Inferential Statistics

The problem of **inferential statistics** rises when data refer not to the whole population as subject of the study (the statistical population), but they refer only to a part of it. However, the aim of the research is to draw conclusions on the whole population.

In the production field, these varieties of needs are frequent in a great amount of circumstances and in very different areas of interests.

Example

- ⋆ It has to be checked if the switches, which go out from the production chain, respond adequately to certain kinds of destructive stress tests. The aim, therefore, is to know the percentage of switches that do not respond to our expectations. In a similar case, it is not possible to check all the produced switches.
- ⋆ It is desirable then, starting from a sample, to accurately determine which will be the percentage of faulty switches in the whole production.
- ⋆ Starting from that, one might also want to find out the percentage of faulty switches. This percentage, identified through the sample, will support or reject the hypothesis that the ratio of defectiveness is part of the previously specified parameters.

Inferential statistics techniques provide the guidelines and the formulas to obtain these kinds of information.

These techniques are divided into three fundamental subgroups:

1. Techniques for the point estimation;
2. Confidence intervals valuation;
3. Statistical hypothesis test.

The point 1 techniques are often specific to the single problem that has been analysed. These varieties of techniques usually lead to intuitive results. On the other side, the points 2 and 3 approaches are less intuitive, even though they use similar techniques in the various applications. Afterwards, the techniques which regard the points 2 and 3, will be deepened according to common and frequent problems.

It is supposed to analyse a population in order to obtain information about a characteristic (X) known to be normally distributed with unknown mean μ and known σ standard deviation. That is:

$$X \sim N(\mu,\ \sigma)$$

It is of interest to know the value of the unknown parameter μ. In order to have this information, all the population requests should be measured and then averaged out.

Depending on the size of the population and on the kind of population, this aim could become complex and hard to achieve, or almost impossible.

An alternative could be drawing a sample of n elements of the population and then calculating the sample mean ($\overline{x}$). Afterwards, this value will be used as the best estimation or the polite choice as regards the value of μ.

If we repeat the same process, the second time we will almost certainly obtain a different value of $\overline{x}$.

If we repeat the process k times, we will obtain a $\overline{x}_1$, $\overline{x}_2$, ..., $\overline{x}_k$ series of k samples means. All of them will almost certainly be different and probably near to the true value of the unknown parameter (μ).

To $k \to \infty$ it is possible to obtain, in addition, the distribution of the random variable of the sample mean of a sample with size n. But $k \to \infty$ is still equivalent to measure the whole population.

Fortunately it is possible to prove that, if X_i is a generic value of our population and if $X_i \sim N(\mu,\ \sigma)$, then, drawing a sample $X_1, X_2, \ldots, X_i, \ldots, X_n$ of sample size n:

$$\overline{X} \sim N\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right)$$

In other words the sample mean random variable, calculated on a sample of size n drawn from a population normally distributed with mean μ and standard deviation σ, is also **normally distributed with the same mean and with standard deviation $\sqrt{n}$ times smaller than the standard deviation of the population**.

As consequence, if this is right, it is also possible to calculate an interval in which it will be included the 95% of the possible requests for the same sample mean random variable

$$P\left\{\mu - \frac{\sigma}{\sqrt{n}} \cdot z_{0.975} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}} \cdot z_{0.975}\right\} = 0.95$$

Whose result is:

$$P\left\{\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{0.975} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{0.975}\right\} = 0.95$$

Where $z_{0.975}$ represents the quantiles that leaves at its left the 97,5% in a $N(0,1)$.

Therefore, instead of providing a single value for the estimation of μ, once that the realization of $\overline{X}$ (called $\overline{x}$) is known, one can provide an interval of possible values for μ. This result is provided with a certain confidence level (in this case, the 95%).

As the normal distribution is symmetrical, then $z_{0.975} = -z_{0.025}$. For this reason the previous formula can be rewritten as:

$$P\left\{\overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{0.025} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{0.975}\right\} = 0.95$$

The meaning of the confidence interval has to be understood like follows.

If we draw out an infinity of samples of the population, and every time we calculate the corresponding 95% confidence intervals, then the 95% of these confidence intervals will contain the true value of the population mean μ.

In other words, the 5% of the intervals should not contain the true value of the population mean.

The 95% is not the only confidence interval that can be taken in consideration. An interval could be considered, for example, 90%, or 99% or even 99.9%.

A 99% interval would have the unquestionable advantage of containing the true value of the mean (μ) in the 99% of cases, but this interval will be wider of the corresponding 95% interval of a little less smaller quantity of $2\sigma/\sqrt{n}$.

A way to reduce the wideness of the interval, with all confidence levels being equal, consists in increasing the sample size n.

This case has a limit because there is a square root in the denominator. As consequence, for each of the added sample unit, the advantage will be smaller and smaller.

In other words, an optimal confidence interval does not exist. Each confidence interval represents a compromise between the interval wideness, the confidence level and the number of samples unities.

As before, it is supposed a normally distributed population, with unknown mean μ and known variance σ^2 that is:

$$X \sim N(\mu,\ \sigma)$$

It is supposed to check the hypothesis that the population mean is equal to a certain value (μ_0). In this case, a null hypothesis (H_0) is assumed:

$$H_0 :\ \mu = \mu_0$$

This hypothesis is in contrast to an alternative hypothesis (H_A):

$$H_A :\ \mu \neq \mu_0$$

The alternative hypothesis is rarely shown as $H_A :\ \mu = \mu_A$, with $\mu_A \neq \mu_0$.

It has to be found a criterion that allows us to choose one alternative between the following two alternatives:

- ⋆ to accept H_0 and to refuse H_A;
- ⋆ to refuse H_0 and to accept H_A.

Knowing that:

- ⋆ the distribution of X is known: $X \sim N(\mu,\ \sigma)$;
- ⋆ σ is supposed to be known and constant ;
- ⋆ μ is unknown and constant.

If a sample of n elements from the population has to be extracted and the analysed dimension X has to be measured, it is possible to affirm that: the random variable $\overline{X}$, given from the sample mean of the observations, will be normally distributed around μ with standard deviation equal to $\frac{\sigma}{\sqrt{n}}$:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The random variable (**statistics test**) Z is defined as :

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If H_0 is true, prior to the extraction of the sample, Z will be:

$$Z \sim N(0, 1)$$

After the experiment, the z value (the realization of Z) will be then, presumably, "close to zero" and included with probability of 95% in the interval ± 1.96.

If H_0 is not true (but H_A is true), prior to the extraction of the sample, Z will be:

$$Z \sim N\left(\frac{\mu_A - \mu_0}{\sigma/\sqrt{n}},\ 1\right)$$

After the experiment, the z value (the realization of Z) will be then, presumably, "far from zero" and external from the interval ± 1.96.

The z value resulting from the experiment, will be called z_0.

The null hypothesis is accepted if, after the test, the z_0 value is included in the interval ± 1.96.

