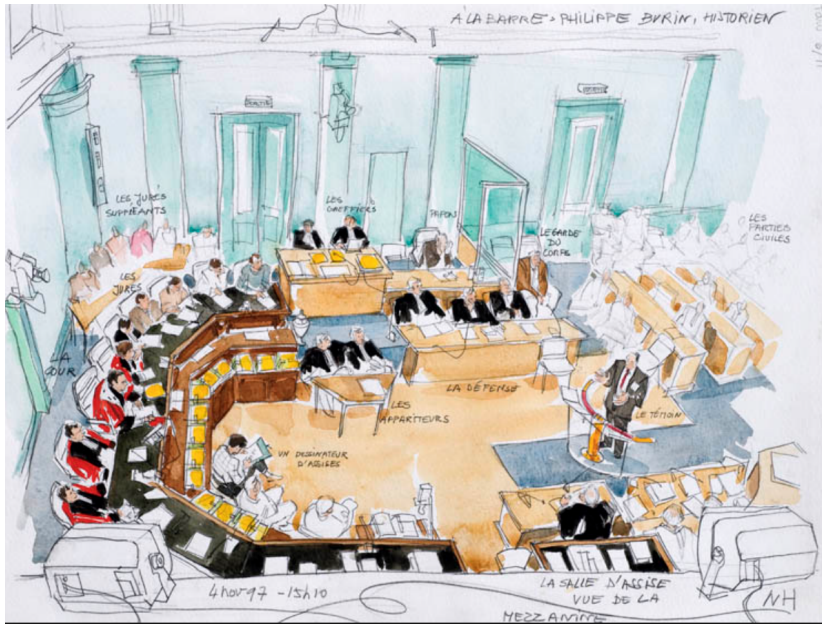
We have seen that we decide that the null hypothesis is false if the results of the observation are extreme under the null hypothesis. The significance level (or **p-value**) gives us a formal way of measuring the “strength” of the evidence that a sample provides against the null hypothesis and in support of the alternative hypothesis.

Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement into statistical testing terms.

"There are two types of errors a jury can make :
- Acquitting a guilty person
- Convicting an innocent person"

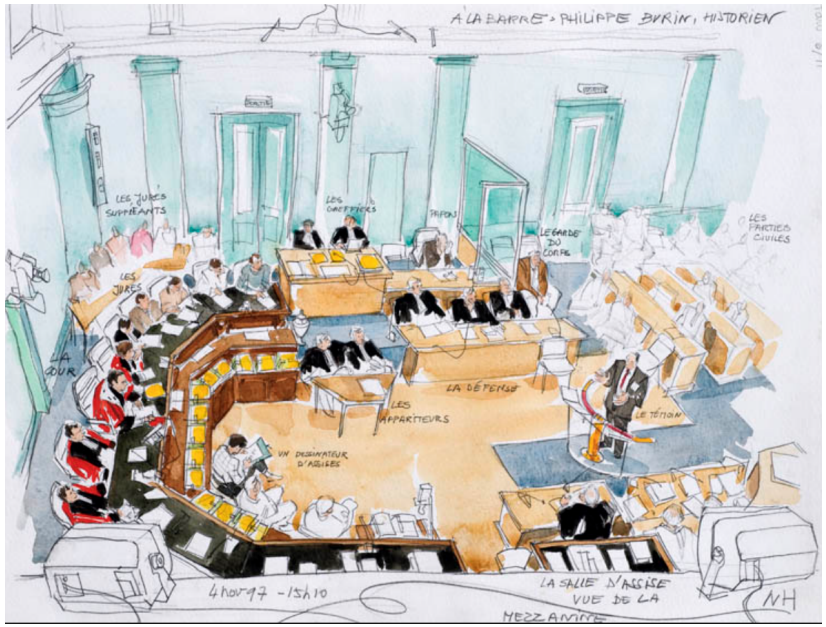


Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement into statistical testing terms.

"There are two types of errors a jury can make :
- Acquitting a guilty person
- Convicting an innocent person"



"Acquitting a guilty person" corresponds to a Type II error, because we find no evidence to reject a false H_0.

"Convicting an innocent person" corresponds to a Type I error, since we (incorrectly) find evidence in the data to reject a true H_0.

In our legal system, we are generally more concerned with a Type I error (convicting an innocent person) than a Type II error (acquitting a guilty person). As in our legal system, there is a trade-off between the two types of errors when we test hypotheses.

The p-value

The p-value in a statistical test is the probability, when the null hypothesis is true, of obtaining a sample as extreme as (or more extreme than) the observed sample.

The smaller the **p-value**, the stronger the statistical evidence is **against the null hypothesis** and in favor of the alternative.

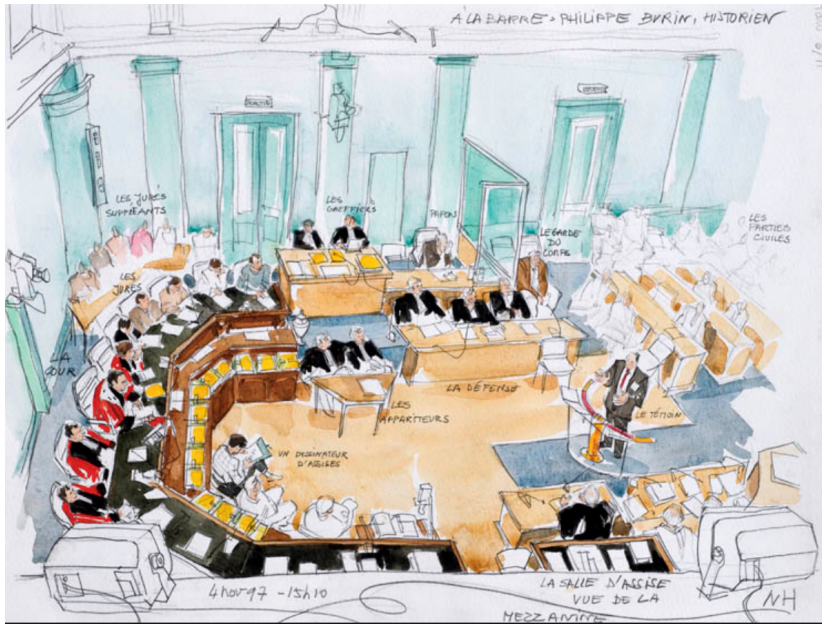
We also say that
the p-value is the probability of being wrong when rejecting H_0.

Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement into statistical testing terms.

"The evidence provided must indicate the suspect's guilt beyond a reasonable doubt."



The "evidence" is the sample data (statistics) and its probability of occurring if the individual is innocent (p-value).

