Student's Guide on Statistical Analysis

Lynn Delcon
PhD student at the VUB
Lynn.Delcon@vub.be

1 Introduction

This draft aims to provide a guide to Master's students in conducting statistical analysis during their Master thesis. We explain the theory and provide python codes on the most used statistical analysis throughout the paper. We start with the description of data (Section 3), followed by the three conditions of application for parametric tests (Section 4):

- independence of observations (Subsection 4.1),
- normality (Subsection 4.2),
- homogeneity of variances (Subsection 4.3).

We then elaborate on the most frequent parametric tests:

- Confidence interval (exact and asymptotic) (Subsection 5.1),
- Student test (independent, Welch's t-test and dependent) (Subsection 5.2),
- One-way anova (classic and Welch's anova) (Subsection 5.3).

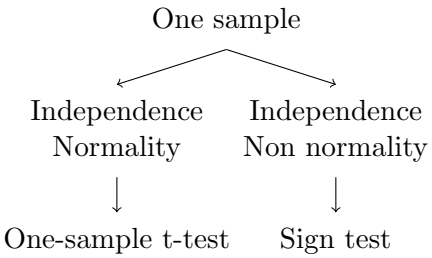
From those parametric tests, we provide their non parametric alternatives:

- Sign test (Subsection 6.1),
- Wilcoxon test (rank-sum and signed-rank) (Subsection 6.2 & 6.3),
- Paired Sign test (Subsection 6.4),
- Kruskal-Wallis test (Subsection 6.5).

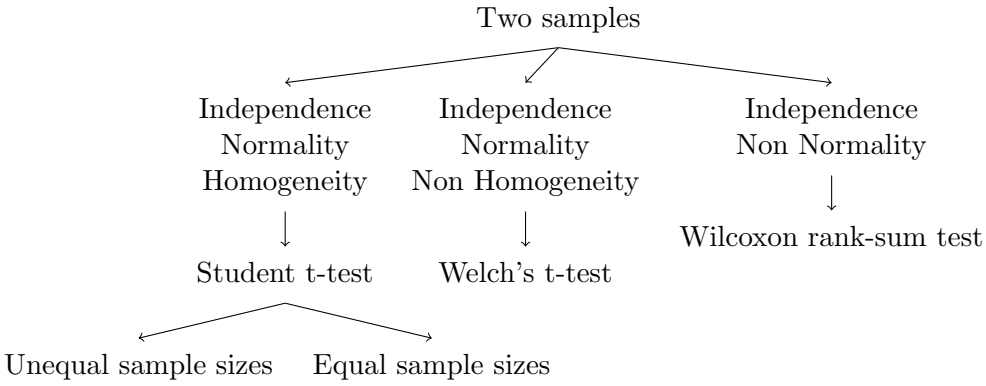
Finally, we conclude with some tips for the writing of the statistical part of a Master thesis. The following section provides an overview of the tests covered in this draft, notably, the situations specific to each test.

2 Overview

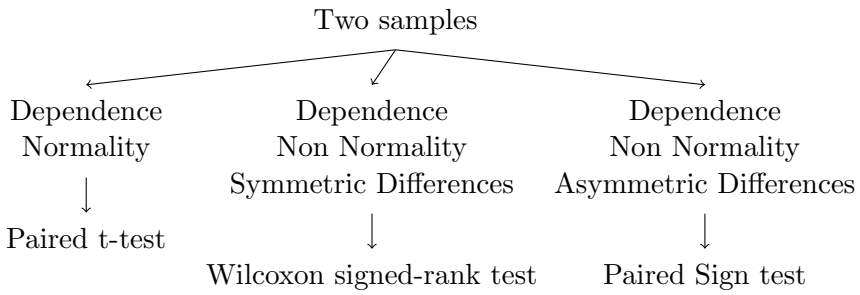
The confidence interval is of interest when we want to test the accuracy of an assumed normal distribution of mean μ and variance σ^2 to characterize our sample.
The Student test is a parametric test that compares the distribution of one sample to a theoretical mean (Subsection 5.2.1) or the distributions of two samples to each other according to their means. Regarding the one-sample t-test, the observations must be independent and normally distributed. If the sample is large enough ($n \geq 30$), we can use the Theorem Central Limit (Theorem 4.1) and bypass the normality condition. If the observations are not normal and the sample size is to small to apply the later theorem, we use the non parametric alternative, the sign test (Subsection 6.1).



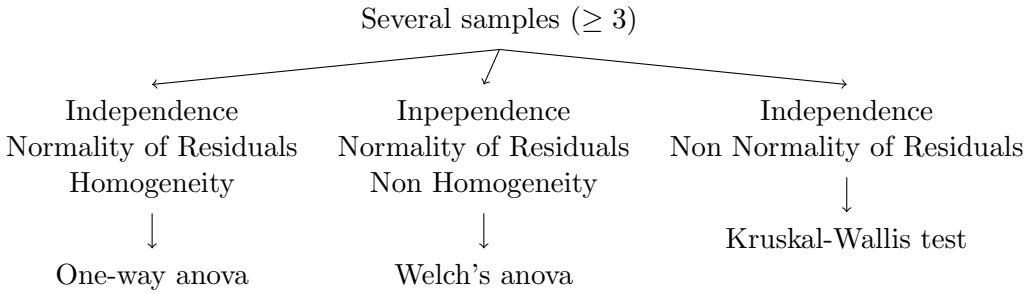
The Student test for two samples requires the independence (between and within samples, Subsection 4.1), normality and homogeneity of variances conditions. When all three conditions are met, the classic independent Student test (Subsection 5.2.2) provides statistics for equal and unequal sample sizes. If both samples satisfy the independence and normality conditions but the two variances are not homogeneous, we apply the Welch's t-test (Subsection 5.2.3). When the sample sizes are large enough ($n_1, n_2 \geq 30$), we can bypass the normality condition thanks to Theorem 4.1, but, if the sample sizes are small and the independent observations are not normal, we rather use the non parametric alternative test, the Wilcoxon rank-sum test (Subsection 6.2).



Finally, if both samples are dependent, we consider the paired Student test (Subsection 5.2.4) under the normality condition. If the normality is not satisfied but the differences between observations are symmetrically distributed around zero, we consider the non parametric alternative, the Wilcoxon signed-rank test (Subsection 6.3). On the other hand, if the differences are not symmetrically distributed around zero, we consider the paired Sign test (Subsection 6.4).



The one-way anova allows to compare several samples with respect to their means, at least two samples, but more often at least three samples. This parametric test is also built on the independence, normality of residuals and homogeneity of variances conditions. (We do not address the anova for dependent samples in this draft.) If the independence and the normality of residuals are met but not the homogeneity of variances, we use the Welch's anova (Subsection 5.4). If the residuals are not normally distributed, we use the non parametric alternative, the Kruskal-Wallis test (Subsection 6.5).



Parametric tests are more powerful than the non parametric ones if we meet the conditions of application. The power of a test refers to the ability to reject the null hypothesis when the later is false. If the **normality** condition of application is **not satisfied**, the **parametric tests** will provide **invalid results** while the non parametric ones deliver valid results. For small sample size, it is also recommended to use non parametric tests given that the normality condition is rarely verified with a few observations. Lastly, non parametric tests are robust to outliers, which is not the case for parametric tests.

Contents

3 Descriptive Statistics

The first stage of a statistical analysis is the visualization of our data. The typical tool to display the observations is the boxplot (Figure 1).

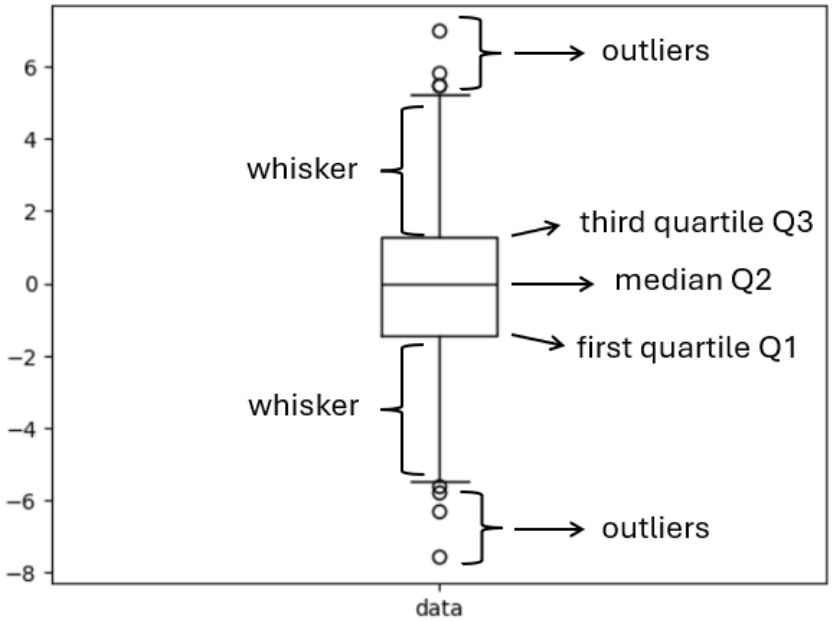


Figure 1: Boxplot of a normal sample ($\mu = 0$, $\sigma^2 = 4$) of size 1000.

Let us explain the legend of the above figure. First, recall that a quartile is the value below which a multiple of 25% of the observations are situated. We have four quartiles, the first one is denoted as Q_1 on the figure and delimits the value below which 25% of the observations are located. The second quartile refers to the value below which 50% of the observations lie. This second quartile is also called the median. The third quartile Q_3 delineates the value below which 75% of the observations are. Finally, the last quartile Q_4 represents the value below which all observations are situated.

Remark 3.1. (Quartiles) The value of a quartile, or a quantile (Definition B.1), does not necessary corresponds to an observation of the sample.

On the above figure, we can also observe the whiskers of the boxplot. Those whiskers are defined using the InterQuartile Range (IQR), that is

$$\text{IQR} = Q_3 - Q_1.$$

From the IQR, we delineate the lower whisker as

$$\text{lower whisker} = Q_1 - 1.5 \times \text{IQR},$$

and the upper whisker as

$$\text{upper whisker} = Q_3 + 1.5 \times \text{IQR}.$$

Those two whiskers, thresholds can be used to detect *outliers* for symmetric and light tailed distributions (Definition 3.2).

Definition 3.1. (Outlier) An outlier is defined as an observation whose value is *far* from the global trend.

The global trend can be characterized by the mean but also, and more accurately, the median. Indeed, the mean is highly influenced by the outliers (extreme values) while the median is not. The adjective *far* in the boxplot context refers to observations whose values are below the lower whisker or above the upper whisker.

Definition 3.2. (Tails of a distribution) The tails of a distribution refers to the two extremities of the density. For instance, the standard normal density ($\mu = 0$, $\sigma^2 = 1$) is a light tailed distribution. If the variance increases, we will have a more spread density implying heavier tails, that is, more area (observations) under both extremities, discarded from the mean.

If the distribution is rather asymmetric and/or heavy tailed, it is recommended to use the distance, thresholds, provided by the robust 3σ rule.