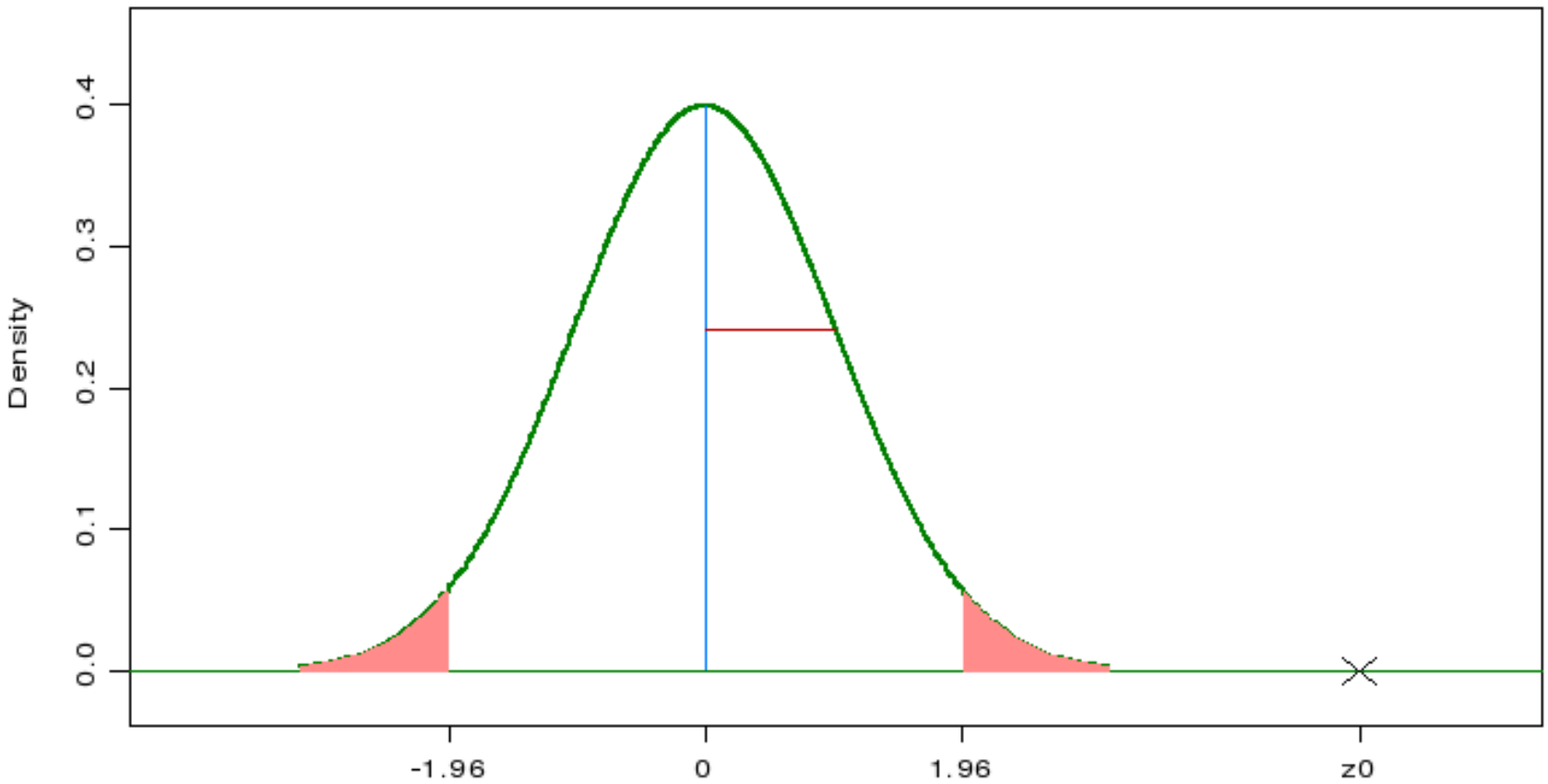
Alternatively, the null hypothesis is rejected in favour of the alternative hypothesis if z_0 is not included in the interval ± 1.96.

The null hypothesis is accepted if the z_0 value is included in the interval ± 1.96 independently from it is on the right or on the left of the value 0.

An alternative graphic representation consists in bending the left half in its right side. It consists then in considering the refusal area of the hypothesis like it was just on the right side of the curve
In this case the z value has to be considered in absolute value.

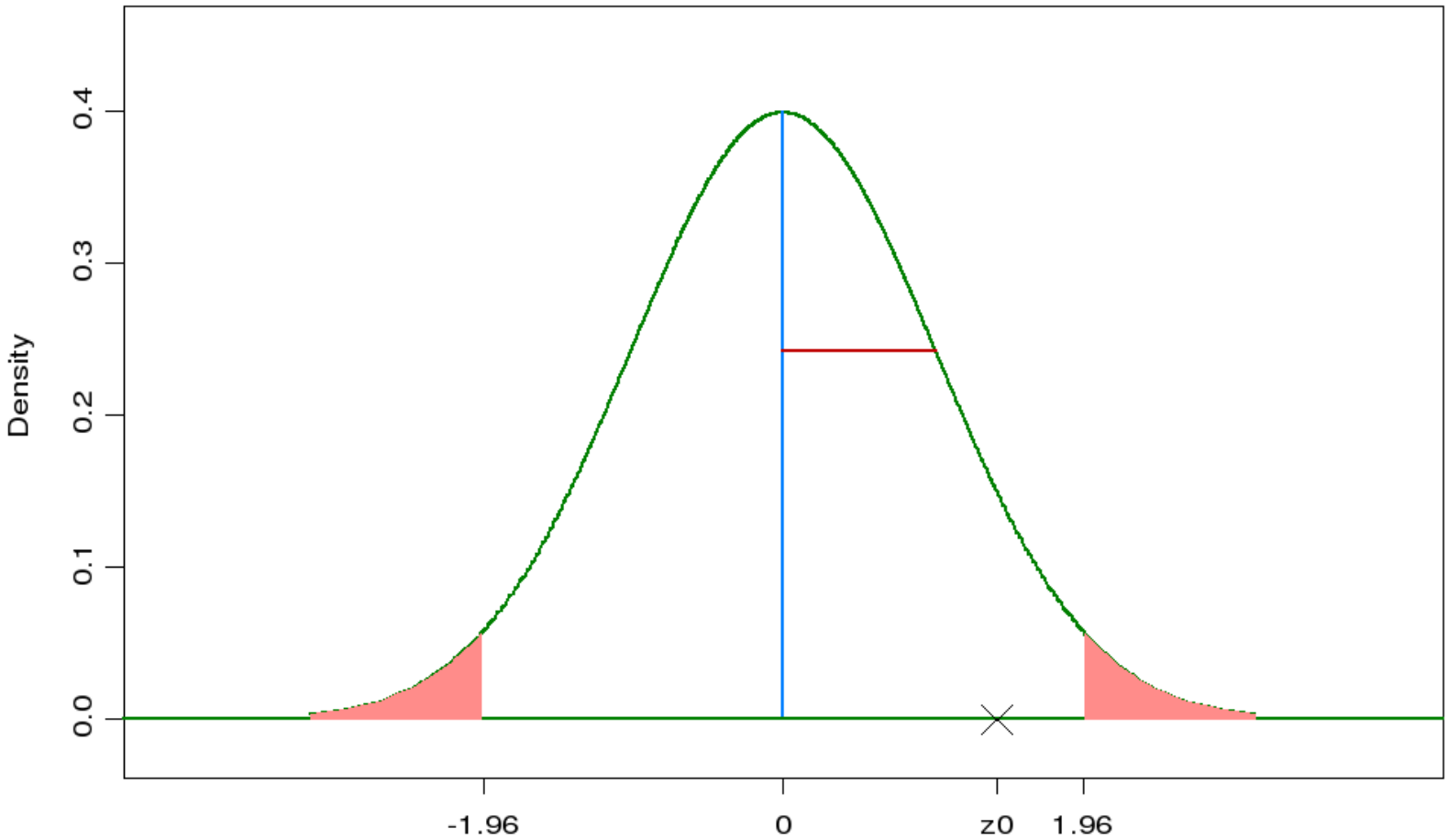
The null hypothesis H_0 is refused

(The curve represents the Z distribution if H_0 is true)