Number of beats according to the Caffeine/Non-caffeine group

$$H_0 : \mu_C = \mu_{NC} \text{ versus } H_1 : \mu_C - \mu_{NC} > 0$$

P-value is the probability, calculated under the hypothesis H_0, that the random variable $Z = \widehat{\mu}_C - \widehat{\mu}_{NC}$ is greater than the **observed test statistic**

$$z_{obs} = \widehat{\mu}_C^{obs} - \widehat{\mu}_{NC}^{obs}$$

This probability is estimated using random permutations :

$$\text{p-value} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\widehat{\mu}_C^{(b)} - \widehat{\mu}_{NC}^{(b)} > \widehat{\mu}_C^{obs} - \widehat{\mu}_{NC}^{obs}}$$

where B is the number of permutations repeated.

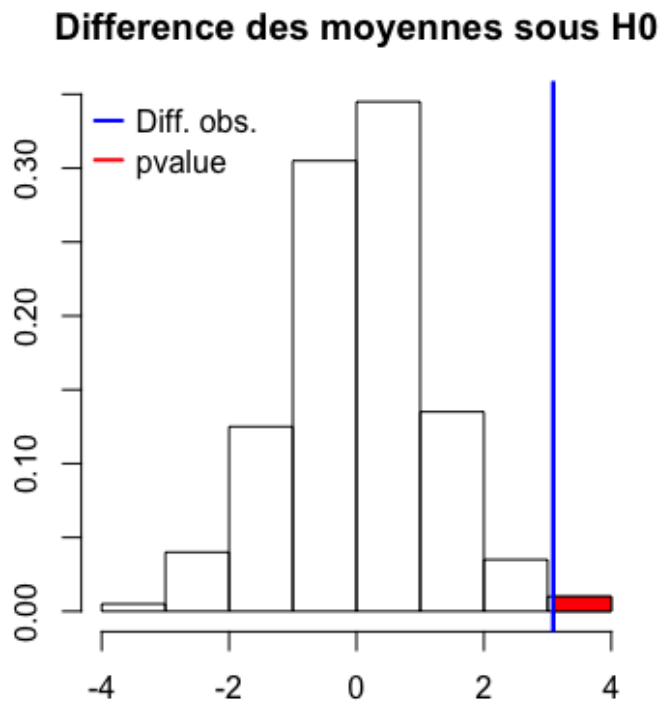
With $B = 10000$, we obtain p-value = 0.01
i.e., there is a 1 in 100 chance of being wrong when we say that caffeine has an effect on heart rate.

Graphic for p-value

$$H_0 : \mu_C = \mu_{NC} \text{ contre } H_1 : \mu_C - \mu_{NC} > 0$$

P-value is the probability, computed under H_0 that the random variable $Z = \widehat{\mu}_C - \widehat{\mu}_{NC}$ is greater than **the observed test statistic**

$$z_{obs} = \widehat{\mu}_C^{obs} - \widehat{\mu}_{NC}^{obs}$$

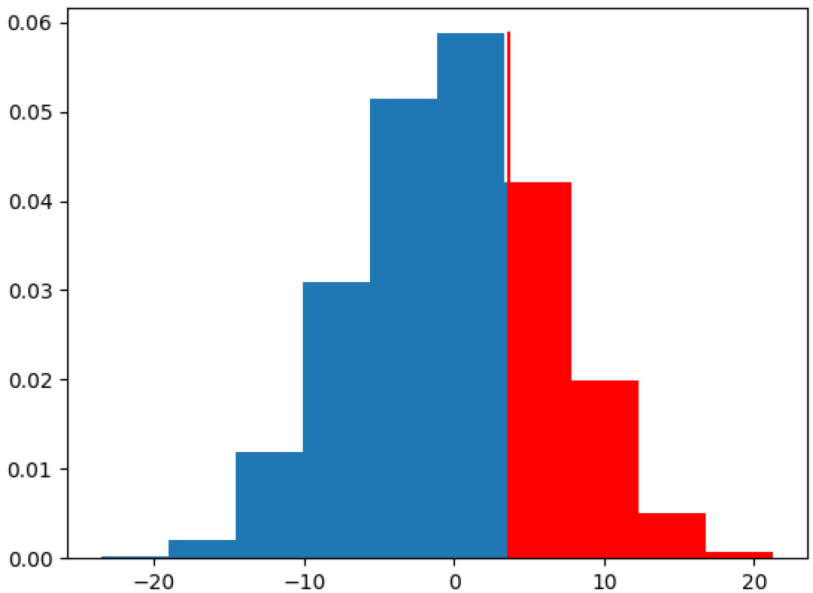


Example : Hollywood

A journalist claims that the budget for horror films made in Hollywood is lower than that for comedies. The average budget for the top 15 comedies made in 2011 was : 3.2, 8, 10, 19, 20, 21, 25, 28, 32.5, 35, 36, 40, 45, 55, 79 ;

that of the top 17 horror films were as follows (in dollars) : 1.4, 4, 5, 5, 16, 20, 25, 25, 30, 37, 38, 40, 40, 42, 50, 50

Using a permutation test, we obtain the histogram below for the test statistic under H_0, where the p-value is shown in red [3]



Que pensez-vous de l'affirmation du journaliste ? Justifier votre réponse.

3. The total area of the histogram is equal to 1.

Significativity

Significance level

For a hypothesis test, the significance level is a fixed value α such that if the p-value is less than α, H_0 is rejected.

The most common significance levels are $\alpha = 0.05$, $\alpha = 0.01$ or $\alpha = 0.10$.

When making a formal decision in a statistical test based on a sample

Reject H_0 — p-value $\leq \alpha$
ie a statistic associated with the sample that is so extreme is unlikely when H_0 is true. This means that we have found evidence to support H_1.

Do not reject H_0 — p-value $> \alpha$
ie the statistic associated with the sample is not too extreme when H_0 is true. This means that the test is inconclusive and that either H_0 or H_1 may be true.

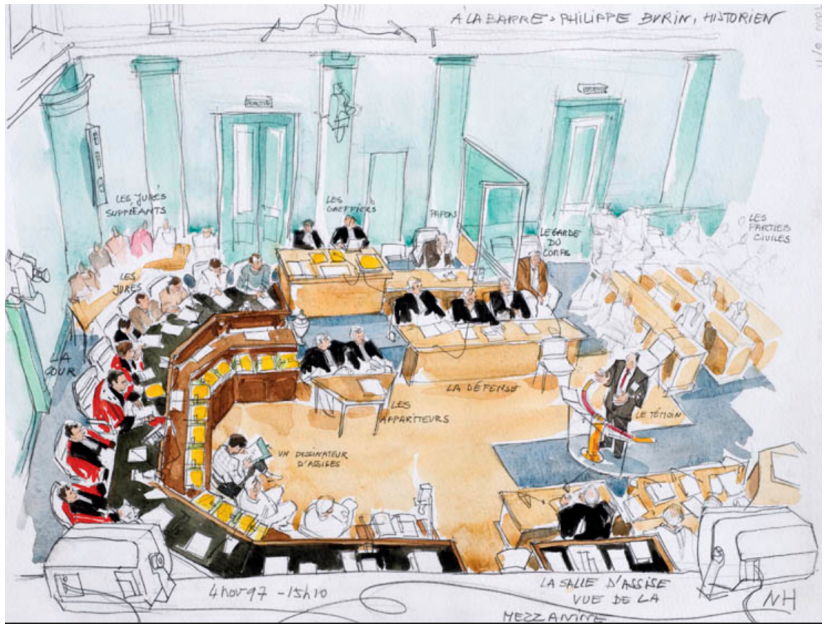
In both cases, be sure to interpret the decision in the context of the experiment.

Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement in terms of statistical testing.

“The evidence provided must indicate the guilt of the suspect beyond a reasonable doubt.”

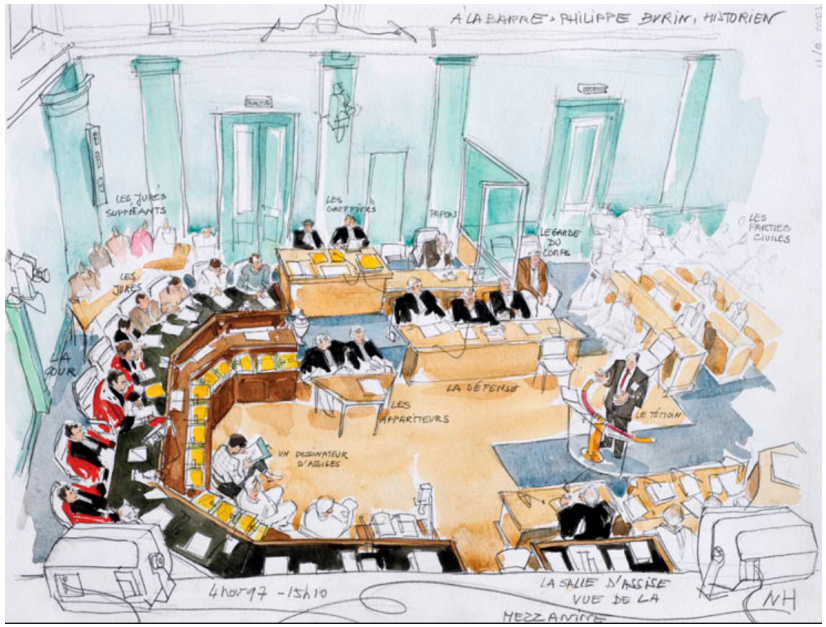


Analogy with justice

It is often useful to think of hypothesis testing as similar to court cases.

Translate the following statement in terms of statistical testing.

"The evidence provided must indicate the guilt of the suspect beyond a reasonable doubt."



"Beyond a reasonable doubt" corresponds to the significance level α. We reject the claim of innocence (H_0) and determine that the suspect is guilty (H_1) when the evidence (p-value) has a very low probability of occurring (less than α) if the suspect is truly innocent.

pvalues in more general cases

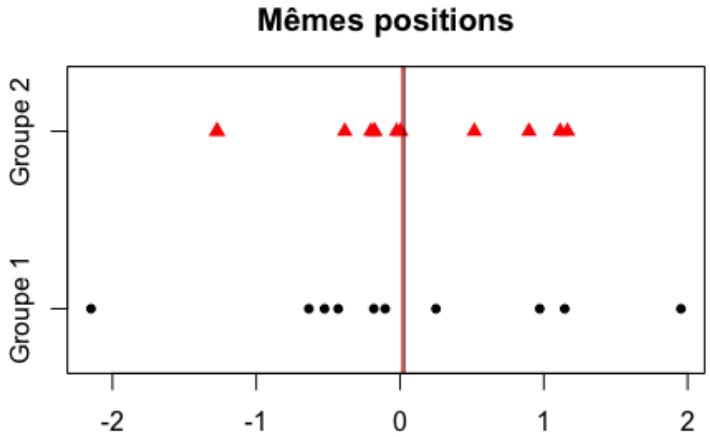
Let us denote S_n the test statistic and assume the S_n ditribution is F_n.

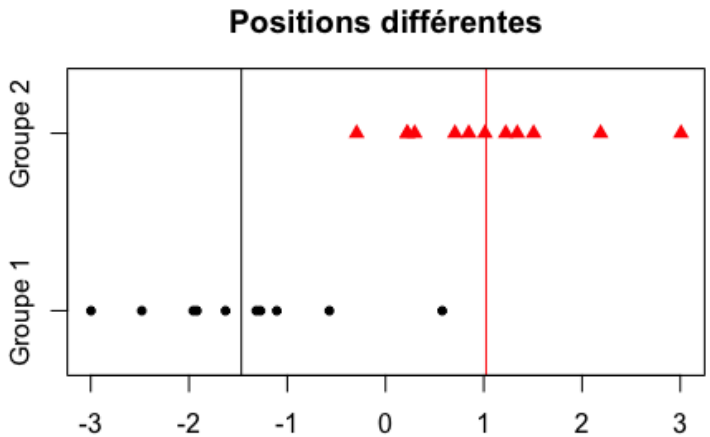
- irlughlwdkjfv

Outline

Nonparametric tests for position

- When the observed samples are small (n<30) [4], nonparametric (or distribution-free) tests are used.
- To compare/test positions
  1 or 2 groups : Wilcoxon-Mann-Whitney test (or Mann Whitney U test)
  more than 2 groups : Kruskal–Wallis test

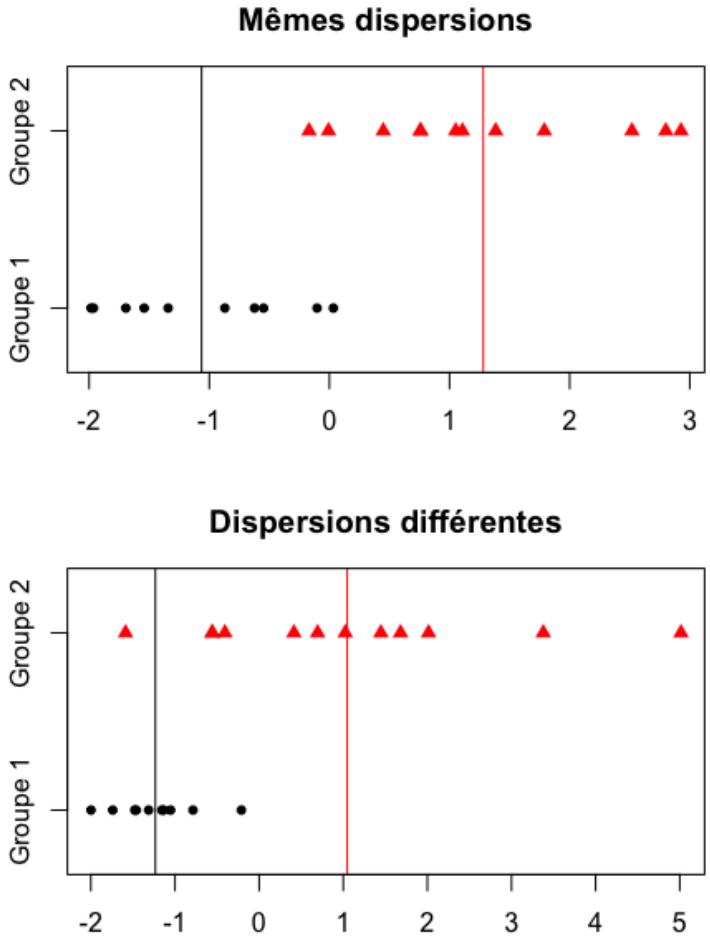




4. *and the assumption of normality is not verified*

Nonparametric tests for dispersion

- When the observed samples are small (n<30) [5], nonparametric (or distribution-free) tests are used.
- To **compare/test dispersions** : **Siegel-Tukey test**