3.1 The Robust 3σ Rule (outlier)

This rule uses robust quantities to define a distance threshold below or above which a value is considered as an outlier. Those quantities are the median and the Median Absolute Deviation (MAD)[1] that are the robust equivalent of the mean and the standard deviation, respectively. From those two quantities we can define an outlier as

$$\text{Value}_{\text{outlier}} \leq \text{median} - 3 \times \text{MAD},$$

or

$$\text{Value}_{\text{outlier}} \geq \text{median} + 3 \times \text{MAD}.$$

This rule gives similar results to the boxplot whiskers if the distribution is symmetric and light tailed but provides more reliable outcomes if the distribution does not meet those conditions.

Python code:

```
import numpy as np
from scipy import stats
med = np.median(data)
MAD = stats.median_abs_deviation(data)
threshold_up = med + 3*MAD
threshold_low = med - 3*MAD
print(len(data > threshold_up),len(data < threshold_low))
```

3.2 Prior Information

The use of boxplots allows to capture some prior information about our samples, especially when we have several samples. Typically, for the parametric tests of at least 2 samples (e.g., Student and anova), displaying the boxplot of each sample allows to choose the kind of hypothesis: -bilateral, -unilateral greater, -unilateral lower. For instance, Figure 2 shows the distribution of two samples.

[1] https://en.wikipedia.org/wiki/Median_absolute_deviation

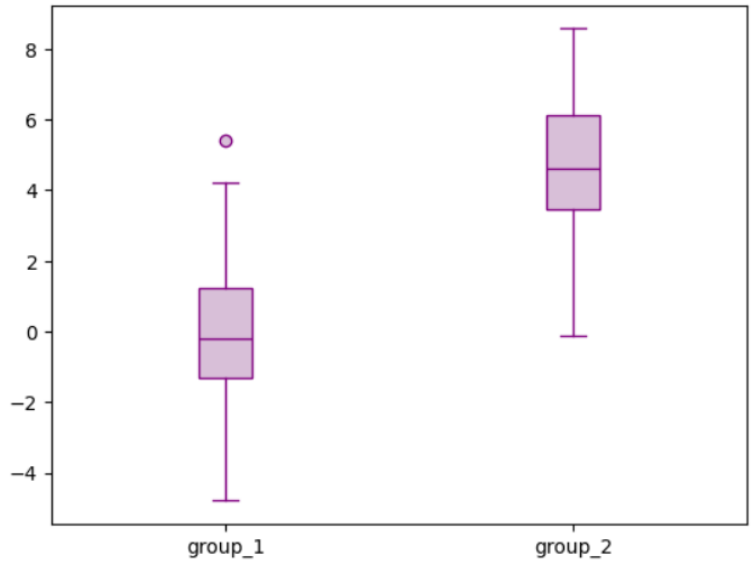


Figure 2: Boxplots of two samples.

From the above figure, we would test if the distribution of the first sample is significantly lower than the second sample. This comparison can be done according to the position measure: mean (parametric test, Section 5) or median (non parametric test, Section 6). Note that the choice of hypothesis can also be selected by comparing the mean or the median values without plotting any graph.

Python code:

```
import matplotlib.pyplot as plt
#one sample
plt.boxplot(data, medianprops=dict(color='black'))
plt.xticks([1], ['data'])
plt.show()
#two samples
data = [group_1,group_2]
plt.boxplot(data)
plt.xticks([1,2], ['group_1','group_2'])
plt.show()
```

3.3 Conclusion

The boxplot tool allows to rapidly have a holistic understanding of the sample distribution such as the symmetry or asymmetry of the distribution, the median, maximum and minimum values. It also displays potential outliers, those extreme values can be considered as outliers if the distribution is symmetric and light tailed but if the distribution is asymmetric and/or heavy tailed, we rather encourage the use of the robust 3σ rule to define outliers. If we indeed detect outliers, the safest is to apply non parametric tests that use robust parameters such as the median and the ranks (Section 6). Lastly, when we have several samples, the boxplot visualization also allows to select the kind of hypothesis, bilateral or unilateral (lower or greater).

4 Conditions of Application

In this section, we explain the importance of the classic conditions of application, especially for the normality and the homogeneity of variances. We also provide python codes to check

empirically those conditions on a data set.

4.1 Independence

Among all discussed tests, the independence within (inside) samples is always required but the independence between samples is not necessarily needed.
The condition of independence within or between samples is easily verified looking at the design of the experiment. The dependence between observations arises typically if the participants are linked, e.g., husband and wife. In this dependence design, there might be a high correlation between the recorded measures of the survey.
If this dependence occurs between two samples but the observations within samples are independent, there exist appropriate tests such as the paired t-test (Subsection 5.2.4) or its non parametric alternatives (Subsection 6.3 & 6.4). Regarding more than two dependent samples, we have the within-subject anova or its non parametric alternative, the Friedman test, but these tests are out of the scope of this draft.
Lastly, if the dependence occurs inside the sample (between observations), the tests presented in this draft cannot be applied. Some examples of those kind of samples are time-series where the observation at time $t = 1$ depends on the observation at $t = 0$, or spatial-series where the observation at position $x = 1$ depends on the observation at $x = 0$. The later situations are not covered in this paper.

4.2 Normality

In this section we will see how to check the normality of our observations (or our residuals in the anova context). The normality check is an important and redundant condition for a *parametric* test to be valid. Before discussing the techniques to verify if we are in the normality condition, we introduce the mathematics behind this condition.
For a sample with independent observations $x_1, ..., x_n$, the normality condition writes as

$$x_i \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu, \sigma^2), \forall i \in \{1, ..., n\}.$$

In other words, our observations follow the same law, i.e., they are identically distributed, and this law is the normal one of mean μ and variance σ^2. In this situation, we have the empirical/estimated mean,

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &\overset{\mathcal{L}}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n}) \\ \Leftrightarrow \frac{\sqrt{n}(\bar{x} - \mu)}{\sqrt{\sigma^2}} &\overset{\mathcal{L}}{\sim} \mathcal{N}(0, 1). \end{aligned}$$

Regarding the empirical variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

we have the following result

$$\frac{n\hat{\sigma}^2}{\sigma^2} \overset{\mathcal{L}}{\sim} \chi^2(n-1).$$

The χ^2 law (Figure 3) is characterized by $n - 1$ degrees of freedom (Definition B.2).

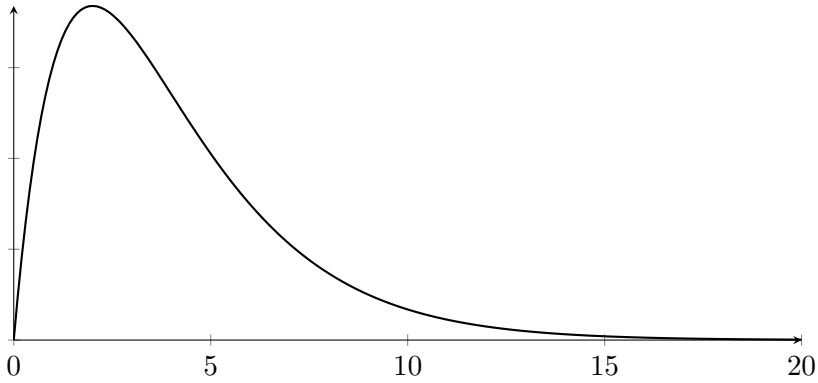


Figure 3: χ^2 density with 4 degrees of freedom.

Lastly, we point out the independence between $\bar{x}$ and $\hat{\sigma}^2$. Based on those theoretical results, we can construct the following test statistic and deduce its law

$$\frac{\sqrt{n}(\bar{x}-\mu)/\sqrt{\sigma^2}}{\sqrt{n\hat{\sigma}^2/\sigma^2(n-1)}} \underset{\sim}{\mathcal{L}} \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/n-1}}$$

$$\Leftrightarrow \frac{\sqrt{n-1}(\bar{x}-\mu)}{\sqrt{\hat{\sigma}^2}} \underset{\sim}{\mathcal{L}} \mathcal{T}_{n-1}.$$

The final equality comes from the fact that the fraction of the standard normal law and the χ^2 law divided by its degrees of freedom gives the Student law (Figure 4) with the χ^2 degrees of freedom.

Remark 4.1. (Convergence of the Student law) It is useful to note that the Student law converges to a standard normal law for n large

$$\mathcal{T}_{n-1} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{1}$$

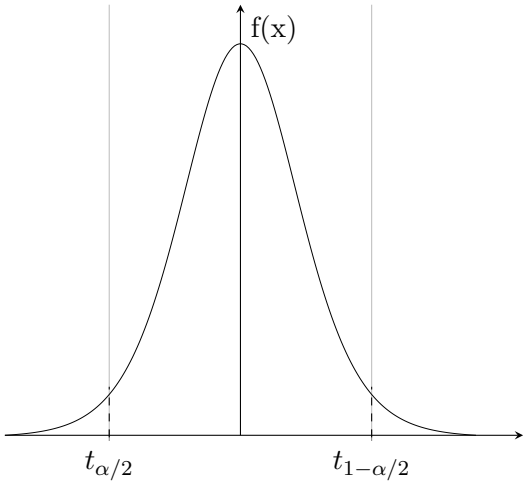


Figure 4: Student density with 10 degrees of freedom.

Let us see now the Central Limit Theorem (CLT).

Theorem 4.1. *(Central Limit Theorem) For n sufficiently large (Remark 4.2), the observations $x_1, ..., x_n$ are independent and identically distributed with respect to a certain law. Assuming that this law admits finite moment of order 2, i.e., $\mathbb{E}\left[x^2\right] < \infty$, we have*

$$\sqrt{n}(\bar{x}-\mu) \underset{\sim}{\mathcal{L}} \mathcal{N}(0,\sigma^2),$$

with $\mu = \mathbb{E}[x_i]$ *and* $\sigma^2 = Var[x_i] \forall i \in \{1, ..., n\}$.

Lastly, for n large, the empirical estimator $\hat{\sigma}^2$ is a good estimator (unbiased) of the true variance σ^2, hence

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sqrt{\hat{\sigma}^2}} \overset{\mathcal{L}}{\sim} \mathcal{N}(0, 1).$$

This statement explains why, with n large, we can bypass the normality condition and still apply a parametric test based on means such as the Student t-test or the anova.

Remark 4.2. (Sufficiently large n) In theory, there is some consensus that n is considered as sufficiently large when $n \geq 30$.

4.2.1 Visual Check

To check the condition of normality, we can do it visually using histograms or QQ-plots, or we can do it formally with a test, notably, the Shapiro-Wilk test. The following histogram (Figure 5) displays a nice bell shape (Gauss density) for a high number of observations (1000 in our case). Note that, first, with less observations, we would probably not have this bell shape even though the observations follow a normal distribution. A second point is the arbitrary choice of the number of bins in the construction of a histogram. Indeed Figure 6 displays the wiggly behavior of the histogram when using lots of bins.

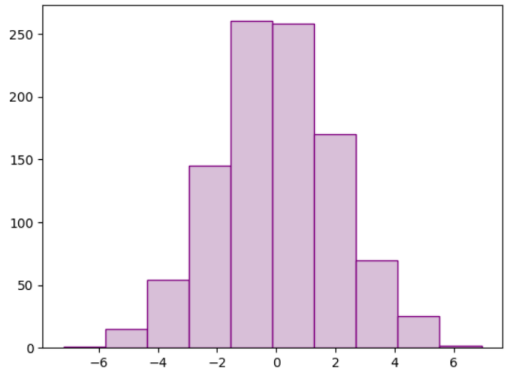


Figure 5: Histogram of 1000 observations drawn from a normal distribution with $\mu = 0$ and $\sigma^2 = 4$. Number of bins: 10.