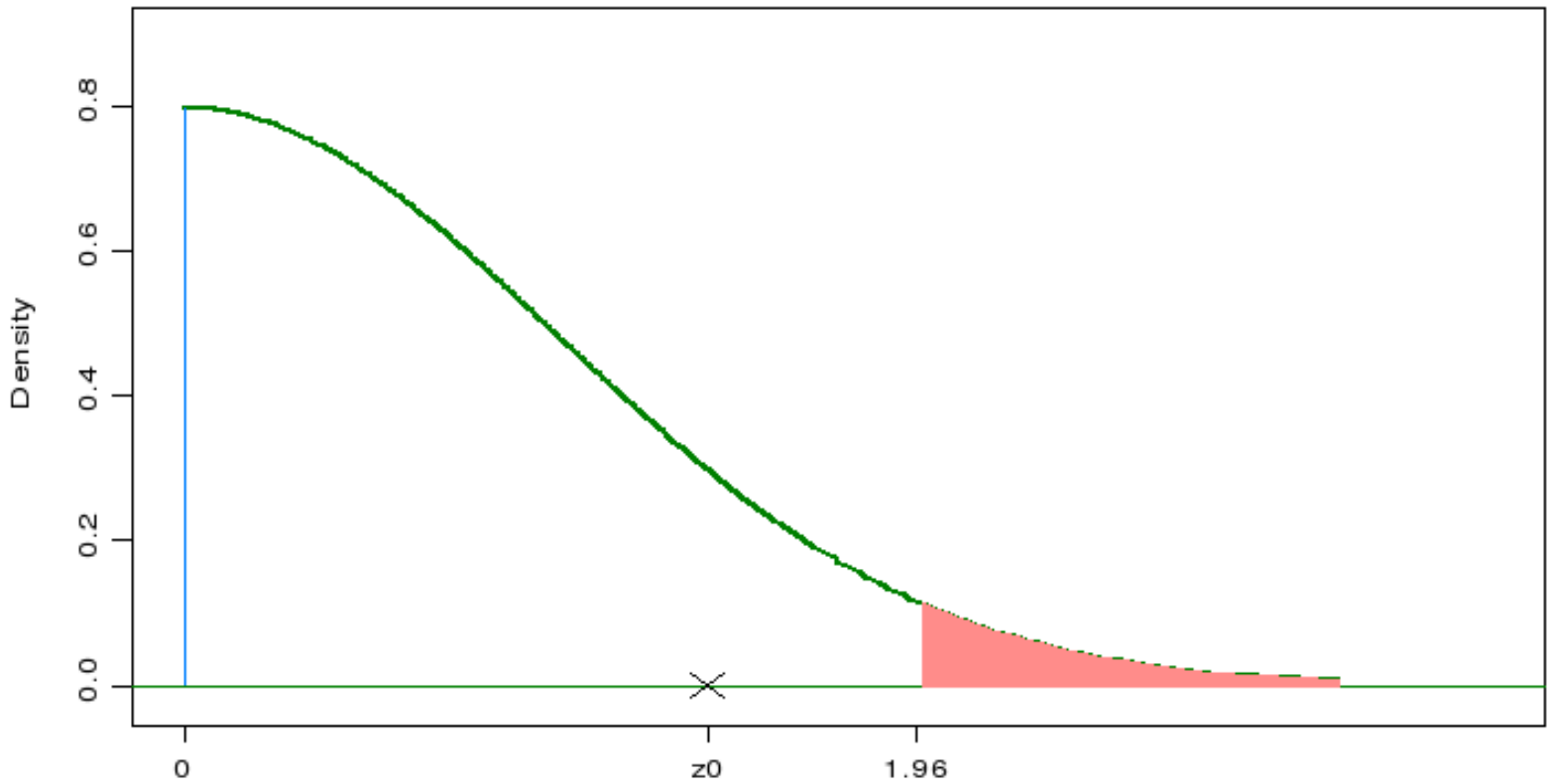
The null hypothesis H_0 is accepted

(The curve represents the Z distribution if H_0 is true)



The null hypothesis H_0 is accepted

(The curve represents the Z' distribution, the Z absolute value, if H_0 is true)



Important notes.

The choice of the value 1.96 in the previous formulas is not unjustified, even though it is, in a certain way, arbitrary.

If the null hypothesis (H_0) is valid, with the value 1.96, it is possible to have a 0.05 (5%) probability to commit an error and to refuse this hypothesis.

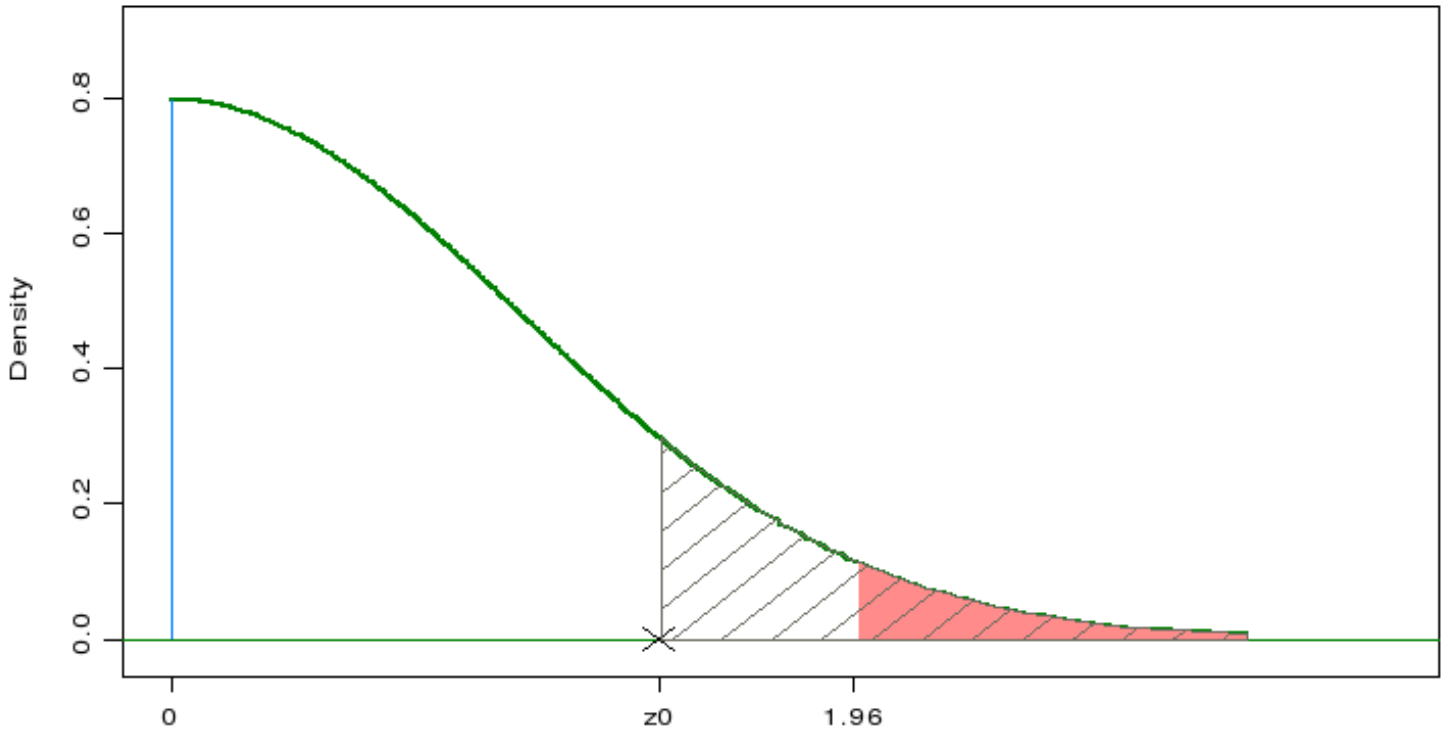
This probability is usually shown with the symbol α, its corresponding error is called Type I error (or error of the first kind).

In the statistical hypothesis tests, the Type I error is the only kind of error which is directly verifiable by the reasearcher. Indeed, in the case taken into consideration, when α increases or decrease, also the "cut off" value (1.96) increases or decreases.

When α is fixed and, as consequence, the cut-off value is fixed too, the test is called "of level α".

The value 1.96 leaves at its right area a 5% probability.

If the null hypothesis is accepted, or rather if the z_0 absolute value is at the left of the value 1.96, then the area on the right of the z_0 absolute value must be greater than 5%.



Thence, another similar criterion to check the null hypothesis consists in measuring the area on the right of the z_0 value. This measure has to be taken on the distribution that we would have if the null hypothesis is true.

If $P\{Z' > z_0|H_0\} \geq 0.05 \Rightarrow$ we accept H_0

If $P\{Z' > z_0|H_0\} < 0.05 \Rightarrow$ we refuse H_0

The z_0 of this formula, that refers to the Z', distribution, are the z_0 absolute values (previously seen), referred to the Z distribution.

The quantity $P\{Z' > z_0|H_0\}$ is called **p-value**. This criterion has the advantage of being release from the specific value 1.96. In the same way, it is also release from any other value that isolates, on its right, an area of probability α different from 5%.