5. *and the assumption of normality is not verified*

Nonparametric tests of goodness of fit to a distribution or comparison of distributions

- Goodness-of-fit tests
  discrete distributions : chi2 test
  continuous distributions : Shapiro-Wilk test for the Gaussian distribution, Kolmogorov test in other cases
- Tests for comparing distributions : Kolmogorov-Smirnov test

Outline

Test de Wilcoxon-Mann-Whitney

- The Mann Whitnay Wilcoxon test aims at comparing the position of two distributions $\mathbf{x} = \{x_1, \cdots, x_n\}$ et $\mathbf{y} = \{y_1, \cdots, y_m\}$.

$$H_0 : \theta_x = \theta_y \text{ contre } H_1 : \theta_x \neq \theta_y$$

  where θ is for instance the median.
- No assumption about the distribution is requiered.
- The idea is similar to the one of permutation tests : the two samples are gathered into one $\mathbf{z} = \{\mathbf{x}, \mathbf{y}\}$ and a rank is attributed to each observation $\mathbf{z}$:

$$z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n+m)}$$

- The Wilcoxon-Mann-Whitney test statistic is

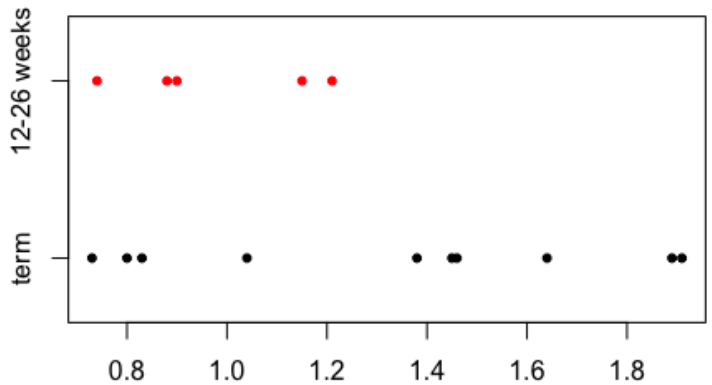
$$S_n = R_x - \frac{n(n+1)}{2}$$

  where R_x is the sumof ranks from sample $\mathbf{x}$.
  If both samples have the same position, then Si les deux échantillons ont la même position, ranks from $\mathbf{x}$ and $\mathbf{y}$ are equal and,
  $S_{n,m} \simeq (n+m)(n+m+1)/4 - \frac{n(n+1)}{2} = nm/2$.
- Thera are some tables of the distribution of the test statistics.

Test de Wilcoxon-Mann-Whitney, exemple

Des constantes de perméabilité d'une membrane du placenta sont mesurées à term ($p12$) et entre 12 et 26 semaines de gestation ($p26$)

$$H_0 : \theta_{term} = \theta_{12-16} \text{ contre } H_1 : \theta_{term} > \theta_{12-26}$$



- If H_0 is true, it is expected that $S_{n_x} \simeq 25 * 26/4 = 31.5$.
- Or $S_{n_x} = 35$ et d'après la table $pv = 0.2$.

```
p12 = np.array([0.80, 0.83, 1.89, 1.04,
1.45, 1.38, 1.91, 1.64, 0.73, 1.46])
p26 = np.array([ 1.15, 0.88, 0.90, 0.74, 1.21])

import statsmodels
from statsmodels.stats.nonparametric import rank_compare_2indep
statsmodels.stats.nonparametric.rank_compare_2indep(p12, p26)
```

Mann-Whitney test, special cases

- Tie
  When several observations are equal, they are assigned their average rank.
  Example of the effect of caffeine

C						245		246	$\cdots$
NC	242	242	242	244	244	245		246	$\cdots$
Rank	1	2	3	4	5	6	7	8	9	$\cdots$
Average rank	2	2	2	4.5	4.5	6.5	6.5	8.5	8.5	$\cdots$

- Large samples

$$U = \min(S_x, S_y)$$

$$\frac{U + 1/2 - nm/2}{\sqrt{(n+m)(n+m+2)/12}} \sim \mathcal{N}(0,1)$$

Paired samples

Samples are said to be paired if the observations are coupled.

- Measurements before and after treatment on the same individuals.
- Observations on siblings (or fathers/sons)

In this case, we calculate the difference between the two measurements and compare the position of the difference to 0.

Example : Hamilton Depression Rating Scale, for 9 patients with symptoms of anxiety and depression, measured before (x) and after administration of a sedative.

$$H_0 : \theta_{av} - \theta_{ap} = 0 \text{ versus } H_1 : \theta_{av} - \theta_{ap} > 0$$

before	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
after	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
difference	0.952	-0.147	1.022	0.430	0.620	0.590	0.490	-0.080	0.010

```
> x = c(1.83, 0.50, 1.62, 2.48,1.68,1.88,1.55,3.06,1.30)
> y = c(0.878,0.647,0.598,2.05,
1.06, 1.29, 1.06, 3.14, 1.29)
> wilcox.test(x, y, paired = TRUE,  alternative = "greater")
Wilcoxon signed rank test
data:  x and y
V = 40, p-value = 0.01953
alternative hypothesis: true location
shift is greater than 0
```

Outline

Comparison of 3 or more unpaired samples

- The **Kruskall-Wallis test** allows us to compare the position of 3 or more samples $\mathbf{x_1} = \{x_{11}, \cdots, x_{1n_1}\}, \cdots, \mathbf{x_k} = \{x_{k1}, \cdots, x_{kn_k}\}$

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \text{ versus } H_1 : \text{at least one of the } \theta \text{ is different}$$

- The Kruskall-Wallis test performs a nonparametric analysis of variance. It is used for small samples or when the normality assumption is not verified.
- The test statistic is given by

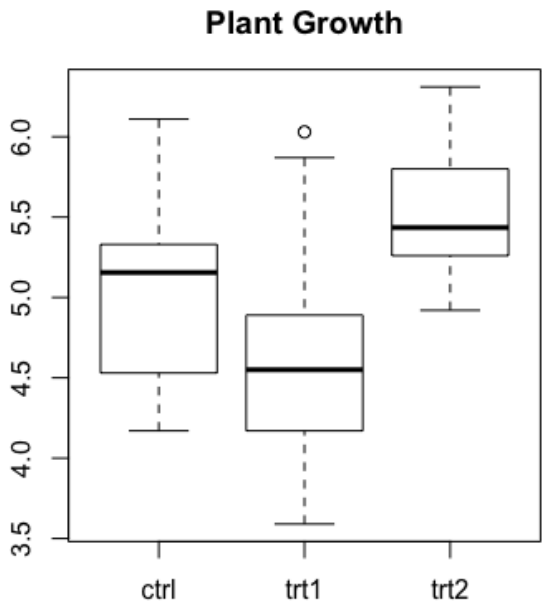
$$S = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_{i.}^2}{n_i} - 3(n+1)$$

where $n = \sum_{i=1}^{k} n_i$ is the total number of observations and $R_{i.}$ is the sum of the ranks of sample i.