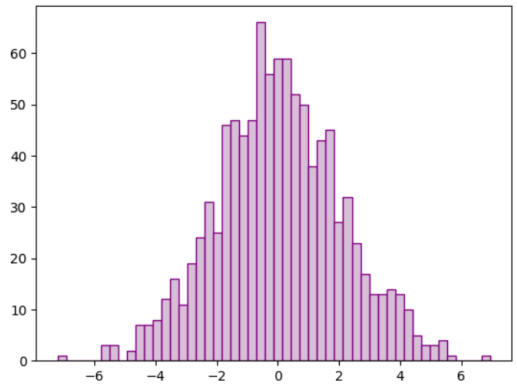


Figure 6: Same observations with a higher number of bins: 50.

When using a QQ-plot (Quantile-Quantile plot), we are comparing the distribution of the quantiles of our observations (y-axis) against the quantiles of the standard normal law (x-axis) (Figure 7). If the point cloud follows the diagonal line (black line), then we can conclude the normality of our observations. As for the histogram, it can be quite subjective. Therefore, we will introduce the normality test of Shapiro-Wilk.

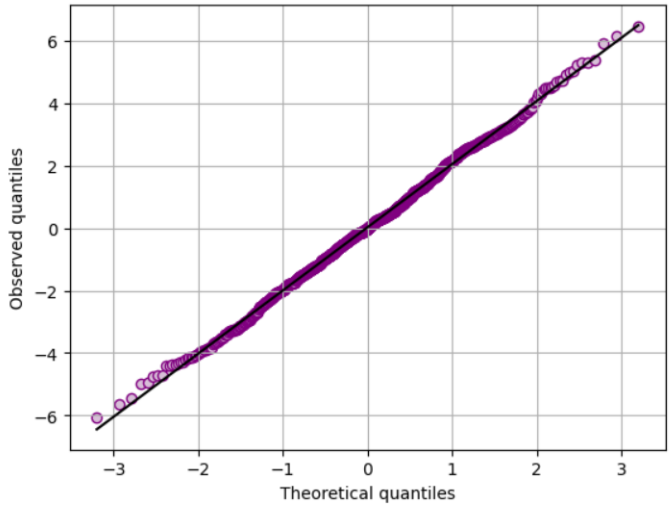


Figure 7: QQ-plot of 1000 observations drawn from a normal distribution with $\mu = 0$ and $\sigma^2 = 4$.

Python code:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
#generation of data
data = np.random.normal(loc=0,scale=2,size=1000)
#classic histogram
plt.hist(data,bins=12)
plt.show()
#classic QQ plot
stats.probplot(data, dist="norm", plot=plt)
plt.grid(True)
plt.show()
```

4.2.2 Formal Check

The Shapiro-Wilk test is the most used test to check statistically the normality of a distribution. The hypothesis testing is formulated as

$$\begin{cases} \mathrm{H}_0 : \text{the distribution of the observations is normal,} \\ \mathrm{H}_1 : \text{the distribution of the observations is not normal.} \end{cases}$$

Hence, if the p-value is higher than the chosen α level, the distribution is considered as normal while if the p-value is less than α, we conclude that the observations are **not** drawn from a normal distribution.

Remark 4.3. This test is disputed for several reasons: for large n, the test will be too sensitive and reject the null hypothesis of normality even though the sample is indeed from a normal distribution. Thanks to the CLT (Theorem 4.1), we can bypass this test of normality for n large. For n small, it can be insensitive to non-normal behavior and therefore conclude to the non-reject of the null hypothesis even though the sample is not normal.

Python code:

```
from scipy.stats import shapiro
stat, p_value = shapiro(data)
```

4.2.3 Conclusion

The normality check must include both techniques: visual (histogram or QQ-plot) and formal (Shapiro-Wilk test). For $n \geq 30$, we can use Theorem 4.1 to state the normality of our empirical estimator $\bar{x}$. For $n \leq 30$, if both techniques lead to the conclusion of non normality of our observations, we will then use non parametric tests that do not require this assumption.

4.3 Homogeneity of Variances

When we are working with ≥ 2 groups of observations, we need the homogeneity between the different variances to be able to apply the Student test or the anova. Those tests compare the sample distributions according to their means.

4.3.1 Student Test

Let us explain why the variances must be homogeneous to detect a difference between those means in the context of two independent samples. We have both samples, $X_1 \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$. If μ_1 and μ_2 are significantly different from each other but so are the variances, we are comparing different distributions according to their means but also their variances. Looking at the test statistic (Subsection 5.2.2), we have the ratio between the mean difference and the sum of variances

$$t_{emp} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n}}}.$$

Therefore, if both variances are very different, assuming $\hat{\sigma}_1^2 << \hat{\sigma}_2^2$ (Figure 8), the denominator will explode and the test statistic t_{emp} will be near zero, so the mean difference will not be detected by the classic test.

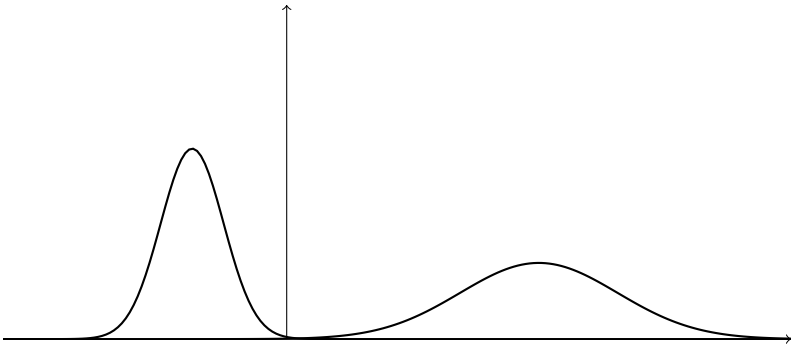


Figure 8: Two normal densities with different means and different variances: $\sigma_1^2 = 1$ (left) and $\sigma_2^2 = 6.25$ (right) (heterogeneity of variances).

The test used to check the homogeneity between variances is called the Levene's test[2]. The hypothesis testing of this test is

$$\begin{cases} \mathrm{H}_0 : \text{the variances are homogeneous,} \\ \mathrm{H}_1 : \text{the variances are not homogeneous.} \end{cases}$$

Hence, a significant p-value (less than α), implies the rejection of H_0 and so the non homogeneity of variances.

Python code:

[2] https://en.wikipedia.org/wiki/Levene%27s_test

```
import numpy as np
from scipy import stats
stat, p = stats.levene(group_1, group_2, center='mean')
```

4.3.2 One-way ANOVA

In the anova context, the test statistic is the ratio between the inter-group variances and the intra-group variances (Subsection 5.3)

$$\frac{\text{Var}_{\text{inter group}}}{\text{Var}_{\text{intra group}}}.$$

Indeed, if we have three samples/groups, we have the variance inside each sample (intra-group) and the variance between the three groups (inter-group). The test statistic is high when the numerator, inter-group variance, is high and when the denominator, intra-group variance, is low. Therefore, if the variance inside each group is high (the denominator), the statistic might be too small for the test to detect the mean difference between groups. This explains why we want an homogeneity between the three variances.
The null hypothesis of the Levene's test states the equality of variances, hence, for a significant p-value (less than α), we reject H_0 which implies the non homogeneity of variances.

Python code:

```
import numpy as np
from scipy import stats
stat, p = stats.levene(group_1, group_2, group_3, center='mean')
```

5 Parametric Tests

In this section we will see the motivation to construct a Confidence Interval (CI) as well as the most used parametric tests: the Student test, also called t-test, and the ANOVA (ANalysis Of VAriance). The appellation *parametric* refers to the use of parameters such as the mean μ and the variance σ^2. Those parameters are called population parameters or true parameters (Definition C.1).
When we apply a parametric test, we typically assume the normality behavior on the distribution of the observations. If those assumptions are not met, the statistical test is not valid. For instance, the p-value is not reliable.

5.1 Confidence Interval

In the case of independent and normally distributed observations, we are able to build a CI around a chosen parameter. The later interval provides bounds within which the true parameter of interest is located. We can typically construct a CI around the mean population or the variance population. We focus here on the mean parameter, hence, the CI answers the question *Is my sample drawn from a normal distribution of mean μ_0?*. The hypothesis testing is then

$$\begin{cases} \text{H}_0 : \mu = \mu_0 \\ \text{H}_1 : \mu \neq \mu_0, \end{cases}$$

with μ the true mean that our sample estimates with $\bar{x}$ and μ_0 the hypothesized true mean.
Let us discuss the construction of such interval. The CI is based on the **independence** and **normality** conditions. In the following we distinguish between *exact* CI, that is, for small n, and *asymptotic* CI, for large n.

5.1.1 Exact Confidence Interval

For n small, we use the quantiles of the Student law to construct the interval. To do that, recall the theoretical results from Subsection 4.2 that states

$$\frac{\sqrt{n-1}(\bar{x}-\mu_0)}{\sqrt{\hat{\sigma}^2}} \overset{\mathcal{L}}{\sim} \mathcal{T}_{n-1}.$$

This result implies that the later statistic has a probability of $1-\alpha$, if H_0 is true, to be included within the two quantiles $t_{\alpha/2,n-1}$ and $t_{1-\alpha/2,n-1}$ of the Student law (Figure 4). We then write it mathematically as

$$\mathbb{P}_{H_0}\left[t_{\alpha/2,n-1} \leq \frac{\sqrt{n-1}(\bar{x}-\mu_0)}{\sqrt{\hat{\sigma}^2}} \leq t_{1-\alpha/2,n-1}\right] = 1-\alpha.$$

Therefore, the interval writes as

$$\left[\bar{x} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n-1}} \leq \mu_0 \leq \bar{x} + t_{1-\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n-1}}\right].$$

In practice, if μ_0 is included in the interval, we do not reject the null hypothesis of $\mu = \mu_0$ at the α level. If μ_0 is not included in the CI, we reject the null hypothesis at the α level.

Remark 5.1. (Interpretation of the CI) The CI at level $1-\alpha$ with $\alpha = 0.05$ is the probability that if we run the experiment 100 times, 95 intervals will contain the true parameter μ_0. Note that the probability is not stated on the parameter μ_0 given that it is a fixed number, it has no probability.

Python code:

```
from scipy.special import stdtr, stdtrit
import math
import numpy as np
mean_sample = np.mean(data)
var_sample =np.var(data)
t_alpha_2 = stdtrit(len(data)-1, 0.05/2)
t_1_alpha_2 = stdtrit(len(data)-1, 1-0.05/2)
lower_ic = mean_sample + t_alpha_2*(math.sqrt(var_sample)/math.sqrt(len(data)-1))
upper_ic = mean_sample + t_1_alpha_2*(math.sqrt(var_sample)/math.sqrt(len(data)-1))
```

5.1.2 Asymptotic Confidence Interval

When n is large and the observations are still independent, we do not need to check the normality condition (Theorem 4.1), we simply use the quantiles of the normal distribution ($z_{\alpha/2}$ and $z_{1-\alpha/2}$). We then obtain the following asymptotic interval

$$\mathbb{P}_{H_0}\left[\bar{x} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right] \to 1-\alpha.$$

As for the exact CI, if μ_0 is included in the interval, it means that $\mu = \mu_0$ at the α level. If this is not the case then we reject the null hypothesis.

Python code:

```
from scipy.stats import norm
import math
import numpy as np
mean_sample = np.mean(data)
```

```
var_sample =np.var(data)
alpha_2 = 0.05/2
z_alpha_2 = norm.ppf(alpha_2)
z_1_alpha_2 = norm.ppf(1-alpha_2)
lower_ic = mean_sample + z_alpha_2*(math.sqrt(var_sample)/math.sqrt(len(data)))
upper_ic = mean_sample + z_1_alpha_2*(math.sqrt(var_sample)/math.sqrt(len(data)))
```

5.2 Student Test

The test of Student[3], or t-test, answers the question *Are the two samples from the same population?* through the analysis of the means. We can either compare an empirical mean (from our observations) to a theoretical mean μ_0, this is the one-sample t-test. Or we can compare two empirical means, this refers to the two-sample t-test. The later test can be done on independent as well as dependent groups. Let us start with the one-sample t-test.

5.2.1 One-sample t-test

Once the **independence** and **normality** of the observations are satisfied, we formulate the bilateral hypothesis testing as

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0. \end{cases}$$

The mean μ in this hypothesis testing is estimated/inferred with $\bar{x}$. We then compute the following test statistic t_{emp} (empirical/observed t):

$$t_{emp} = \frac{\sqrt{n-1}(\bar{x} - \mu_0)}{\sqrt{\hat{\sigma}^2}}.$$

Under the normality condition and H_0, we saw (Subsection 4.2) that t_{emp} is drawn from a Student law of $n-1$ degrees of freedom. The p-value is then formulated as

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{\alpha/2,n-1} \leq t_{emp} \leq t_{1-\alpha/2,n-1}\right]. \tag{2}$$

Two remarks can be made regarding this probability.

Remark 5.2. (Under the null hypothesis) The above probability $\mathbb{P}_{\mathrm{H}_0}$ is valid only under the null hypothesis. That is, assuming that H_0 is true, we have our population parameter $\mu = \mu_0$, therefore our mean estimator $\bar{x}$ is drawn from $\mathcal{N}(\mu_0, \sigma^2/n)$ (using also the normality condition). Then, if H_0 is true, t_{emp} is a random variable following the Student law. If this probability is very rare, less than α, we can then reject H_0.

Remark 5.3. (Interpretation) The bilateral p-value is the probability, under the null hypothesis H_0, to observe that our statistic t_{emp} is included in the interval formed by the two quantiles of the Student law of $n-1$ degrees of freedom at a chosen α level. Here we can state a probability on our statistic given that it is a random variable (it varies according to the observations).

Remark 5.4. (How to find the p-value) The p-value can be found looking at specific tables built on the running of lots of experiments. Those experiments allow to assign probabilities to empirical statistics (e.g., t_{emp}) with respect to the specified law (here it is the Student law for instance) and the null hypothesis of mean equality.

We reject the null hypothesis if this probability (eq. 2) is very small, less than a small chosen α level. In other words, the probability to obtain a test statistic whose value is included between the two quantiles is very rare. We then conclude with the reject of the null hypothesis, implying that $\mu \neq\neq \mu_0$.

[3] https://en.wikipedia.org/wiki/Student%27s_t-test

Remark 5.5. (Non reject of H_0) If the p-value is greater than α, we do not reject the null hypothesis but we never accept it, we only conclude that we do not have enough proofs to reject H_0.

If we have prior information on the direction of the difference between μ and μ_0, through descriptive statistics for instance (Section 3), we can formulate a unilateral hypothesis. If our empirical mean $\bar{x}$ is higher than the theoretical μ_0, we formulate the null hypothesis H_0 as what we want to reject, that is

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0. \end{cases}$$

The probability to test the null hypothesis is then

$$\mathbb{P}_{H_0}\left[t_{emp} \leq t_{1-\alpha,n-1}\right].$$

If this probability (p-value) is very small, smaller than the α level, we reject H_0.
Lastly, if prior information shows that $\bar{x} < \mu_0$, we formulate H_0 as what we want to reject, that is

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0. \end{cases}$$

Our probability of interest will then be

$$\mathbb{P}_{H_0}\left[t_{emp} \geq t_{\alpha,n-1}\right].$$

Same reasoning as before, if this p-value is smaller than α, we reject the null hypothesis.

Python code:

```
import scipy.stats as stats
mu_0 = 0
t_stat, p_value = stats.ttest_1samp(data, popmean=mu_0, alternative="two-sided")
t_stat, p_value = stats.ttest_1samp(data, popmean=mu_0, alternative="greater")
t_stat, p_value = stats.ttest_1samp(data, popmean=mu_0, alternative="less"')
```

5.2.2 Independent Two-sample t-test

When we want to apply a t-test on two **independent** groups, we need the **normality** and the **homogeneity of variances** conditions. If the later condition is not met, we can use the Welch's t-test that applies a correction on the degrees of freedom of the Student law for heterogeneous variances (Subsection 5.2.3). Once all three conditions are satisfied, we can pursue with the t-test.
The bilateral hypothesis testing is formulated as follows

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \mu_0 \\ H_1 : \mu_1 - \mu_2 \neq \mu_0, \end{cases}$$

with μ_0 a theoretical mean, often taken as 0 to immediately compare the difference between μ_1 and μ_2. Note that the provided python code at the end of this subsection is based on $\mu_0 = 0$.
Given that we have two samples, they might or might not have the same sample size. Let us look at the first situation, that is, $n_1 = n_2$.

Equal sample sizes

For $n_1 = n_2 = n$, we use the following test statistic under H_0

$$t_{emp} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n}}}.$$

Under the conditions of application and the null hypothesis, this random variable t_{emp} follows the Student law of degree of freedom $2n - 2 = 2(n - 1)$. This is explained by the following

$$\bar{x}_1 \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2/n)$$
$$\bar{x}_2 \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2/n).$$

Then, we have

$$\bar{x}_1 - \bar{x}_2 \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2 + \sigma_2^2}{n}\right).$$

Under the null hypothesis, $\mu_1 - \mu_2 = \mu_0$. Hence, we have

$$(\bar{x}_1 - \bar{x}_2) - \mu_0 \overset{\mathcal{L}}{\sim} \mathcal{N}\left(0, \frac{\sigma_1^2 + \sigma_2^2}{n}\right).$$

The Student law is then deduced as explained in Subsection 4.2.

Unequal sample sizes

When the sample sizes are not equal, $n_1 \neq n_2$, we correct the statistic t_{emp} using a pooled standard deviation

$$t_{emp,pooled} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\hat{\sigma}^2_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\sqrt{\hat{\sigma}^2_{pooled}} = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}.$$

The later test statistic $t_{emp,pooled}$ follows (under H_0 and the conditions of application) a Student law of $n_1 + n_2 - 2$ degrees of freedom. We can see that the difference between the two test statistics is the degrees of freedom of the Student law.
Now that we defined the test statistic for each situation, we can formulate the bilateral p-value for **equal sample sizes** as

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp} \in [t_{\alpha/2,2(n-1)}, t_{1-\alpha/2,2(n-1)}]\right],$$

and for **unequal sample sizes** as

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,pooled} \in [t_{\alpha/2,n_1+n_2-2}, t_{1-\alpha/2,n_1+n_2-2}]\right].$$

If the probability is very low, less than a small α, we reject the null hypothesis of mean equality and therefore, we consider both samples as drawn from different distributions according to the position measure, that is, the mean. If we have prior information (e.g., looking at the boxplot (Subsection 3.2) or simply looking at both means), we can formulate unilateral hypothesis. For a prior information stating that $\bar{x}_1 - \bar{x}_2 > \mu_0$, we formulate the null hypothesis as the opposite

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 \leq \mu_0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 > \mu_0. \end{cases}$$

Then, the p-value for **equal sample sizes** is

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp} \leq t_{1-\alpha,2(n-1)}\right],$$

and the p-value for **unequal sample sizes** is

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,pooled} \leq t_{1-\alpha,n_1+n_2-2}\right].$$

Lastly, if we observe that $\bar{x}_1 - \bar{x}_2 < \mu_0$, we have the following hypothesis testing

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 \geq \mu_0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 < \mu_0. \end{cases}$$

The p-value for **equal sample sizes** is

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp} \geq t_{\alpha,2(n-1)}\right],$$

and the p-value for **unequal sample sizes** is

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,pooled} \geq t_{\alpha,n_1+n_2-2}\right].$$

Python code:

```
import numpy as np
from scipy import stats
#for mu_0 = 0
t,p =stats.ttest_ind(group_1,group_2, equal_var=True, alternative="two-sided")
t,p =stats.ttest_ind(group_1,group_2, equal_var=True, alternative="greater")
t,p =stats.ttest_ind(group_1,group_2, equal_var=True, alternative="less")
```

5.2.3 Welch's t-test

When the homogeneity of variances condition is not satisfied, we can apply the Welch's t-test that suggests the following test statistic for equal and unequal sample sizes under the **independence** and **normality** conditions

$$t_{emp,\mathrm{Welch}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}.$$

The main difference with the test statistic of the Student test lies in the degrees of freedom of the Student law. Those degrees of freedom ν are computed via the Welch–Satterthwaite equation[4]

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)}}.$$

Hence, the Welch's test statistic follows approximately a Student law of degrees of freedom ν. Hence, The bilateral p-value writes as

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,\mathrm{Welch}} \in \left[t_{\alpha/2,\nu}, t_{1-\alpha/2,\nu}\right]\right].$$

We can define the unilateral p-values using the same logic as the Student test above taking into account the different degrees of freedom ν.

Python code:

```
import numpy as np
from scipy import stats
#for mu_0 = 0
t,p =stats.ttest_ind(group_1,group_2, equal_var=False, alternative = "two-sided")
t,p =stats.ttest_ind(group_1,group_2, equal_var=False, alternative = "greater")
t,p =stats.ttest_ind(group_1,group_2, equal_var=False, alternative = "less")
```

[4] https://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation

5.2.4 Paired t-test

The dependent two-samples t-test, or paired t-test, is of interest when we have **dependence** between samples, for instance, the participants of group 1 are men and the participants of group 2 are their wives. In this case, there might be a high correlation between the recorded measures (e.g., answers to survey). To take this correlation into account, we compute the difference of the observations between both groups using the following quantity D_i

$$D_i = x_{i,1} - x_{i,2},$$

with $x_{i,1}$ the ith observation in group 1 and $x_{i,2}$ the ith observation in group 2. From this quantity, we can define the average difference $\bar{D}$

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,1} - x_{i,2}),$$

with $n = n_1 = n_2$. We define the associated (unbiased) variance $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2.$$

The differences must be **independent** and **normally** distributed, $D_i \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_0, \sigma^2)$. We formulate the bilateral hypothesis as

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 = \mu_0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 \neq \mu_0, \end{cases}$$

and the test statistic as

$$t_{emp,paired} = \frac{\bar{D} - \mu_0}{\sqrt{\hat{\sigma}^2}}.$$

Note that μ_0 is often taken as 0 to immediately compare both samples. Under the normality and independence conditions on D and under the null hypothesis, this statistic is drawn from a Student law with $n-1$ degrees of freedom, therefore we formulate the bilateral p-value as

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,paired} \in [t_{\alpha/2,n-1}, t_{1-\alpha/2,n-1}]\right].$$

If we have prior information on the direction of the mean inequality (Subsection 3.2), we use unilateral hypothesis. For $\bar{x}_1 - \bar{x}_2 > \mu_0$, the unilateral hypothesis is formulated as

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 \leq \mu_0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 > \mu_0, \end{cases}$$

with the associated p-value

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,paired} \leq t_{1-\alpha,n-1}\right].$$

For prior information showing $\bar{x}_1 - \bar{x}_2 < \mu_0$, the unilateral hypothesis writes as

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 \geq \mu_0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 < \mu_0, \end{cases}$$

with the following p-value

$$\mathbb{P}_{\mathrm{H}_0}\left[t_{emp,paired} \geq t_{\alpha,n-1}]\right].$$

Python code:

For $\mu_0 = 0$:

```
from scipy import stats
t_stat, p_value = stats.ttest_rel(group_1, group_2, alternative='two-sided')
t_stat, p_value = stats.ttest_rel(group_1, group_2, alternative='greater')
t_stat, p_value = stats.ttest_rel(group_1, group_2, alternative='less')
```

5.3 One-way ANOVA

We will now consider the ANOVA[5] that refers to the analysis of variances of at least 2 groups, but more frequently, at least 3 given that the Student test is used for 2 groups. This analysis answers the question *Are my samples from the same population?* according to their means. The design of the experiment that fits this analysis is the following. We have **independent** samples and we have one independent variable (x_1) that is a categorical variable.

Definition 5.1. (Categorical variable) A variable is categorical when it admits nominal levels/responses, e.g., the eyes color is a categorical variable with the levels/possible responses: brown, blue, green.

For instance, we have 20 participants and we measure their sizes (e.g., in centimeter, this is our y's numerical observations) and note their eyes color (that is our categorical variable x_1). We can then separate our 20 participants in groups of brown, blue and green eyes color. We then estimate the mean and variance of those three groups and use them to do a statistical comparison of the group distributions.
The bilateral hypothesis testing is formulated as follow

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 = ... = \mu_0 \\ \mathrm{H}_1 : \mu_i \neq \mu_j \text{ for } i, j \in \{1, ..., l\} \end{cases}$$

The null hypothesis states that all the means from 1 to l are equal to μ_0, meaning that there is not effect of the factor level. The alternative hypothesis, H_1, means that there is at least two means that are different from each other among all the means, that is, at least one level of the factor has a different impact on the observations compared to the other levels.
Mathematically speaking, our observations y_{ij} are indexed by $i \in \{1, ..., l\}$ that refers to the level of the categorical variable x_1 and by $j \in \{1, ..., n\}$ that refers to the participant ID. The equation of the one-way ANOVA under the null hypothesis is

$$y_{ij} = \mu_0 + \epsilon_{ij}$$

given that $\mu_i = \mu_0$, $\forall i \in \{1, ..., l\}$. In practice, this equation becomes

$$y_{ij} = \bar{y}_{..} + e_{ij},$$

with $\bar{y}_{..}$ the overall mean (overall participants regardless of the level i). The equation representing the alternative hypothesis, i.e., $\mu_i \neq \mu_0$, is

$$\begin{aligned} y_{ij} &= \mu_0 + \alpha_i + \epsilon_{ij} \\ &= \mu_i + \epsilon_{ij} \end{aligned}$$

with μ_0 the theoretical mean, α_i the effect of level i and ϵ_{ij} the error of participant j in the group i. Here, $\mu_i = \mu_0 + \alpha_i$. The ϵ_{ij} must be **independent and identically distributed** (i.i.d) according to the **normal** law

$$\begin{aligned} \epsilon_{ij} &= y_{ij} - \mu_i \\ &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2). \end{aligned}$$

[5] https://en.wikipedia.org/wiki/Analysis_of_variance

The explanation to test the normality condition on the residuals and not on the observations is the following. Each group i of observations has its own mean μ_i that can be equal to μ_0 (under H_0) or significantly different (under H_1) (that question is answered by the test), we then have

$$y_{ij} \overset{\mathcal{L}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2).$$

Hence, even though the law of each observations is normal, the different means will lead to a mixture of normal distributions when we test the normality hypothesis on all the observations. A mixture of normal distributions is not necessarily normal, if the means are different, we would have a multimodal distribution while a normal distribution is unimodal by definition. Therefore, by subtracting each mean μ_i to the observations y_{ij}, we center the distribution around zero, that gives us the residuals ϵ_{ij}. In practice, the equation representing the alternative hypothesis is

$$y_{ij} = \bar{y}_{i.} + e_{ij},$$

with $\bar{y}_{i.}$ the mean specific to each level i and e_{ij} the residuals defined as

$$\begin{aligned} e_{ij} &= y_{ij} - \bar{y}_{i.} \\ &\overset{\mathcal{L}}{\sim} \mathcal{N}\left(0, \sigma^2\left(1 - \frac{1}{n_i}\right)\right), \end{aligned}$$

with n_i the sample size of level i. To test the **normality of the residuals**, we use the Shapiro-Wilk test and the visual check (Subsection 4.2)

Python code:

```
#computation of residuals
r_1 = group_1 - np.mean(group_1)
r_2 = group_2 - np.mean(group_2)
r_3 = group_3 - np.mean(group_3)
residuals = np.hstack((r_1,r_2,r_3))
#Apply Shapiro test and QQ-plot visualization on residuals
```

As explained in Subsection 4.3, the model also requires the **homogeneity of variances**. if the Levene's test is significant, we cannot pursue with the classic anova. Instead, we use the Welch's anova (Subsection 5.4). Let us continue with the classic anova assuming variance homogeneity. The test statistic used to compare the impact of the different levels i of x_1 is computed using the Sum of Squares (SS) method. The Sum of the Squares Total (SST) is defined as follows

$$\text{SST} = \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2.$$

This quantity can be further decomposed as

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}\left((y_{ij} - \bar{y}_{i.})^2 + (\bar{y}_{i.} - \bar{y}_{..})^2\right) + 2\sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}). \end{aligned}$$

Here we simply add and subtract the same quantity $\bar{y}_{i.}$. Noting that the last product term is zero

$$\begin{aligned} 2\sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..}) &= 2\sum_{i=1}^{l}(\bar{y}_{i.}-\bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.}) \\ &= 2(l\bar{y}_{..}-l\bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.}) \\ &= 0. \end{aligned}$$

Here we just put the term $(\bar{y}_{i.}-\bar{y}_{..})$ outside of the sum over j given that it does not depend on it. Noting that $n\bar{x}=\sum_{k=1}^{n}x_k$, we apply this identity on the later term and we obtain the result. Therefore, we have

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 \\ &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}\left((y_{ij}-\bar{y}_{i.})^2+(\bar{y}_{i.}-\bar{y}_{..})^2\right) \\ &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2+\sum_{i=1}^{l}n_i(\bar{y}_{i.}-\bar{y}_{..})^2 \\ &= \text{SS}_{\text{intra group}}+\text{SS}_{\text{inter group}} \\ &= \text{SSE}+\text{SS}_{\text{factor}}. \end{aligned}$$

The *intra group* term refers to the sum of the squared error inside each group/level, also called the SS of the errors (SSE). The *inter group* term refers to the sum of the squared between the groups/levels, also called the SS explained by the factor/the independent categorical variable (x_1).
Under the null hypothesis, we have that $\bar{y}_{i.}\sim\mathcal{N}(\bar{y}_{..},\sigma^2/n_i)$ for each level $i=\{1,...,l\}$ (explained in Subsection 4.2), therefore

$$\frac{\bar{y}_{i.}-\bar{y}_{..}}{\sqrt{\sigma^2/n_i}}\overset{\mathcal{L}}{\sim}\mathcal{N}(0,1),$$

hence, squaring the later expression,

$$\frac{n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sigma^2}\overset{\mathcal{L}}{\sim}\chi^2(1).$$

We can define the following random variables so that they follow a χ^2 law (Figure 3):

$$\frac{\text{SS}_{\text{factor}}}{\sigma^2}\overset{\mathcal{L}}{\sim}\chi^2(l-1),$$

the SS of the factor is endowed with $l-1$ degrees of freedom, i.e., the number of levels minus one, and the random variable

$$\frac{\text{SSE}}{\sigma^2}\overset{\mathcal{L}}{\sim}\chi^2(n-l),$$

with $n-l$ degrees of freedom, i.e., the number of observations/participants minus the number of levels of x_1.
We are now able to define the test statistic as

$$\begin{aligned} \frac{\frac{\text{SS}_{\text{factor}}}{\sigma^2(l-1)}}{\frac{\text{SSE}}{\sigma^2(n-l)}} &\overset{\mathcal{L}}{\sim}\frac{\chi^2(l-1)/(l-1)}{\chi^2(n-l)/(n-l)} \\ &\overset{\mathcal{L}}{\sim}F(l-1,n-l) \end{aligned} \tag{3}$$

with $\mathcal{F}(l-1,n-1)$ the Fisher law (Figure 9) with the two specified degrees of freedom.

Remark 5.6. (Sum of squares and Fisher law) The test statistic (eq.3) is the ratio of the squared difference between each level mean $\bar{y}_{i.}$ and the overall mean $\bar{y}_{..}$ ($\text{SS}_{\text{factor}}$) and the squared error (SSE). This statistic is small when the numerator is small, that is, $\bar{y}_{i.} \approx \bar{y}_{..} \forall i$. On the contrary, when $\bar{y}_{i.} \neq \bar{y}_{..}$ for at least one level i, the statistic becomes greater although we do not know the direction of this difference ($<, >$) because of the square. Hence, the only hypothesis we can check is the bilateral one. This is in line with the Fisher law. Indeed, the Fisher density (Figure 9) is a right tailed density, therefore the only quantile of interest is the right quantile, i.e., $1-\alpha$. This explains the p-value below and the non unilateral hypothesis.

We can also write the test statistic using the Mean Sum of Squared (MSE) which is simply the SS divided by their respective degrees of freedom

$$\text{MSE}_{\text{factor}} = \frac{\text{SS}_{\text{factor}}}{l-1},$$

$$\text{MSE}_{\text{error}} = \frac{\text{SSE}}{n-l},$$

so that we have

$$\frac{\text{MSE}_{\text{factor}}}{\text{MSE}_{\text{error}}} \sim \mathcal{F}(l-1, n-l).$$

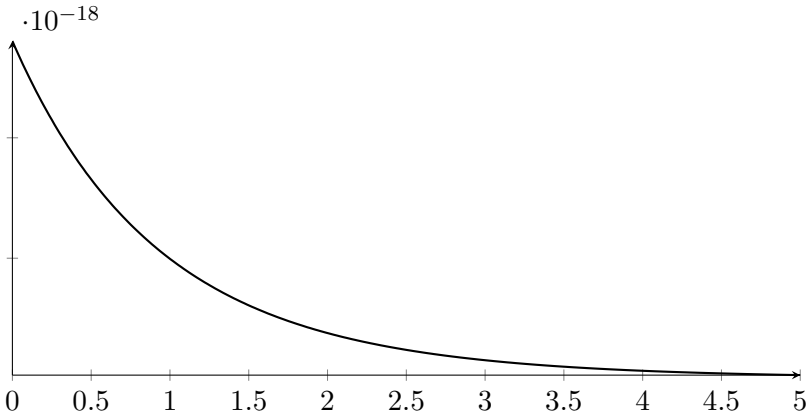


Figure 9: Fisher(-Snedecor) density with numerator degrees of freedom $d_1 = 2$ and denominator degrees of freedom $d_2 = 27$.

Now that we have our test statistic, we formulate the p-value under the bilateral null hypothesis as

$$\mathbb{P}_{\text{H}_0}\left[\frac{\text{MSE}_{\text{factor}}}{\text{MSE}_{\text{error}}} \leq \mathcal{F}_{1-\alpha}(l-1, n-l)]\right].$$

Once again, if this probability is very rare, less than a small α level, we reject H_0. We conclude then that our test the means $\bar{y}_{i.}$ are not all equal to each other. Therefore, the groups are not from the same population according to the parametric position measure.

Python code:

```
import numpy as np
from scipy.stats import f_oneway
f_oneway(group_1,group_2,group_3)
```

If the anova is significant, the next step is to investigate where the mean inequality comes from, we can do that via pairwise comparisons.

5.3.1 Multiple comparisons test

The most popular post-hoc test, i.e., after the anova test, is the Tukey test, also called Tukey HSD (Honestly significant difference). The hypothesis formulation for the Tukey test is as follows

$$\begin{cases} \mathrm{H}_0 : \mu_i = \mu_j \\ \mathrm{H}_1 : \mu_i \neq \mu_j \text{ for } i, j \in \{1, ..., l\}. \end{cases}$$

The statistic of this test under the **independence**, **normality** and **homogeneity of variances** conditions (same as the one-way anova conditions) and the null hypothesis is the following

$$\begin{aligned} q_{emp,ij} &= \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{\mathrm{MSE}/n}} \\ &\overset{\mathcal{L}}{\sim} \mathcal{Q}(l, n-l), \end{aligned}$$

with $\mathcal{Q}$ the Studentized range law, l the number of groups/means to compare and $n - l$ the degrees of freedom associated to the MSE.
Now that we know the law of our test statistic, we can construct the p-value as follows

$$\mathbb{P}_{\mathrm{H}_0}\left[q_{emp,ij} \leq \mathcal{Q}_{1-\alpha}(l, n-l)\right],$$

which is the same for any i,j from 1 to l. If the test statistic is higher than the quantile $1 - \alpha$, we reject the null hypothesis at this α level.

Remark 5.7. (Studentized range law) The hypothesis is formulated in a bilateral manner, however, the Studentized range law is a right tailed law, i.e., the only interesting quantile to look at is the right one, that is, the $1 - \alpha$ quantile.

Remark 5.8. (Correction of the α inflation) When applying several statistical tests such as the Tukey tests to compare two-by-two the 3 means, we are increasing the type I error, i.e., the error of rejecting H_0 when we should not (false positive), which is exactly the α level. Therefore, there exist some corrections of this level such as the Bonferroni correction which simply consists of lowering the α level by dividing it by the number of pairwise comparisons.
However, in the context of the Tukey test, the statistic follows a Studentized range law that takes into account the number of pairwise comparisons and already realizes a correction for the false positive problem.

Python code:

```
from scipy.stats import tukey_hsd
tukey_hsd(group1, group2, group3)
```

5.4 Welch's ANOVA

If the homogeneity of variances condition is not satisfied but the **independence** and **normality of residuals** conditions are met, we can apply the Welch's anova. The adjusted F statistic for heterogeneous variances is

$$F_{\mathrm{Welch}} = \frac{\sum_{i=1}^{l} \frac{\hat{\sigma}_i^2}{n_i}(\bar{y}_{i.} - \bar{y}_w)^2/(l-1)}{1 + \frac{2(l-2)}{l^2-1}\sum_{i=1}^{l} \frac{(1-(\frac{\hat{\sigma}_i^2}{n_i})/W)^2}{n_i - 1}},$$

with

$$W = \sum_{i=1}^{l} \frac{\hat{\sigma}_i^2}{n_i}, \quad \bar{y}_w = \frac{\sum_{i=1}^{l} \frac{\hat{\sigma}_i^2}{n_i}\bar{y}_{i.}}{\sum_{i=1}^{l} \frac{\hat{\sigma}_i^2}{n_i}}.$$

This adjusted F_{Welch} statistic is simply the ratio of weighted variances

$$F_{\text{Welch}} = \frac{\text{Weighted Var}_{\text{inter group}}}{\text{Weighted Var}_{\text{intra group}}}.$$

Regarding the law of this test statistic, we compute the degrees of freedom of the weighted test statistic via the Welch–Satterthwaite equation[6]

$$\nu \approx \frac{\left(\sum_{i=1}^{l} \frac{\hat{\sigma}_j^2}{n_j}\right)^2}{\sum_{i=1}^{l} \frac{\left(\frac{\hat{\sigma}_i^2}{n_i}\right)^4}{n_i - 1}}.$$

Hence, F_{Welch} follows a Fisher law of $l-1$ degrees of freedom for the numerator and ν for the denominator. Then, as for the one-way anova, the p-value writes as

$$\mathbb{P}_{\text{H}_0}\left[F_{\text{Welch}} \leq \mathcal{F}_{1-\alpha}(l-1, \nu)\right].$$

Here, once again, the only quantile of interest is the right one, $1-\alpha$, given that the Fisher density is right tailed and the statistic is always positive because of the square. If this probability is very small, less than α, we reject the null hypothesis of mean equality.

Python code[7]*:*

```
import pandas as pd
import pingouin as pg
#example of dataframe input for the test
data = pd.DataFrame({
    'score': [23, 25, 27, 30, 19, 21, 34, 36, 38, 22, 20, 18],
    'groups': ['A']*4 + ['B']*4 + ['C']*4
})
#Welch's anova
pg.welch_anova(dv='score', between='groups', data=data)
```

As for the classic one-way anova, in the situation of significant test statistic, we need to detect where the difference between means occurs, therefore we apply a multiple comparison test adapted to the heterogeneity of variances, that is, the Games-Howell test.

5.4.1 Mutliple comparison test for non homogeneous variances

The Games-Howell test statistic also follows, as the Tukey test, a Studentized range law. The hypothesis is formulated as

$$\begin{cases} \text{H}_0 : \mu_i = \mu_j \\ \text{H}_1 : \mu_i \neq \mu_j \text{ for } i, j \in \{1, ..., l\}. \end{cases}$$

The test statistic is the following

$$q_{GH,ij} = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j}}} \forall i, j \in \{1, ..., l\}.$$

[6] https://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation
[7] https://pingouin-stats.org/build/html/generated/pingouin.welch_anova.html

We compare the amplitude between each mean i and j. The degrees of freedom of this statistic are also computed with the Welch–Satterthwaite equation

$$\nu_{ij} = \frac{\left(\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j}\right)^2}{\frac{\left(\frac{\hat{\sigma}_i^2}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{\hat{\sigma}_j^2}{n_j}\right)^2}{n_j - 1}}.$$

Hence, the Games-Howell test statistic follows a Studentized range law of degrees of freedom ν. Given that the Studentized range law is a right tailed law, the p-value is formulated with respect to the right quantile

$$\mathbb{P}_{\mathrm{H}_0}\left[q_{\mathrm{GH},ij} \leq \mathcal{Q}_{1-\alpha}(l, \nu)\right],$$

with l the number of groups (levels of the categorical variable) and ν the adapted degrees of freedom. If this probability is very small, we reject the null hypothesis of mean equality and conclude that the two tested means i, j are significantly different from each other. To know the direction of this difference ($\mu_i > \mu_j$ or $\mu_i < \mu_j$), we can simply look at the value of those means or look at the boxplots (Subsection 3.2).

Python code[8]*:*

```
import pandas as pd
import pingouin as pg
#example of dataframe input for the test
data = pd.DataFrame({
    'score': [23, 25, 27, 30, 19, 21, 34, 36, 38, 22, 20, 18],
    'groups': ['A']*4 + ['B']*4 + ['C']*4
})
#if significant anova => Games-Howell test
pg.pairwise_gameshowell(data=data, dv='score',between='groups').round(3)
```

6 Non Parametric Tests

Non parametric tests are useful, notably, when the normality condition of the parametric tests is not encountered. Moreover, they are also used when our sample contains outliers (Subsection 3). Indeed, non parametric tests do not assume any distribution on the observations. They are instead based on ranks such as the median when we want to distinguish samples according to the position measure. The ranks are integers from 1 to n and are assigned to the ordered (according to their values) observations. For instance, once the observations are ordered in an increasing way, the median is the value corresponding to the rank $\frac{(n+1)}{2}$.
We will now see the non parametric alternatives of the above parametric tests.

6.1 Sign Test

The sign test is the non parametric alternative of the one-sample t-test (Subsection 5.2.1) under the **independence** condition. As explained in the introduction of this section, we are now interested in the median parameter and not in the mean parameter anymore.

[8] https://pingouin-stats.org/build/html/generated/pingouin.welch_anova.html

For an independent sample $X := (x_1, ..., x_n)$ and its median ξ, we can write

$$\begin{aligned}\mathbb{P}(X < \xi) &= \frac{1}{2} \\ &= \mathbb{P}(X > \xi).\end{aligned}$$

This is the definition of the median. The proportion of observations X less than the median value ξ is equal to $1/2$, i.e., the half of the observations are below the middle value and the other half of the observations are above the median. With that in mind, we can formulate the bilateral null hypothesis (no prior information) using a reference median ξ_0 as follows

$$\begin{cases} \mathrm{H}_0 : \xi = \xi_0 \\ \mathrm{H}_1 : \xi \neq \xi_0. \end{cases}$$

If H_0 is true, we have

$$\mathbb{P}_{\mathrm{H}_0}(X < \xi_0) = \frac{1}{2}.$$

Given that a probability is also a proportion, we denote the later probability as π. Therefore, the hypothesis can be reformulated as

$$\begin{cases} \mathrm{H}_0 : \pi = 1/2 \\ \mathrm{H}_1 : \pi \neq 1/2. \end{cases}$$

The test statistic will be $n\hat{\pi}$, that is, the number of observations below ξ_0 and this corresponds to a binomial process. The binomial process refers to the counting of observations satisfying a specific condition, here the condition is to fall under the threshold ξ_0. Under the null hypothesis, $\xi = \xi_0$, each observation has a probability of $1/2$ to be below ξ_0, therefore the law of this test statistic for n small ($n < 20$), i.e., the exact law, is

$$n\hat{\pi} \overset{\mathcal{L}}{\sim} \mathcal{B}(n, 1/2).$$

As usual, we now construct the p-value using the quantiles of this binomial law

$$\mathbb{P}_{\mathrm{H}_0}\left[b_{\alpha/2} \leq n\hat{\pi} \leq b_{1-\alpha/2}\right].$$

If this probability is less than a small chosen α level, we consider this probability as rare and then reject H_0, meaning that the median of our sample ξ is different from the reference median ξ_0. We apply the same reasoning as in the parametric framework.
As for the CI (Section 5.1) where we have exact and asymptotic CI, we also have a different law for this test statistic $n\hat{\pi}$ for n large (here, $n \geq 20$)

$$\frac{n(\hat{\pi} - 1/2)}{\sqrt{(n/2)^2}} \overset{\mathcal{L}}{\sim} \mathcal{N}(0, 1).$$

Then, the p-value is written as

$$\mathbb{P}_{\mathrm{H}_0}\left[z_{\alpha/2} \leq \frac{n(\hat{\pi} - 1/2)}{\sqrt{(n/2)^2}} \leq z_{1-\alpha/2}\right].$$

Regarding the two other unilateral hypothesis, the usual reasoning applies. If we have the information that $\xi < \xi_0$, it implies that $\mathbb{P}(X < \xi_0) > 1/2$. Therefore, as usual, the null hypothesis is formulated as the opposite of what we want so that we reject H_0. Therefore, $\mathbb{P}_{\mathrm{H}_0}(X < \xi_0) < 1/2$ and the hypothesis is formulated using the proportion

$$\begin{cases} \mathrm{H}_0 : \pi \leq 1/2 \\ \mathrm{H}_1 : \pi > 1/2. \end{cases}$$

The p-value for n small (exact law) is

$$\mathbb{P}_{\mathrm{H}_0}\left[n\hat{\pi} \leq b_{1-\alpha}\right].$$

For n large, the p-value is

$$\mathbb{P}_{\mathrm{H}_0}\left[\frac{n(\hat{\pi}-1/2)}{\sqrt{(n/2)^2}} \leq z_{1-\alpha}\right].$$

Lastly, the unilateral situation based on the prior information that $\xi > \xi_0$, we have $\mathbb{P}(X < \xi_0) < 1/2$. Therefore, under the null hypothesis, we have $\mathbb{P}(X < \xi_0) > 1/2$ and then the formulation of the hypothesis using the proportion is

$$\begin{cases} \mathrm{H}_0 : \pi \geq 1/2 \\ \mathrm{H}_1 : \pi < 1/2, \end{cases}$$

and the p-value for n small is

$$\mathbb{P}_{\mathrm{H}_0}\left[n\hat{\pi} \geq b_{\alpha}\right].$$

For n large, the p-value is

$$\mathbb{P}_{\mathrm{H}_0}\left[\frac{n(\hat{\pi}-1/2)}{\sqrt{(n/2)^2}} \geq z_{\alpha}\right].$$

Python code:

```
from statsmodels.stats.descriptivestats import sign_test
stat, p = sign_test(data, mu0 = 0)
```

Remark 6.1. This python code does not provide the argument *alternative* to specify the direction of the difference (unilateral testing). It is bilateral by default. If the p-value (p) is significant, less than α, we simply look at the median value of our data and compare it to zero. In this way, we know where the significant direction arises.

6.2 Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is the non parametric alternative of the **independent** two-sample t-test (Subsection 5.2.2).
Similarly to the two-sample Student test, we want to know if our two samples, X_1 and X_2, are drawn from the same population (identically distributed) through the lens of the position parameter, here, the median. The logic is as follows, we put our two samples together and order all the observations $X = X_1 + X_2$ in an increasing way according to their values. Then, we assign a rank to each observations so that the smallest observation has rank 1, the second smallest has rank 2, etc, until the highest observation has rank $n = n_1 + n_2$. Once all the observations are ranked, we separate them according to their initial sample (X_1 or X_2), and we compute the mean of those ranks for each sample,

$$\bar{R}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} R_{i,1},$$

with $R_{i,1}$ the rank i of the observation from X_1 and

$$\bar{R}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} R_{j,2},$$

with $R_{j,2}$ the rank j of the observation belonging to X_2. We also have the overall mean

$$\begin{aligned}\bar{R} &= \frac{1}{n}\left(\sum_{i=1}^{n_1} R_{i,1} + \sum_{j=1}^{n_2} R_{i,2}\right) \\ &= \frac{1}{n}\sum_{r} r \\ &= \frac{1}{n}(1 + ... + n) \\ &= \frac{1}{n}\frac{n(n+1)}{2} \\ &= \frac{n+1}{2},\end{aligned}$$

Which corresponds to the median of the overall sample X.
The formulation of the bilateral hypothesis is as follows

$$\begin{cases} \mathrm{H}_0 : \bar{R}_1 = \bar{R}_2 = \bar{R} \\ \mathrm{H}_1 : \bar{R}_1 \neq \bar{R}_2 \Rightarrow \bar{R}_1 \neq \bar{R}. \end{cases}$$

The later formulation means that if our two samples are identically distributed, their median (rank mean) should be more or less equal to each other and therefore more or less equal to the overall median $\bar{R}$ (H_0). We can then construct a test statistic either on $\bar{R}_1$ or on $\bar{R}_2$ and compare it to $\bar{R} = (n+1)/2$. The test statistic is the following

$$\begin{aligned} w_{emp} &= n_1 \bar{R}_1 \\ &= \sum_{i=1}^{n_1} R_{i,1}. \end{aligned}$$

This statistic follows the Wilcoxon law (for $n_1 \leq 20$ and $n_2 \leq 20$, i.e., exact law) and its formulation explains why this test is called the Wilcoxon rank-sum test. The bilateral p-value writes as

$$\mathbb{P}_{\mathrm{H}_0}\left[w_{emp} \in [w_{\alpha/2}, w_{1-\alpha/2}]\right].$$

For $n_1 > 20$ and $n_2 > 20$, we can use the following asymptotic result

$$\begin{aligned} z_{emp} &= \frac{\bar{R}_1 - \frac{n+1}{2}}{\sqrt{\frac{n_2(n+1)}{12n_1}}} \\ &\overset{\mathcal{L}}{\sim} \mathcal{N}(0,1). \end{aligned}$$

The later result comes from the fact that the statistic $\bar{R}_1$ follows a uniform distribution of expectation $\frac{(n+1)}{2}$ and variance $\frac{n_2(n-1)}{12n_1}$ under H_0. Hence, the p-value becomes

$$\mathbb{P}_{\mathrm{H}_0}\left[z_{emp} \in [z_{\alpha/2}, z_{1-\alpha/2}]\right].$$

Regarding the unilateral hypothesis, for prior information stating that $\bar{R}_1 > \bar{R}_2$ (unilateral greater)

$$\begin{cases} \mathrm{H}_0 : \bar{R}_1 \leq \bar{R}_2 \\ \mathrm{H}_1 : \bar{R}_1 > \bar{R}_2. \end{cases}$$

The p-value in the exact case is

$$\mathbb{P}_{\mathrm{H}_0}\left[w_{emp} \leq w_{1-\alpha}\right].$$

The p-value in the asymptotic case uses the quantile $z_{1-\alpha}$ instead of the Wilcoxon quantile. For prior information stating $\bar{R}_1 < \bar{R}_2$ (unilateral less)

$$\begin{cases} \mathrm{H}_0 : \bar{R}_1 \geq \bar{R}_2 \\ \mathrm{H}_1 : \bar{R}_1 < \bar{R}_2. \end{cases}$$

The p-value in the exact case writes as

$$\mathbb{P}_{\mathrm{H}_0} \left[w_{emp} \geq w_\alpha \right].$$

The p-value in the asymptotic case uses the quantile z_α instead of the Wilcoxon quantile.

Python code:

```
from scipy.stats import ranksums
ranksums(sample1, sample2, alternative= 'two-sided')
ranksums(sample1, sample2, alternative= 'less')
ranksums(sample1, sample2, alternative= 'greater')
```

6.3 Wilcoxon Signed-Rank Test

This test in the non parametric alternative of the **paired** Student test (Subsection 5.2.4) and is used when the normality condition of application is not met. As for the paired Student test, we have two samples X_1 and X_2 and we compute the difference between each ith observations,

$$D_i = x_{i,1} - x_{i,2}.$$

If the distribution of those D_i's is (more or less) symmetric around 0 (check with a boxplot, Figure 10), we can pursue with the test.

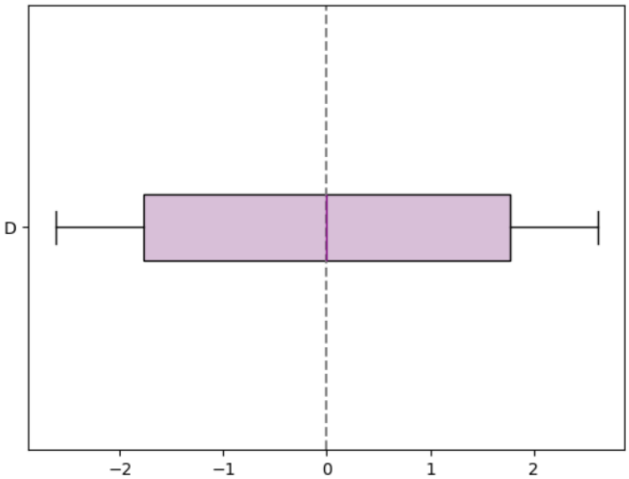


Figure 10: Symmetric distribution around 0.

If this is not the case, we apply the paired sign test (Subsection 6.4). Assuming the D_i's are indeed symmetrically distributed, we define a test statistic that takes into account the rank of only the positive differences,

$$w_{emp} = \sum_{i \in \{1,\ldots,n\}, D_i > 0} R_i,$$

with R_i the rank of D_i. The logic behind this statistic is that we assume X_1 to have higher values than X_2. Hence, they should be drawn from different populations and the magnitude and

frequency of their positive differences should be high. Therefore, the hypothesis is formulated as

$$\begin{cases} \mathrm{H}_0 : X_1 \text{ and } X_2 \text{ have the same distribution} \\ \mathrm{H}_1 : X_1 = X_2 + \delta, \end{cases}$$

with δ a quantity that is non null, i.e., positive or negative. If δ is null, then we are in H_0. Regarding the test statistic, it has been built based on the alternative hypothesis that δ is positive so that $X_1 > X_2$. The test statistic follows a Wilcoxon law under the symmetry of D and under H_0. The p-value, for $n \leq 50$, under the null hypothesis ($\delta = 0$) is as follows

$$\mathbb{P}_{\mathrm{H}_0}\left[w_{emp} \in [w_{\alpha/2}, w_{1-\alpha/2}]\right].$$

If this probability is less than the chosen α level, we consider that $\delta \neq 0$, i.e., X_1 and X_2 are not identically distributed. For $n > 50$, we have the asymptotic statistic under H_0 and under the hypothesis that D has a symmetric distribution around 0,

$$\begin{aligned} z_{emp} &= \sqrt{\frac{24}{n(n+1)(2n+1)}}\left(w_{emp} - \frac{n(n+1)}{4}\right) \\ &\overset{\mathcal{L}}{\sim} \mathcal{N}(0,1). \end{aligned} \tag{4}$$

In what concerns the unilateral hypothesis for $n \leq 50$, if we have prior information telling us that δ is positive, then

$$\begin{cases} \mathrm{H}_0 : X_1 \leq X_2 \\ \mathrm{H}_1 : X_1 > X_2. \end{cases}$$

Hence, the p-value is formulated as

$$\mathbb{P}_{\mathrm{H}_0}\left[w_{emp} \leq w_{1-\alpha}\right].$$

If prior information shows that δ is negative, we have

$$\begin{cases} \mathrm{H}_0 : X_1 \geq X_2 \\ \mathrm{H}_1 : X_1 < X_2, \end{cases}$$

with the associated p-value,

$$\mathbb{P}_{\mathrm{H}_0}\left[w_{emp} \geq w_{\alpha}\right].$$

For $n > 50$, we use the test statistic in eq.4 and formulate the same unilateral p-values as above but with the quantiles of the normal law.

Python code:

```
from scipy.stats import wilcoxon
D = sample1 - sample2
stat,p = wilcoxon(D)
```

6.4 Paired Sign Test

This test is the alternative test of the Wilcoxon signed-rank test when $D = X_1 - X_2$ is not symmetrically distributed around 0. The context is the same as the later test, we have paired samples X_1, X_2 and we use the difference between those observations. We want to know if the two samples are identically distributed, i.e., they tend to have the same values, or if X_1 has values significantly lower than X_2 ($\delta < 0$) or higher than X_2 ($\delta > 0$),

$$\begin{cases} \mathrm{H}_0 : X_1 \text{ and } X_2 \text{ have the same distribution} \\ \mathrm{H}_1 : X_1 = X_2 + \delta. \end{cases}$$

Stated in a bilateral probability manner we have

$$\begin{cases} \mathrm{H}_0 : \mathbb{P}(X_1 < X_2) = \mathbb{P}(X_1 > X_2) = 1/2 \\ \mathrm{H}_1 : \mathbb{P}(X_1 < X_2) \neq \mathbb{P}(X_1 > X_2). \end{cases}$$

We can use the quantity D and rewrite the hypothesis as

$$\begin{cases} \mathrm{H}_0 : \mathbb{P}(D < 0) = 1/2 \\ \mathrm{H}_1 : \mathbb{P}(D < 0) \neq 1/2. \end{cases}$$

The probability $\mathbb{P}(D < 0)$ is the proportion of D located below the median 0. This proportion is denoted as π and allows to once again rewrite the hypothesis as

$$\begin{cases} \mathrm{H}_0 : \pi = 1/2 \\ \mathrm{H}_1 : \pi \neq 1/2. \end{cases}$$

Hence, this proportion π is used to define the test statistic, $n\hat{\pi}$, that is the number of negative differences D_i's. The reader is referred to Subsection 6.1 to see the law under H_0 of the statistic, for n small and large, as well as the formulation of the p-values and of the unilateral hypothesis.

Python code:

```
from statsmodels.stats.descriptivestats import sign_test
D = sample1 - sample2
stat, p = sign_test(D, mu0 = 0)
```

Remark 6.1 also holds here.

6.5 Kruskal-Wallis Test

The Kruskal-Wallis test is the non parametric alternative of the anova (Subsection 5.3). This test is of interest when the normality condition of application of the anova is not satisfied. Still, it requires the **independence** condition on the samples. As for the anova, we are interested in studying the distribution of the (numeric) observations (y) across different levels of the independent categorical variable. For instance, we measure the size of women and their hair color (e.g., blond, brown, red). The typical research question is *Does the hair color influences the size of women?*. Therefore, the hypothesis formulation is as follows

$$\begin{cases} \mathrm{H}_0 : Y_1, Y_2, ..., Y_l \text{ are drawn from the same population} \\ \mathrm{H}_1 : \text{at least one group } Y_i,\ i \in \{1, ..., l\}, \text{ comes from a different population.} \end{cases}$$

This formulation is the same as for the anova but instead of using a parametric measure of the position, i.e., the mean, we use a non parametric measure of the position, that is, the median. We apply the same logic as the Wilcoxon rank-sum test, i.e., we gather all the observations, order them in an increasing way of their values and assign the rank to each ordered observation. The mean rank is $\frac{n+1}{2}$, i.e., the median among all the observations $n = n_1 + n_2 + ... + n_l$.
Stated with respect to the median, we have the following (bilateral) hypothesis

$$\begin{cases} \mathrm{H}_0 : \xi_1 = \xi_2 = ... = \xi_l = \frac{n+1}{2} \\ \mathrm{H}_1 : \xi_i \neq \frac{n+1}{2} \text{ for } i \in \{1, ..., l\}. \end{cases}$$

A candidate for the test statistic is

$$kw_{emp} = \sum_{i=1}^{l} n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2.$$

From that statistic, we see that for $\bar{R}_i \approx \frac{n+1}{2}$, kw_{emp} will be close to zero. On the other hand, if $\bar{R}_i \neq \frac{n+1}{2}$, we have two situations (two unilateral cases). Either $\bar{R}_i < \frac{n+1}{2}$, or $\bar{R}_i > \frac{n+1}{2}$. None of both situations can be captured by the test statistic given the squared difference. Hence, we just look at how high is kw_{emp}. The previous test statistic is a candidate given that we will rather use the following statistic which is a normalized version of the later candidate

$$kw_{emp} = \frac{12}{n(n+1)} \sum_{i=1}^{l} n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2 \overset{\text{asympt.}\mathcal{L}}{\sim} \chi^2_{n-1}.$$

This test statistic follows approximately a χ^2 law (Figure 3) of $n-1$ degrees of freedom which is a right tailed law and therefore the only quantile we are interested in is the right one, i.e., $1-\alpha$. Therefore, the p-value writes as

$$\mathbb{P}_{\mathrm{H}_0} \left[kw_{emp} \leq \chi^2_{1-\alpha,n-1} \right]$$

Remark 6.2. For $l = 3$ (three levels of the variable) and $n_1 \leq 5$, $n_2 \leq 5$ and $n_3 \leq 5$, the sampling distribution under the null hypothesis has been exactly computed but is computationally demanding. Hence we use the approximated law of the statistic, that is, the χ^2 law.

Python code:

```
from scipy import stats
stat,p = stats.kruskal(sample1, sample2, sample3)
```

If the test is significant, we pursue with the Dunn test, that is a non parametric pairwise test, in order to detect where the median inequality comes from.

6.5.1 Non Parametric Multiple Comparisons Test

The hypothesis testing is

$$\begin{cases} \mathrm{H}_0 : \ Y_i \text{ and } Y_j \text{ are drawn from the same population,} \\ \mathrm{H}_1 : Y_i \text{ and } Y_j \text{ come from a different populations.} \end{cases}$$

Mathematically stated

$$\begin{cases} \mathrm{H}_0 : \bar{R}_i = \bar{R}_j \\ \mathrm{H}_1 : \bar{R}_i \neq \bar{R}_j. \end{cases}$$

The pairwise test statistic is formulated as

$$z_{emp,ij} = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}},$$

with $N = \sum_{i=1}^{l} n_i$. This statistic follows a standard normal law. Therefore, we use thoses quantiles to define the p-value as

$$\mathbb{P}_{\mathrm{H}_0} \left[z_{emp,ij} \in [z_{\alpha/2}, z_{1-\alpha/2}] \right].$$

If the later probability is rare, less than α, we reject the mean rank equality between distributions Y_i and Y_j. We repeat this test for each pair i,j and apply a correction of the p-value. Indeed, as explained in Remark 5.8, when applying pairwise comparisons, the type I error increases. It is then recommended to apply a correction such as the Bonferroni correction[9]. This correction consists of either multiplying each obtained p-value by the number of pairwise comparisons, or, lower the α level of rejection by dividing it by the number of comparisons. For instance, assume we have 4 groups, we compute the p-value of each 6 pairwise comparisons. We then multiply each obtained p-value by 6 and reject if the p-values are less than α. Or, we multiply the α level of rejection by 6 and compare the obtained p-values to $6 \times \alpha$.

Python code:[10]

```
!pip install scikit-posthocs
import scikit_posthocs as sp
groups = [[1,2,3,5,1], [12,31,54, 9], [10,12,6,74,11]]
sp.posthoc_dunn(groups, p_adjust = 'bonferroni')
```

7 Tips For Students

The best advice we can give to student while writing the statistical part of their Master thesis is to be transparent. Transparency means reporting the motivations that lead to the choice of a specific test. The motivations refer, notably, to the conditions of application (the visuals and results of the formal tests).

8 Resources

The main resources used in this draft are from the courses provided in the Master of Statistics at the ULB and from the french book *Probabilités et inférence statistique* (2020) by C. Dehon, M. Hallin, D. Paindaveine, C. Thomas-Agnan and C. Vermandele.

[9] https://en.wikipedia.org/wiki/Bonferroni_correction

[10] https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_dunn.html#scikit_posthocs.posthoc_dunn

A Descriptive Statistics

A.1 Prior Information

A.1.1 Fancy Plot

Python code Figure 2:

```
pastel_mauve = "#D8BFD8"
data = [group_1,group_2]
plt.boxplot(data, patch_artist=True,
            boxprops=dict(facecolor=pastel_mauve, color='purple'),
            capprops=dict(color='purple'),
            whiskerprops=dict(color='purple'),
            flierprops=dict(markerfacecolor=pastel_mauve, markeredgecolor='purple'),
            medianprops=dict(color='purple'))
plt.xticks([1,2], ['group_1','group_2'])
plt.show()
```

B Conditions of Application

B.1 Normality

B.1.1 Definitions

Definition B.1. (Quantile) The quantile α (say) of a law refers to the values below which $\alpha\%$ of the observations are located. It is the same idea than the quartiles, but it is not restricted to the quarts of the distribution. Visually, on Figure 4, we have the quantile $t_{\alpha/2}$ that refers to the x values for which the area under the curve in the interval $[f(0), f(x)]$ is $\alpha/2$. Note that for a symmetric distribution such as the normal and the Student distribution, $t_{\alpha/2} = -t_{1-\alpha/2}$.

Definition B.2. (Degrees of freedom) The degrees of freedom is the number of components of a vector that need to be known to fully determine all the components of that vector [11]. For instance, the statistic of the χ^2 law uses the estimator of the mean, $\bar{x}$. To compute this estimator we use all the observations, the n observations. Once we have computed $\bar{x}$, we need to know the $n-1$ observations to deduce from them and from $\bar{x}$ the nth observation. Therefore, we say that the degrees of freedom is $n-1$ for n observations.

Definition B.3. (Statistic) A statistic is a function of the observations x_i. This is a random variable given that it varies with the observations at our disposal.

B.1.2 Fancy Plots

Python code Figure 5:

```
pastel_mauve = "#D8BFD8"  # mauve pastel
border_color = 'purple'   # contour plus foncé
plt.hist(data, bins=50, color=pastel_mauve, edgecolor=border_color)
plt.show()
```

Python code Figure 7:

```
pastel_mauve = "#D8BFD8"
mauve_fonce = 'purple'
```

[11] https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)

```
osm, osr = stats.probplot(data, dist="norm", plot=None)[0]
plt.figure()
plt.scatter(osm, osr, facecolors=pastel_mauve, edgecolors=mauve_fonce)
slope, intercept = stats.probplot(data, dist="norm", plot=None)[1][:2]
x = np.linspace(min(osm), max(osm), 100)
plt.plot(x, slope * x + intercept, color='black')
plt.grid(True)
plt.xlabel("Theoretical quantiles")
plt.ylabel("Observed quantiles")
plt.title("")
plt.show()
```

C Parametric Tests

C.1 Definitions

Definition C.1. (Population parameter) The population parameter refers to a quantity that we will estimate based on a reduced amount of observations (our sample). To fully characterize the parameter we would need all the population (all the observations) which is resource intensive. Therefore, we infer to value of the population parameter using an estimator of the later.

D Non Parametric Tests

D.1 Wilcoxon Signed-Rank Test

D.1.1 Fancy Plot

Python code Figure 10:

```
plt.boxplot(D, vert=False, patch_artist=True,
            boxprops=dict(facecolor="#D8BFD8"),
            medianprops=dict(color="purple"))
yticks = plt.yticks()[0]
yticklabels = [str(int(tick)) if tick != 1 else "D" for tick in yticks]
plt.yticks(yticks, yticklabels, rotation=0)
plt.axvline(0, color='gray', linestyle='--')
plt.show()
```