In the experimental research, the use of an inference test through the sample mean $\overline{x}$ with σ (the population standard deviation) known it is a case in a class by itself.

When the population mean μ is unknown, also the standard deviation σ is unknown. As consequence, it is necessary to use a substitute of the variance (the square of the standard deviation) of the population, **the sample variance s^2 represents its most logic and most reliable estimation**.

The statistic test, in this case, is called t, and it is defined as follow:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

With unknown σ and the use of s **as its replacement**, the statistic test prior to sample extraction does not follow the normal distribution Z, previously seen, but the so called **Student's t** distribution.

For a better comprehension of the fundamental concepts and a proper application of the derived tests, it is important to highlight the specific characteristics that distinguish this distribution to the Gaussian:

- ⋆ **the normal distribution considers the $\overline{X}$ mean sample variation**;
- ⋆ **the Student's t distribution also takes into consideration the sample variability of the standard deviation estimation (s)**.

In order to perform inference on the population mean, starting from sample data, it is necessary to consider at the same time:

- ⋆ **the variability of $\overline{x}$ as estimation of μ;**
- ⋆ **the variability of s as estimation of σ.**

With the increase of the sample size n, s proves to be an always better estimation of σ.

When n is enough great (in theory infinite, in practice over 120-150), s **and σ are almost equivalents**.

The t-distribution's mean is 0 (the quantity at the numerator is the difference between the sample mean random variable and its expected value) and the variation that depends on its degrees of freedom (equal to the dimension of the sample minus 1).

As consequence, it could be affirmed that **when n increases, the result is the convergence of the Student's t distribution towards the Z** standard normal distribution.

With little samples the difference between the Student's **t** statistics value and the correspondent Z statistics value at the same probability α is relevant. On the contrary, over some tens of observations, it is unimportant.

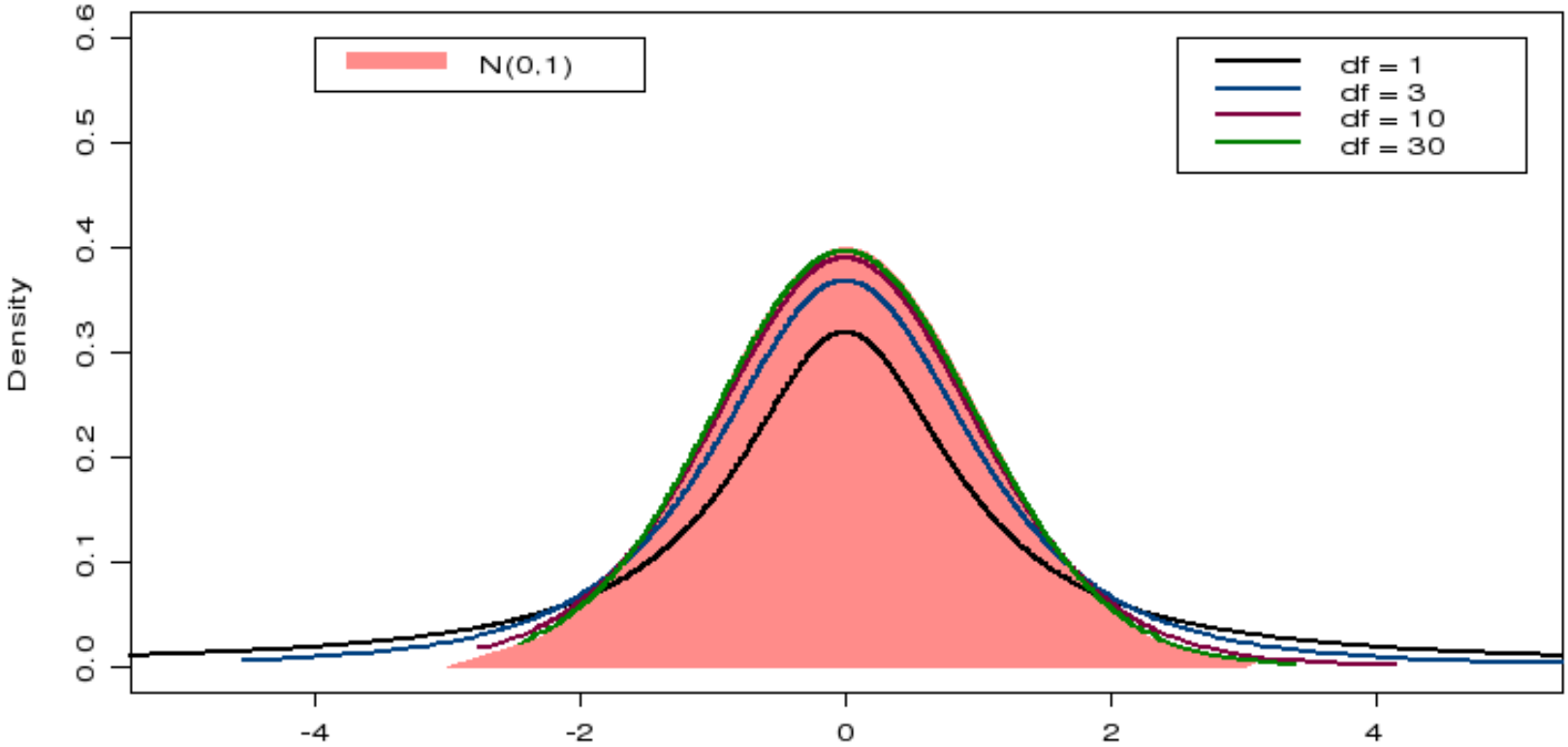
In the experimental practice, **the methods which use the t test usually refer to little samples**, the observations are often ten or less and they rarely reach some tens.

The form of the Student's t distribution is symmetric and bell shaped as the normal distribution, but with a larger dispersion (variability).

It exists a whole family of t-distributions, one for each value of the degrees of freedom (df) parameter.

For an infinite value of df, in practice for a number of data little greater than an hundred, the t distribution curve and the Z distribution curve coincide.

From the mathematic aspect, this means that **the normal distribution represents the limit of the t distribution, when the number of df tends to infinite**.



Standard normal distribution and Student's t distributions.

From the formula $t_{(n-1)} = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$, with μ which, in this case, represents the true and unknown population mean from which the sample is extracted, it is possible to derive the confidence interval at the level $(1-\alpha)\%$, for the true mean μ, with unknown σ.

Once that $\overline{x}$ and s, are calculated, the confidence interval with its extremes becomes