- The test statistic distribution is tabulated for small samples.
- For large samples, the test statistic follows a χ^2 distribution with $k-1$ degrees of freedom.
- In the event of a tie, there is an adjusted formula.

Example

Yield comparison experiment (dry plant weight).



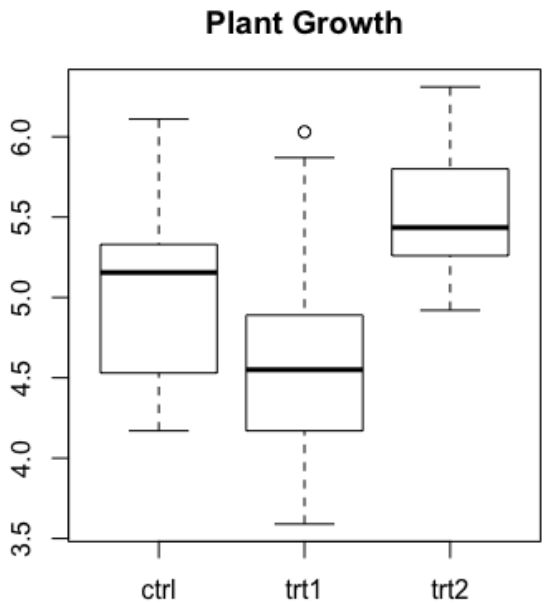
- The boxplots show a difference in distributions between groups.
- The Kruskal-Wallis test returns a p-value of around 0.02.
  Thus, at a 5% risk, we reject the hypothesis of equal distributions.

```
> boxplot(weight~group,data=PlantGrowth,main="Plant Growth")
> kruskal.test(weight~group,data=PlantGrowth)

Kruskal-Wallis rank sum test
data:  weight by group
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842
```

Example (continued)

Yield comparison experiment (dry plant weight).



- The boxplots show a difference in distributions between groups.
- The Kruskal-Wallis test returns a p-value of around 0.02.
  Therefore, at a 5% risk, we reject the hypothesis of equal distributions.

- We now look for significant differences between groups.
- We can use the Wicoxon-Mann-Whitney test to make multiple comparisons. However, we must apply the **Bonferonni correction**.

Outline

Bonferroni correction

- If several hypotheses are tested, the probability of a rare event increases and, consequently, the probability of incorrectly rejecting a null hypothesis (i.e., committing a Type I error) increases.

- The Bonferroni correction compensates for this increase by testing each hypothesis at a significance (or risk) level of α/k, where α is the desired overall level and k is the number of hypotheses.

- In the previous example, there are 3 groups, so $k = 3$. If we perform multiple comparisons at a 5

Correction de Bonferonni, exemple (suite)

```
> pairwise.wilcox.test(PlantGrowth$weight, PlantGrowth$group,
+                      p.adjust.method = « bonferroni »)
    Comparaisons par paires à l'aide du test de somme des rangs de Wilc
     ctrl  trt1
trt1 0,596 -
trt2 0,189 0,027

> grp = PlantGrowth$group
> wgt = PlantGrowth$weight

> wilcox.test(wgt[grp==« ctrl »],wgt[grp==« trt1 »])
Test de somme des rangs de Wilcoxon avec correction de continuité
W = 67,5, valeur p = 0,1986

> wilcox.test(wgt[grp==« ctrl »],wgt[grp==« trt2 »])
    Test de somme des rangs de Wilcoxon
W = 25, valeur p = 0,06301

> wilcox.test(wgt[grp==« trt1 »],wgt[grp==« trt2 »])
    Test de somme des rangs de Wilcoxon
W = 16, valeur p = 0,008931
```

Outline

Friedman tests, more than 2 matched samples

- The **Friedman test** allows you to compare the positions of more than 2 matched samples.
- Example : Effect of hydralazine on lung problems ? (repeated data).

Obs.	Before	48 hours after	6 months after
1	22.2	5.4	10.6
2	17.0	6.3	6.2
3	14.1	8.5	9.3
4	17.0	10.7	12.3

- **Test hypotheses**

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J \text{ versus } H_1 : \text{at least one of } \theta \text{ is different}$$

where θ_j is generally the median.

Friedman tests

- **Test hypotheses**

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_J \text{ versus } H_1 : \text{at least one of } \theta \text{ is different}$$

where θ_j is generally the median.

- To construct the test statistic, ranks are assigned by row (i.e., for each individual) and then the sum of these ranks is calculated by column.
If the null hypothesis is true, we expect the sums of the ranks by column to all be of the same order.

Obs.	Before		48 hours after		6 months after	
	Value	Rank	Value	Rank	Value	Rank
1	22.2	3	5.4	1	10.6	2
2	17.0	3	6.3	2	6.2	1
3	14.1	3	8.5	1	9.3	2
4	17.0	3	10.7	1	12.3	2
Sum		12		5		7

- The **test statistic** is then written as

$$\mathcal{F} = \frac{12}{nJ(J+1)\sum_{j=1}^{J} r_j^2 - (3n(J-1)}$$

where n is the number of subjects and r_j is the sum of the ranks in column j

- The distribution of $\mathcal{F}$ is given in Friedman's table.

Friedman test, example

Obs.	Before		48 hours after		6 months after	
	Value	Rank	Value	Rank	Value	Rank
1	22.2	3	5.4	1	10.6	2
2	17.0	3	6.3	2	6.2	1
3	14.1	3	8.5	1	9.3	2
4	17.0	3	10.7	1	12.3	2
Sum		12		5		7

- We obtain $\mathcal{F} = 6.5$.
- According to the table, the test is significant at a risk of $\alpha = 5\%$ if $\mathcal{F} > 6.5$ for n=4 and J=3.

k=3

N	α <.10	α ≤.05	α <.01
3	6.00	6.00	–
4	6.00	6.50	8.00
5	5.20	6.40	8.40
6	5.33	7.00	9.00
7	5.43	7.14	8.86

We conclude that hydralazine has a significant effect.