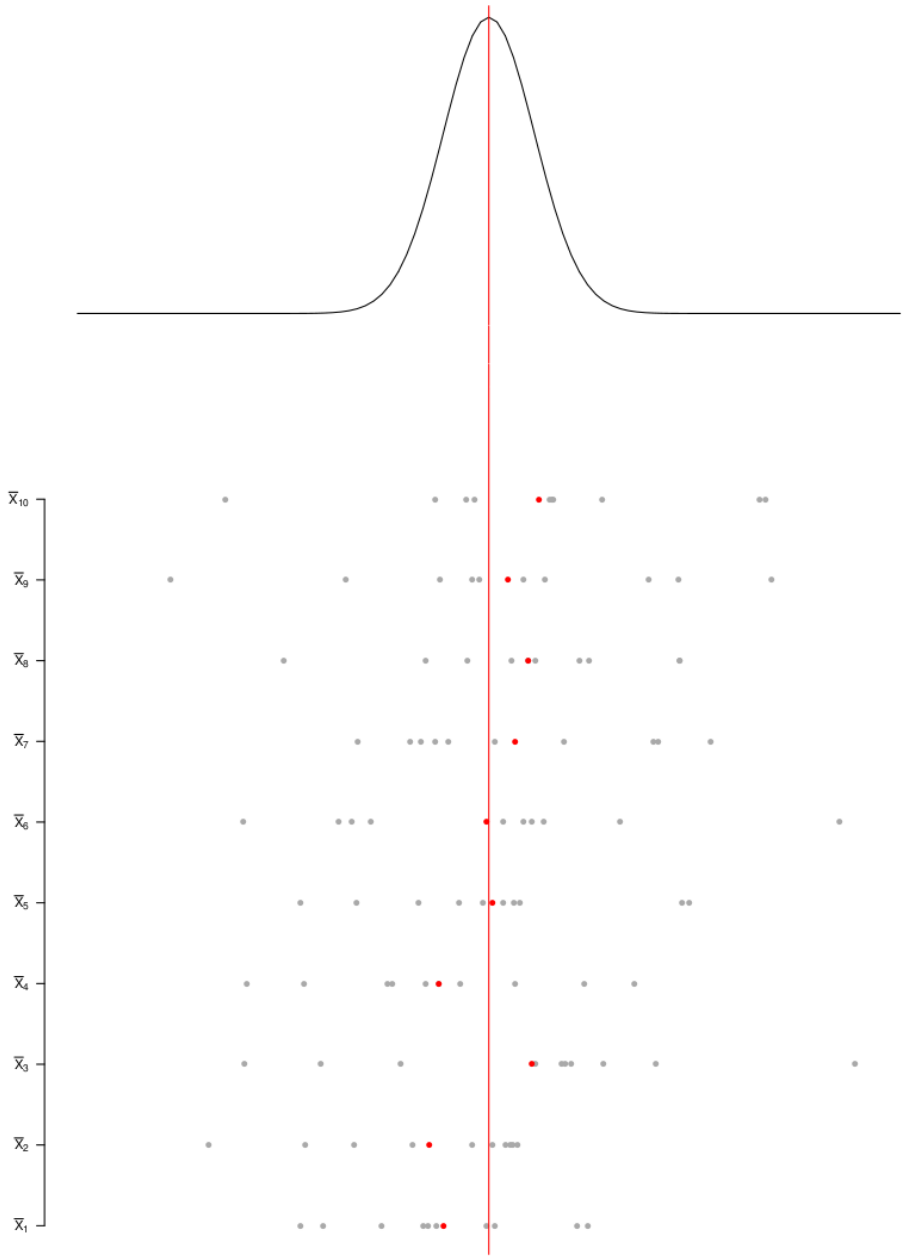
$$\overline{x} - t_{\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2};n-1}$ indicates the t-distribution value with $n-1$ degrees of freedom that "leaves at its right" a $\frac{\alpha}{2}$ probability.

The applicability conditions to obtain the Student's t distribution, and therefore, for its use in the tests (we will see them later) and in the confidence intervals are:

- ⋆ **the distribution of the single sample units (X) is normal**;
- ⋆ **the observations are independently collected**.

T-tests are robust. This means that **their distribution is approximately valid also for data distribution markedly drawn away from the normality.**
This is valid especially if the sample size is large (see Central Limit Theorem).


$\bar{X}_{10}$
$\bar{X}_9$
$\bar{X}_8$
$\bar{X}_7$
$\bar{X}_6$
$\bar{X}_5$
$\bar{X}_4$
$\bar{X}_3$
$\bar{X}_2$
$\bar{X}_1$

In general, a test is defined robust as regards the applicability condition, when its results can be accepted even though the applicability condition is not rigorously verified.

In the applied statistics, there are three versions of t-test used to check means:

- ⋆ **It checks if the true mean of values of a population is equal to an hypothesize** or expected value (comparison between the observed mean and the hypothesized mean);
- ⋆ **It checks if the difference between the true means of values of two independent populations is equal to an hypothesized value** (comparison between independent samples);
- ⋆ **It checks if the difference between the true means of values of two dependent populations is equal to an hypothesized value** (comparison between dependent samples).

In a bilateral test, in order to specify the hypothesis about the μ population mean as regards the expected mean, it will be used the previously seen symbols:

- **The null hypothesis H_0** shown as: $H_0 : \mu = \mu_0$
- **The alternative hypothesis H_A** shown as: $H_A : \mu \neq \mu_0$

Where:

- μ is the value (true and unknown) of the population mean from which the sample is extracted and from which will be calculated the sample mean;
- μ_0 is the hypothesized value of the mean used as point of mark for the comparison.

With the same symbols it is possible to show an unilateral statistical hypothesis test that the population mean, from which the sample is extracted is equal to μ_0 like follows. This is against the alternative that it is smaller than μ_0 will be:

- ⋆ **Null hypothesis H_0**: $H_0 : \mu = \mu_0$
- ⋆ **Alternative hypothesis H_A**: $H_A : \mu < \mu_0$

Notes.

- ⋆ Unilateralism and bilateralism of a test do not depend on the null hypothesis but on the alternative hypothesis. As a matter of fact, the complete formulation is: test with bilateral and unilateral alternatives.
- ⋆ The difference between the bilateral and unilateral tests is not only applied to t-test but also to Z-test and to all the tests that can make this difference.

Effectively.

After having defined the quantity

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Then **under the hypothesis in which H_0 is true**, the t quantity will be able, prior to extract the sample, to assume a random value that comes from the Student's t distribution with $n - 1$ degrees of freedom.

So, before the experimentation, it is expected that the t value will be "concentrated around zero".

In the hypothesis that H_0 is not true (H_A is true), we will have:

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{(\overline{X} - \mu)}{\frac{s}{\sqrt{n}}} + \frac{(\mu - \mu_0)}{\frac{s}{\sqrt{n}}} \sim \left[t_{n-1} + \frac{(\mu - \mu_0)}{\frac{s}{\sqrt{n}}} \right]$$

Before the experimentation, it is expected that the t value will be approximately far from zero.

The t value resulting from the experimentation is called t_0 or rather the t random variable realization.
According to the formulated alternative hypothesis (unilateral or bilateral), there are two alternatives that respond to the test.

Bilateral hypothesis:

- H_0 is accepted with an error level of $\alpha = 0.05$ when the t_0 value is between the interval $\pm t_{n-1;0.025}$. Where $\pm t_{n-1;0.025}$ is the value of the distribution t_{n-1} that leaves at its right the 2.5% of its overall area.
- Alternatively, the null hypothesis is refused in favour of the alternative hypothesis.

Unilateral hypothesis:

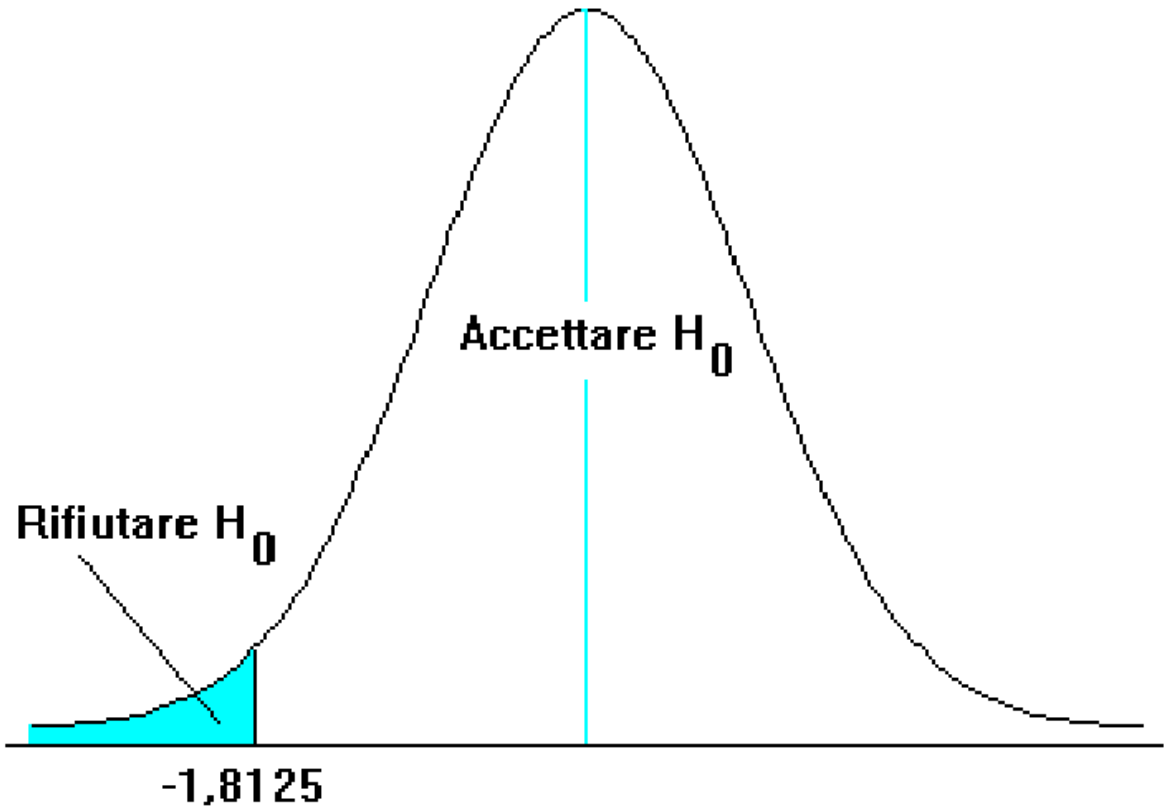
- The null hypothesis is accepted with an error level of $\alpha = 0.05$ if $t_0 \geq t_{n-1;(1-0.05)}$ This happens, as in the example, because the alternative hypothesis (H_A) establishes that $\mu < \mu_0$.
- Alternatively, if $t_0 < t_{n-1;(1-0.05)}$, the null hypothesis is refused in favour of the alternative hypothesis.

A test **is unilateral or a one-sided test**, when the researcher asks himself if a mean is larger than the other, excluding that it could be smaller.

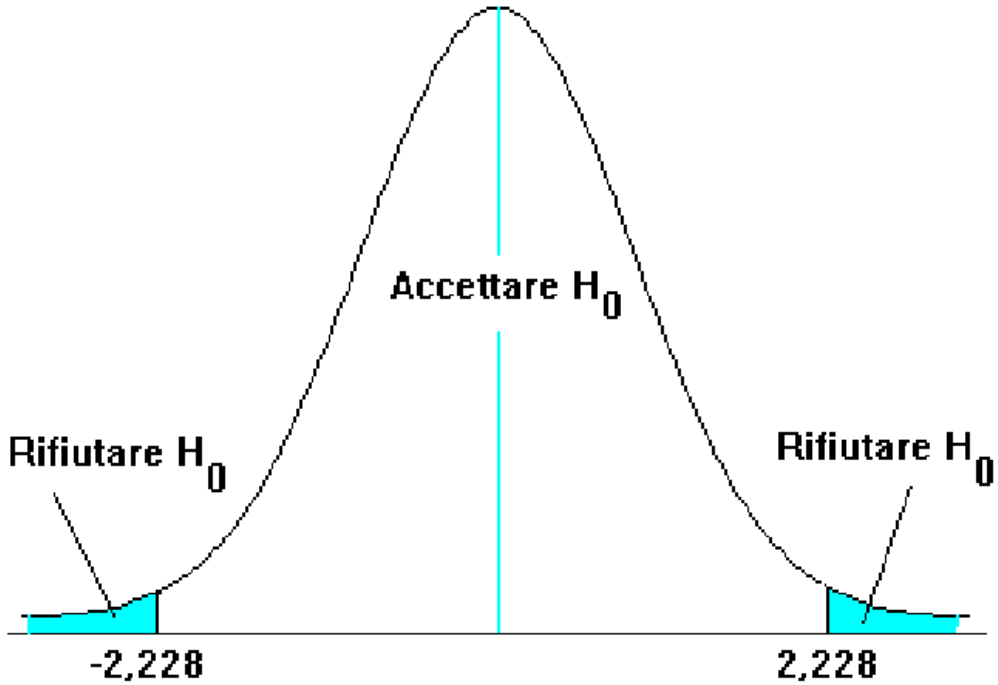
A test **is bilateral or a two-sided one**, when the researcher asks himself if a significant difference between the two means exists. In this case he does not have information about which one is the larger or the smaller.

In the case of the one-sided test, the refusing area will be in only a part of the distribution (at the left when the sign is negative, at the right when it is positive). In the case of the two-sided test, the refusing area will be symmetrically distributed in the two parts

The probability to demonstrate a significant difference is higher in the case of the one-sided test compared to the two-sided test. With a technical term **the two-sided test is more conservative, while the one-sided test is more powerful**.



The unilateral test for a probability associated to a level of statistical significance of 5% with 10 df.



The bilateral test for a probability associated to a level of statistical significance of 5% with 10 df.

If the test is unilateral it is possible, with the same data, to refuse the null hypothesis. On the contrary, if the test is bilateral, this is not possible.

It is possible to use the p-value standard to accept or refuse the null hypothesis both for the t test and for the Z test and all the other statistics tests. In this case, the distribution area has to be measured in the case in which the null hypothesis is valid, for values of the statistics tests that go beyond the value that has been found for the same test. For example, with an alternative unilateral hypothesis like $H_A : \mu < \mu_0$, and a result t_0 of the statistics test, then:

If $P\{t_{n-1} > t_0 | H_0\} \geq 0.05 \Rightarrow$ we accept H_0

We $P\{t_{n-1} > t_0 | H_0\} < 0.05 \Rightarrow$ we refuse H_0

If the p-value results to be greater than 0.05, then the null hypothesis is accepted. Otherwise, the null hypothesis is rejected.

Student's t test can be also used to compare the means of two samples.

The comparison between two means can be done:

- ⋆ **with two dependent samples**;
- ⋆ **with two independent samples**.

It is important to distinguish the two situations which depend on the way the two compared measurements are obtained. The two tests have differences:

- ⋆ In the t test application procedures;
- ⋆ In the way the subjects variability effects are measured.

The characteristic which distinguishes the comparison between the two dependent samples is to be able to associate each of the samples observations to just one of the observations of the other sample. The two groups have always, necessarily, the same number of data.

T test for two dependent samples is required when the same unities have to be analysed in different conditions.

For example, to estimate the mean productivity of the staff (in terms of the mean time for each operation) at the beginning and at the end of the working day and to check if this mean changes or not.

The comparison between the means of the two series of observations is easy: **the analysis is applied to a new series of data. This series is the result of the differences between the elements of each couple**.

For the Student's t test, **in the case of a bilateral test** the null hypothesis H_0 is usually that the differences mean is equal to an hypothesize value which is usually 0:

$$H_0 : \ \delta = \delta_0$$

while the alternative hypothesis H_A is:

$$H_A : \ \delta \neq \delta_0$$

In an **unilateral test**, **the null hypothesis** H_0 is that the differences mean is greater than or equal to the hypothesize one (it is often 0):

$$H_0 : \ \delta \geq \delta_0$$

while **the alternative hypothesis** H_A is that the difference is less than the hypothesize one (it is often 0). It can be written like follows:

$$H_A : \ \delta < \delta_0$$

On the opposite case, **the null hypothesis** H_0 is that the differences mean is less than or equal to the hypothesize one (it is often 0):

$$H_0 : \ \delta \leq \delta_0$$

while **the alternative hypothesis** H_A is that the difference is greater than the hypothesize one (it is often 0). It can be written like follows:

$$H_A : \ \delta > \delta_0$$

In order to choose between the null hypothesis H_0 and the alternative hypothesis H_A, the statistical significance of the means of the differences is checked by the ratio

$$t_{(n-1)} = \frac{\overline{d} - \delta_0}{\frac{s_d}{\sqrt{n}}}$$

where:

- $\overline{d}$ is the mean of the differences;
- δ_0 is the hypothesize mean difference: it is often, but not necessarily, equal to 0;
- s_d is the standard sample deviation calculated on the differences;
- n is the number of differences. It also corresponds to the number of couples of data (that usually is half of the number of the observations).

The confidence interval of the differences mean between the two dependent samples is calculated (with the same symbols of the previous formula) by

$$\delta = \overline{d} \pm t_{\frac{\alpha}{2};n-1} \cdot \frac{s_d}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2};n-1}$ indicates the t distribution value with $n-1$ df that leaves at its right a probability amount equal to $\alpha/2$.

In this formula it is included, at the level $1-\alpha$, the true mean δ of the differences.

According to various authors, in the applied field, when the number of couples of data is over 40 (others indicates 50 or even 150), the Student's t distribution is adequately approximated to the standard normal distribution.