Friedman test, example

- Friedman test

```
> y = matrix(c(22.2,17.0,14.1,17.0,5.4,6.3,8.5,10.7,10.6,6.2,9.3,12.
+            nrow=4,ncol=3,
+            dimnames = list(1:4,c("before", '48h', "6 months")))
> friedman.test(y)

    Friedman's sum of ranks test

data:  y
Friedman's chi-square = 6.5, df = 2, p-value = 0.03877
```

- Multiple comparisons

```
> wilcox.test(y[,1],y[,2],paired=TRUE,alternative="greater")$p.value
[1] 0.0625
> wilcox.test(y[,2],y[,3],paired=TRUE,alternative="greater")$p.value
[1] 0.9375
> wilcox.test(y[,1],y[,3],paired=TRUE,alternative="greater")$p.value
[1] 0.0625
```

Only four individuals are observed : we are at the limit of the validity of the Wilcoxon test (six individuals).

Outline

Parametric tests

- Parametric tests also allow us to compare positions and dispersions.
- They are based on statistics calculated from observations.
  For example, the empirical mean

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

- Unlike statistics on the ranks of observations, we can only determine the distribution of these (empirical) statistics in certain specific cases ; typically Gaussian samples if $n < 30$.
- Comparison of means : **Student's t-test** and **analysis of variance**
- Comparison of variances : **Fisher-Snedecor test**

- Parametric tests are more powerful [6] than nonparametric tests. They should therefore be preferred.

6. A test is $\mathcal{T}_1$ **more powerful** than a test $\mathcal{T}_1$ if, for the same risk level α, its type II error is lower.

Outline

Student's t-test (comparison with a reference value)

- Student's t-test is a test for **comparison of means**.
- Let $\{X_1, \cdots, X_n\}$ be n independent random variables with the same distribution F and unknown mean μ.
- Based on a realization $\mathbf{x} = \{x_1, \cdots, x_n\}$, we would like to decide whether μ is significantly greater than a reference mean μ_0.
- We therefore set the **test hypotheses**

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

We can also consider the one-sided alternative $H_1 : \mu < \mu_0$ or the two-sided alternative $H_1 : \mu \neq \mu_0$.

- The **test statistic** is

$$T = \frac{\bar{X} - \mu_0}{\sqrt{s^2/n}}$$

where

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

- Reminder : to choose between the two hypotheses, we need to calculate the p-value. However

$$\text{p-value} = P_{\mu=\mu_0}(T \geq T_{obs})$$

T_{obs} is the value of T for the observations.

Test statistic law

- Reminder : to choose between the two hypotheses, we need to calculate the p-value

$$\text{p-value} = \mathbb{P}_{\mu=\mu_0}(T \leq T_{obs})$$

- To calculate this probability, we therefore need to know the distribution of S under H_0.

- **If X_i follows a Gaussian distribution** (also called a normal distribution), then the distribution of T is a Student's distribution with $(n-1)$ degrees of freedom.
- **If n is large**, then we can approximate the distribution of T by a Gaussian distribution with mean 0 and variance 1.

- In other cases, we do not know how to write the distribution of T and we use a nonparametric test.

Normality tests

- We have seen that we know the distribution of the Student's t-test statistic if X_i follows a Gaussian distribution.
- In practice, we use a normality test to verify this hypothesis.
- **Normality test**

$$H_0 : X \textit{ follows a Gaussian distribution} \quad \text{against} \quad H_1 : X \textit{ follows a different distribution}$$

- There are several normality tests, but the most powerful [7] is the Shapiro-Wilk test.
- Its test statistic is written as the ratio of two variance estimators that are equal to each other only if X follows a Gaussian distribution.

7. A test is $\mathcal{T}_1$ **more powerful** than a test $\mathcal{T}_1$ if, for the same risk level α, its type II error is lower.

Student's t-test, example

- Example : measurement of depression after administration of treatment

before	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
after	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
difference	0.952	-0.147	1.022	0.430	0.620	0.590	0.490	-0.080	0.010

We want to know if the treatment is effective.

- Test hypotheses

$$H_0 : \mu_{diff} = 0 \text{ versus } H_1 : \mu_{diff} < 0$$

- The sample is small, so we will test normality at a risk of $\alpha = 5\%$ before choosing the comparison test.

```
> x = c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
> y = c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)
> z = y-x
> shapiro.test(z)
    Shapiro-Wilk normality test
data:  z
W = 0.92139, p-value = 0.4039  # W: test statistic
```

- The p-value of the normality test is greater than α, so we can assume that the sample is Gaussian.

Student's t-test, example

- Example : measurement of depression after administration of treatment

before	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
after	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
difference	-0.952	0.147	-1.022	-0.430	-0.620	-0.590	-0.490	0.080	-0.010

We want to know if the treatment is effective.

- Test hypotheses

$$H_0 : \mu_{diff} = 0 \text{ versus } H_1 : \mu_{diff} < 0$$

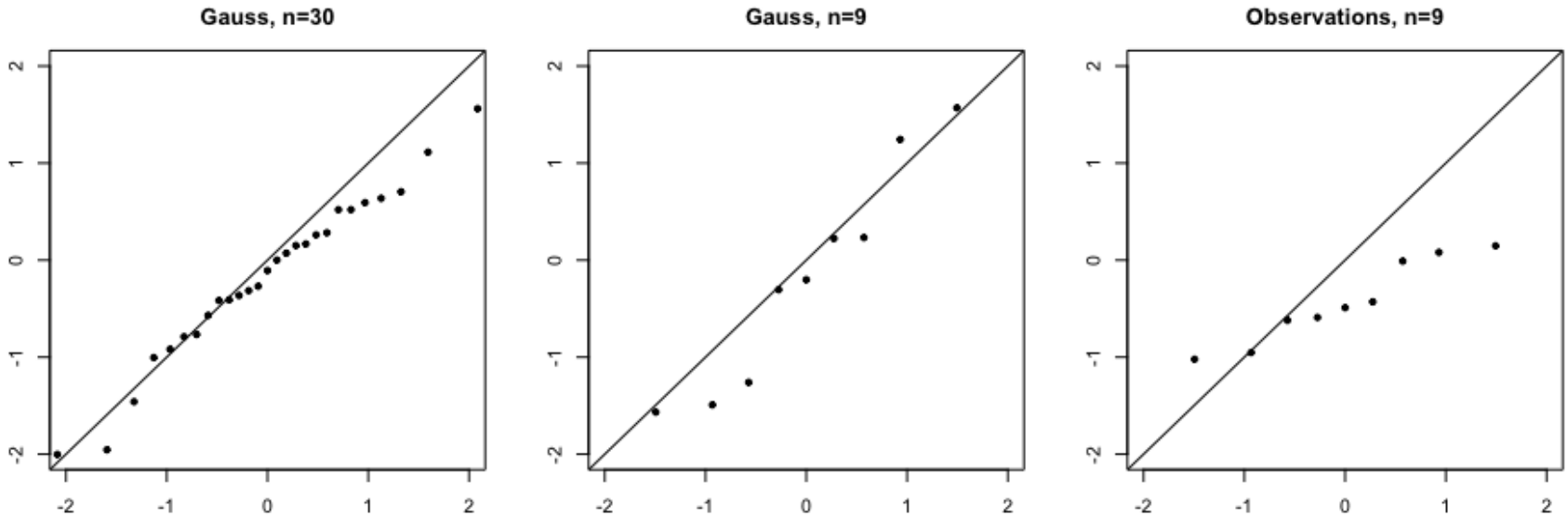
- The p-value of the normality test is greater than α, so we can assume that the sample is Gaussian.
- We therefore choose the Student's t-test (which is more powerful than a non-parametric test).

```
> t.test(z,alternative="lower") # other alternatives can be chosen
    One Sample t-test
data:  z
t = -3.0354, df = 8, p-value = 0.008088 # t: test statistic
alternative hypothesis: true mean is less than 0
```

- The p-value is around 0.01, so the test is significant, meaning that the treatment is effective.

Warning for normality testing

- When samples are (very) small, it is often difficult to measure deviation from normality.
- A common graphical representation for assessing deviation from normality is the **quantile-quantile plot (qqplot)**.



If the sample is from a Gaussian distribution, then the points should be close to the diagonal line.

- For depression observations, the Shapiro test is not significant, i.e., we do not reject the hypothesis of normality, yet the qqplot does not support this hypothesis.

Student's t-test, comparison of two means

- A Student's t-test is also used to compare the means of two independent samples.
- Let $\{X_1, \cdots, X_n\}$ be n independent random variables with the same distribution F with unknown mean μ_X and $\{Y_1, \cdots, Y_m\}$ m independent random variables with the same distribution G with unknown mean μ_Y
- Based on the realizations $\mathbf{x} = \{x_1, \cdots, x_n\}$ and $\mathbf{y} = \{y_1, \cdots, y_m\}$, we would like to decide whether μ_X and μ_Y are significantly different or not.
- We therefore set the **test hypotheses**

$$H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X > \mu_Y$$

  We can also consider the one-sided alternative $H_1 : \mu_X < \mu_Y$ or the two-sided alternative $H_1 : \mu_X \neq \mu_Y$.

- The **test statistic** is

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S}}$$

  where

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

  and S is a variance (see below).

Student's t-test distribution

The definition of the test statistic T and its distribution depend on the situation

- The variances of X and Y are equal or not.
- The variables X and Y follow a Gaussian distribution or not.
- The sample sizes n and m are large or not.

Student's t-test, equal variances

- If **the variances are equal**, then the variance S is defined by

$$S = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{(n+m-2)}$$

- If **the variables X and Y follow a Gaussian distribution**, then the distribution of T is a Student's t-distribution with $n+m-2$ degrees of freedom.

- We test the equality of variances using a Fisher-Snedecor test.

Outline

Fisher-Snedecor test for comparing variances

- The Fisher-Snedecor test is a parametric test that allows us to compare the variances of two independent samples. Let $\{X_1, \cdots, X_n\}$ be n independent random variables with the same distribution F and unknown variance σ_X^2 and $\{Y_1, \cdots, Y_m\}$ m independent random variables with the same distribution G and unknown variance σ_Y^2.
- The **test hypotheses**

$$H_0 : \sigma_X = \sigma_Y \text{ versus } H_1 : \sigma_X \neq \sigma_Y$$

- The test is based on the following **test statistic**

$$F_{n,m} = \frac{s_X^2/(n-1)}{s_Y^2/(m-1)}$$

- If the samples are drawn from Gaussian distributions, then the distribution of $F_{n,m}$ is a Fisher distribution with $n-1, m-1$ degrees of freedom.
- For more than 2 samples, the **Bartlett test** is used.

Fisher-Sndecor test, example

79 urine samples were analyzed to determine whether certain physical characteristics of urine could be linked to the formation of calcium oxalate crystals.
Urine osmolarity. Osmolarity is proportional to the concentration of molecules in solution.

```
from statsmodels.stats.stattools import omni_normtest
from statsmodels.stats.weightstats import ztest

osmo_gp0 = df["osmo"].iloc[np.where(df["target"]==0)[0]]
osmo_gp1 = df["osmo"].iloc[np.where(df["target"]==1)[0]]

# Normality test by group
W0 = statsmodels.stats.stattools.omni_normtest((osmo_gp0-np.mean(osmo_g
print(" pvalue, gp 0 : ", W0[1]) # returns pv = 0.02
W1 = statsmodels.stats.stattools.omni_normtest(osmo_gp1)
print(" pvalue, gp 1 : ", W1[1]) # returns pv = 0.49

# Z test
comp_var = statsmodels.stats.weightstats.ztest(osmo_gp0, osmo_gp1)
print("z test pvalue : ", comp_var[1]) # returns pv = 2.99e-7
```

We conclude that we can assume that the variable `osmo` follows a Gaussian distribution in group 0 only and variances are equal.

Student's t-test for 2 samples, examples

79 urine samples were analyzed to determine whether certain physical characteristics of urine could be linked to the formation of calcium oxalate crystals.
Urine osmolarity. Osmolarity is proportional to the concentration of molecules in solution.

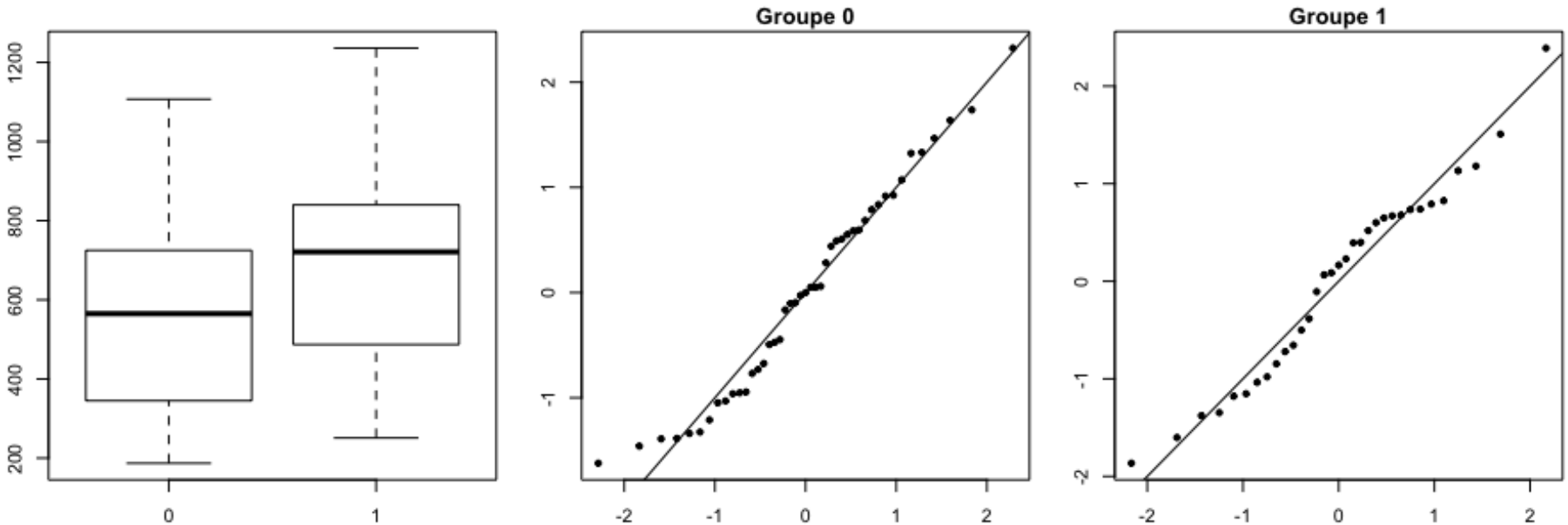
```
comp_mean = statsmodels.stats.weightstats.ttest_ind(osmo_gp1,
      osmo_gp0,
      alternative="larger")
print("t test pvalue : ", comp_mean[1]) # returns pv = 2.30e-7
```

We conclude that osmolarity is significantly higher in group 1 than in group 0 (control).