In the variance calculation formula, with a difference in the results that a great number of researchers consider unimportant, **when the dimension of n measures some tens, it is possible to replace the critic t value with the Z value associated to the prefixed probability α. This situation happens both in the one-sided tests and in the two-sided tests**.

In a great amount of cases, it is not possible and even plausible to get two dependent samples.
The only possible strategy for the analysis of data is **to compare two independent samples and two samples formed by different individuals**.

In this case **the result is the increase of the variability within the two groups** but it is also possible to obtain three advantages:

- ⋆ **a different number of observations in the two groups can be used**;
- ⋆ **to have data that are easier the expression of the random variability**;
- ⋆ **to be able to use samples taken from different units for the comparison**.

With the statistical significance test for two independent samples, it is checked **the same hypothesis of the case of paired data, even though it is expressed in a different way**.

It is fundamental to understand that:

- ⋆ for two dependent samples, the calculation are done just on the differences;
- ⋆ in the case of two independent samples **the calculation are done on the two series of observations**.

In a **bilateral or a two-sided test**, **the null hypothesis** H_0 is that the difference between the two populations (A and B) is equal to δ_0. Quite always δ_0 is equal to 0 in the practice use.

The null hypothesis can be written as:

$$H_0 : \mu_A - \mu_B = \delta_0 \quad \text{or} \quad H_0 : \mu_A = \delta_0 + \mu_B$$

and its **bilateral alternative hypothesis** H_A can be written as:

$$H_A : \mu_A - \mu_B \neq \delta_0 \quad \text{or} \quad H_A : \mu_A \neq \delta_0 + \mu_B$$

An example **of one-sided or unilateral test** is:

$$H_0 : \mu_A - \mu_B \leq \delta_0 \quad \text{or} \quad H_0 : \mu_A \leq \delta_0 + \mu_B$$

against the alternative hypothesis:

$$H_A : \mu_A - \mu_B > \delta_0 \quad \text{or} \quad H_A : \mu_A > \delta + \mu_B$$

In the case of two independent samples, in its easier formulation, the value of t is obtain like follows:

$$t_{(n_A+n_B-2)} = \frac{(\overline{X}_A - \overline{X}_B) - \delta_0}{\sqrt{S_p^2 \cdot (\frac{1}{n_A} + \frac{1}{n_B})}}$$

And the degrees of freedom of t are equal to
$(n_A - 1) + (n_B - 1) = (n_A + n_B - 2)$, or $(N - 2)$.

- $\overline{X}_A$ and $\overline{X}_B$ are the means calculated respectively on the (sub) sample A and on the (sub) sample B;
- $\delta_0 = (\mu_A - \mu_B)$ is the hypothesized difference between the means. It is expressed in the null hypothesis;
- n_A and n_B are the number of observation in the (sub) samples A and B;
- $N = n_A + n_B$;
- S_p^2 is the (*pooled*) variance of the compared groups.

The pooled variance (S_p^2) is the result of the ratio between the sum of the two deviances and the sum of the respective df:

$$S_p^2 = \frac{\sum_{i=1}^{n_A} (X_{A_i} - \overline{X}_A)^2 + \sum_{i=1}^{n_B} (X_{B_i} - \overline{X}_B)^2}{n_A - 1 + n_B - 1}$$

where:

- X_{A_i} and $\overline{X}_A$ are respectively the data and the sample mean of the group **A**;
- X_{B_i} and $\overline{X}_B$ are respectively the data and the sample mean of the group **B**;
- n_A and n_B are the number of observation of samples **A** and **B**.

The pooled variance **is a weighted average variance, calculated starting from the two deviances and from their df**. The pooled variance assigns a proportionally higher importance to the group which has a greater number of data.

$$t_{(n_A+n_B-2)} = \frac{(\overline{X}_A - \overline{X}_B) - \delta}{\sqrt{S_p^2 \cdot (\frac{1}{n_A} + \frac{1}{n_B})}}$$

From the above t test formula it is possible to deduce that, in order to calculate the test it is not necessary that all the data (of the extracted selection unities) are available.

The formula can be calculated starting from the means, the standard deviations and the sample size of the two samples.

The Student's t test (in the form we have just seen) is a **parametric statistics test**. In other words, it is based on the characteristics of the **normal distribution** which is defined by **parameters** like the **mean** and the **variance**, the **symmetry** and the **kurtosis**. The application of this test, in order to be considered valid, **required the three following conditions**:

- ⋆ **the independence of the data within and between the samples**;
- ⋆ **the homogeneity of the variances** (the comparison between two or more means is valid **only if** the samples extracted from the populations have a similar variance);
- ⋆ **the data (or the residual compared to the mean) are normally distributed**.

Whit **two independent samples**, the most important applicability condition is the variances uniformity because, compared to it, the t **test is less robust**.

In order to **calculate the S^2 pooled**, it is necessary that condition of **homoscedastic** is realised. This means that **the two variances should be statistically equal**.

The hypothesis of independent collection of data depends on the plan of the experimentation. The hypothesis of normality of the data or of the error (the data residuals from their mean) can be violated without serious effects on the power and on the result of the test; this happens unless there is a serious asymmetry but **the equality of the two independent samples variance should be always checked**.

From the intuitive point of view, this concept can be easily explained. The variance can be considered as a reliability estimation of a mean that represents the population.

The data that are really variable (with a wide variance) have less reliable means for the same number of observations. This happens because they are more variable, as their data. To compare the two means it is necessary that their reliability is similar.

For the application of the t-test, the homoscedastic between two groups (**A** and **B**) is checked with a **bilateral test**, where the null hypothesis H_0 and the alternative hypothesis H_A are:

$$H_0 : \sigma_A^2 = \sigma_B^2 \qquad\qquad H_A : \sigma_A^2 \neq \sigma_B^2$$

The most popular tests to check the homoscedastic are three:

- ⋆ the **F-test** or the ratio between the two variance;
- ⋆ the **Levene's test**;
- ⋆ the **Bartlett's test**.

It should be possible to use **Student's t test for two independent samples** only if **it is demonstrated that the null hypothesis is true** and therefore that the two groups have statistically equal variances.

The **bilateral F-test**, is the most difficult and the first that has been purposed. It is founded on the **ratio between the larger sample variance (s^2) and the smaller sample variance**:

$$F_{(n_1-1);(n_2-1)} = \frac{S_1^2}{S_2^2}$$

where:

- S_1^2 is the larger variance;
- S_2^2 is the smaller variance;
- n_1 is the number of observations in the group with the larger variance;
- n_2 is the number of observations in the group with the smaller variance.

Under the hypothesis in which the two variances are equal (the null hypothesis H_0 is true), the result of the ratio between them should be close to 1. It is obviously admitted a certain tolerance, because the estimation of the two samples variances is never exact.

In the practice, the variances are often estimated on small samples, formed by few units of observations. As consequence, the ratio between the two variances is a sample estimation, that could vary from **one to infinite** or **from one to zero**.

In order not to use both the measures, which will give a redundant information, it has been chosen the values distribution which is more sensitive to the variations: that from one to infinite.

If the hypothesis H_0 is valid, then the quantity $F_{(n_1-1);(n_2-1)}$ will be a F of Fisher-Snedecor random variable with $(n_1 - 1)$, $(n_2 - 1)$ degrees of freedom. It will tend to concentrate most part of its distribution around 1.

If, on the contrary, the hypothesis H_A is valid, then $F_{(n_1-1);(n_2-1)}$ tends to have greater values than the expectation.

It will be necessary, then, to check with which probability a F random variable tends to have values equal or greater than the one obtained by the samples data. If this probability is little (this means, less than α), then the null hypothesis H_0 is rejected in favour of the alternative one. Otherwise, the null hypothesis that the two variances are not statistically different, is accepted.

The **Levene's Test** is an **alternative method**. It can be used also to integrate the analysis that has been done with the F-test when a more thorough valuation on the two variances homogeneity wants to be found.

Levene's test is considered from certain statisticians **more robust than F-test** as regards the non-normality of the distribution. It is mainly spread because of its introduction in certain statistical tools.