Fisher-Sndecor test, example

79 urine samples were analyzed to determine whether certain physical characteristics of urine could be linked to the formation of calcium oxalate crystals.
Urine osmolarity. Osmolarity is proportional to the concentration of molecules in solution.

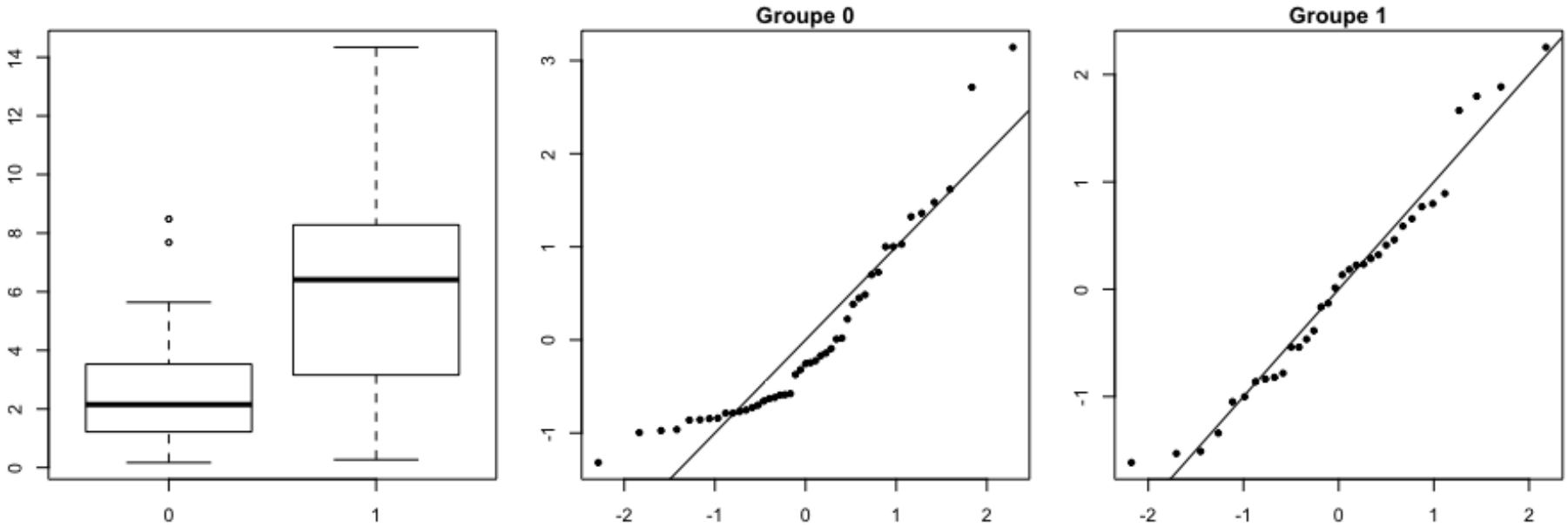
A graphical analysis allowed the results to be anticipated.



The boxplot shows similar dispersions in both groups and the qqplots show a slight deviation from normality.

Urine data (continued)

The data also contains (of course) a measurement of calcium content.



- The center graph shows a deviation from normality for the observations in group 0.
- The left graph shows that the variances of the two groups are different.

Urine data (continued)

For the calcium content of the control group, the Shapiro Wilk test confirms the graphical analysis : the p-value is around 1e-4.

```
> sh1 = shapiro.test(scale(urine$calc[urine$r==1]))
> print(paste("Shapiro test p-value for group 1:",
      round(sh1$p.value,2)))
[1] "Shapiro test p-value for group 1: 0.35"
> sh0 = shapiro.test(scale(urine$calc[urine$r==0]))
> print(paste("Shapiro test p-value for group 0:",
      round(sh0$p.value,4)))
[1] "Shapiro test p-value for group 0: 1e-04"
```

In this case, we should continue with a nonparametric test (see below) even though it is known that *the Student's test is not very sensitive to the non-normality of the parent populations and, for samples of the same size, to the inequality of variances.*

```
> W = wilcox.test(urine$calc[urine$r==1],urine$calc[urine$r==0],
      alternative="greater")
> print(paste("P-value of the Wilcoxon test:",round(W$p.value,6)))
[1] "P-value of the Wilcoxon test: 4e-06"
```

We conclude that the calcium content in the urine of group 1 is significantly higher than that of group 0.

Student's t-test, different variances

- If **the variances are different**, then the variance S is defined by

$$S = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

- If **the variables X and Y follow a Gaussian distribution**, then the distribution of T is a Student's distribution with κ degrees of freedom, with

$$\kappa = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}$$

- Note : if n and m tend towards infinity, the Student's distribution with κ degrees of freedom tends towards a Gaussian distribution.
  In the **case of large samples**, the Student's t-distribution can always be approximated by a Gaussian distribution with mean 0 and variance 1.

Outline

Required sample size : what criteria ?

If you want to compare a quantitative criterion (systolic blood pressure, biological data, etc.) between two treatment groups, you must define :

- the expected difference between the means of group A and group B,
- the standard deviations of these means,
- the risk α (5%, if you are making a single comparison ; in the case of multiple statistical tests, this risk must be adjusted),
- the power ($1 - \beta$, at least 80

Example : If you want to evaluate the effectiveness of a new treatment on systolic blood pressure compared to the standard treatment. In the literature, you have found the following hypotheses : the mean systolic blood pressure is 150.2 $\pm$ 40.6 mmHg (40.6 is the standard deviation) with the standard treatment and 140.5 mmHg with the new treatment with an identical standard deviation. For an alpha risk of 5

Online calculation :
`https://biostatgv.sentiweb.fr/?module=etudes/sujets#`

Required sample size : calculation

- If we denote Δ as the expected difference between the means of group A and group B

$$n = 2(z_\alpha - z_\beta)^2 \frac{\sigma^2}{\Delta^2}$$

- In general, we choose $\alpha = 5\%$ and $1 - \beta = 80\%$, hence
  Type I error : z_α = 1.96 for a two-tailed test and 1.69 for a one-tailed test
  Type II error : z_β = 0.84

References