This test compares the residuals either from the means or from the medians of the two groups i.e. the differences (d_i) compared to the means (or medians) of the groups:

$$d_i = X_i - \overline{X}$$

after

- ⋆ having taken the square root

$$d_i = (X_i - \overline{X})^2$$

- ⋆ or taking them in absolute value

$$d_i = |X_i - \overline{X}|$$

in order to **eliminate the negative signs**.

The two methods provide **different results**: the variance is larger and then the power is smaller with the use of the square of the residuals.

The second method, that uses the residual in absolute value **is the most used one** mainly in the computer programs. It has similar tests in the non-parametric statistics.

In order to compare the variance of the two groups (**A** and **B**),

with null hypothesis

$$H_0 : \sigma_A^2 = \sigma_B^2$$

and with bilateral alternative hypothesis

$$H_A : \sigma_A^2 \neq \sigma_B^2$$

Levene's propose consists in the application of the Student's t test to the two series of residuals (in square or in absolute value). It assumes that, **if their means values are significantly different, the two variances of the original data are different too**.

If, **by the use of the residuals mean**, the null hypothesis is refused

$$H_0 : \mu_A^2 = \mu_B^2$$

in order to accept the alternative hypothesis

$$H_A : \mu_A^2 \neq \mu_B^2$$

(where the μ_K^2 in this case, represent the means residuals in square or in absolute value in the two groups).

implicitly it comes that:

on **the original datai** the null hypothesis is refused $H_0 : \sigma_A^2 = \sigma_B^2$

in favour of the alternative hypothesis $H_A : \sigma_A^2 \neq \sigma_B^2$

Also **Levene's test can be useful to check the unilateral hypothesis**

the null hypothesis $H_0 : \sigma_A^2 \geq \sigma_B^2$

against the alternative hypothesis $H_A : \sigma_A^2 < \sigma_B^2$

In order to make the Bartlett's test, it is enough to apply the unilateral t-test to the residuals means.

Bartlett's test uses the χ^2 distribution.

In order to check the hypothesis of homoscedastic between the two independent samples, the synoptic table of $\chi^2_{(1)}$ **with one degrees of freedom is used**.

It is **important** to remember that it is necessary to act cautiously in the use of the inference for the homoscedastic. As consequence, this is also valid to conclude that two variances are similar and to apply Student's t test on the means. **Statistics tests are generally structured not to refuse the null hypothesis**, unless there is evidence of the contrary: **The null hypothesis is refused only when it is proved that there is a little probability to find, by coincidence, differences as large as the found one, or larger**.

In order to have a significant test, if the researcher has few data, the differences between the two groups have to be really large: **A test with few data is not very powerful. This means that it has few possibilities to refuse the null hypothesis, even when it is known that it is false**.

As consequence, mainly in these cases, **when the null hypothesis H_0 is not refused, it is not possible to affirm that it is true**. It is only impossible to affirm that it is false because of the few information.

In operational terms it is possible to go out from this illogical use of the tests for homoscedastic thanks to **a more complete and detailed valuation of the estimated probability α**.
With few data, if the probability α is little higher than 0.05 (for example 0.10 or 0.15) it is not possible to refuse the null hypothesis. If it would have been possible to have a greater number of data, with an high probability, it would have been possible to refuse the null hypothesis.

In practice, according to certain authors, it would be better to accept the null hypothesis only when the calculated probability α is high, higher than 30%. With a greater number of data, this probability can be lowered to 20%. Unfortunately, it does not exist specific rules to define a large or small sample and to decide to which level of probability α it is possible to affirm that the null hypothesis is true. The choice is based on the statistical common sense, that can only come from the experience.

The tests that have been analysed so far are based on the conditions of normality of data (even though, sometimes, not in a "strong" way because of the robustness of the t test to the deviation of the conditions of normality).

Anyway, in the real world, the natural phenomenon does not always (sometimes rarely) behave like Gaussian random variables.

It is obvious that a certain argument cannot be taken for granted and has to be, on the contrary, checked on the data.

With this aim, different statistics tests have been produced. Some of them are the numerical type and others the graphic type.

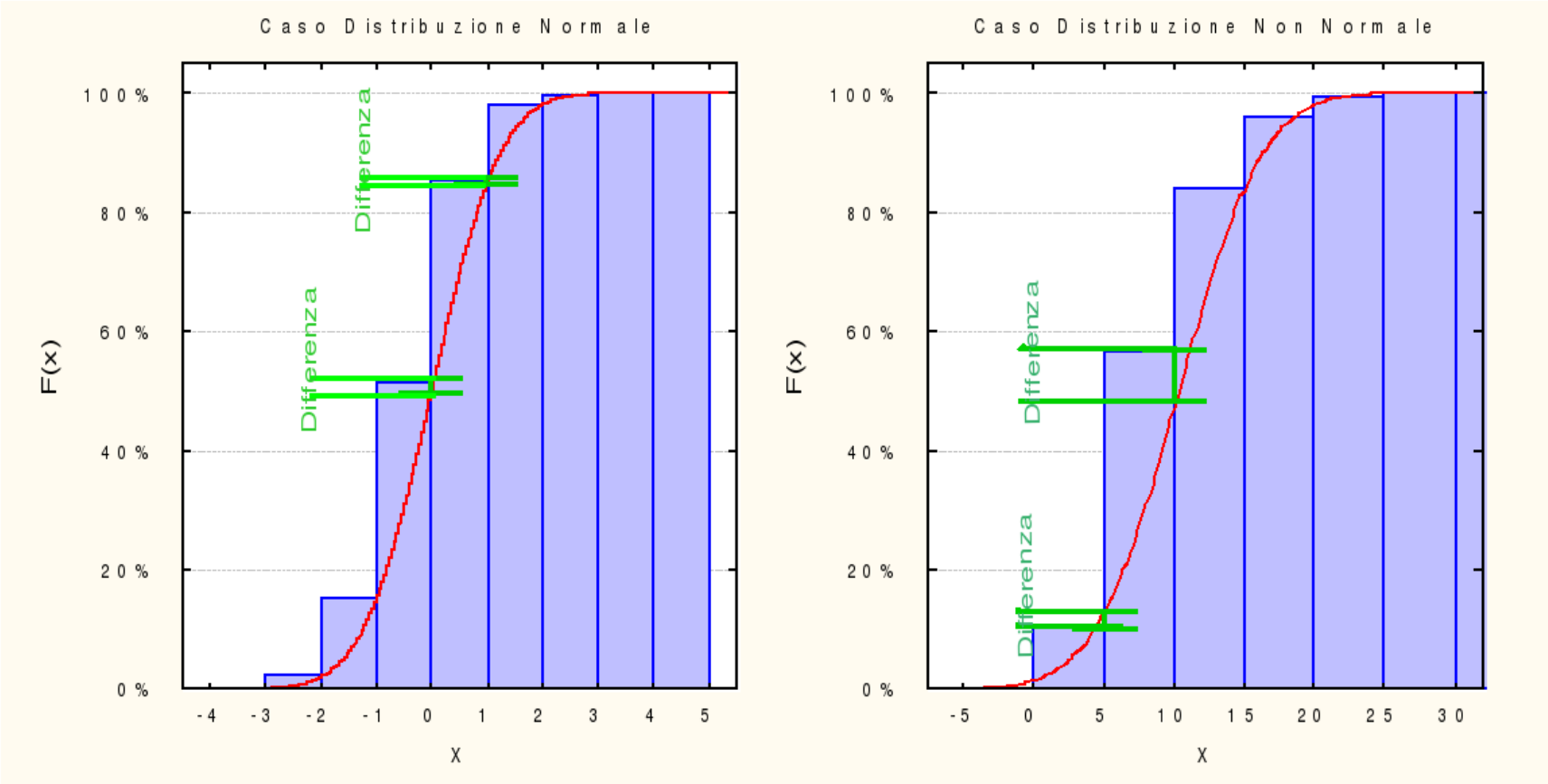
Some of the most known numerical tests of normality are:

- ⋆ Kolmogorov-Smirnov test
- ⋆ Kolmogorov-Smirnov test with Lilliefors' tabulated probability
- ⋆ Shapiro-Wilk W test
- ⋆ Anderson-Darling test

The most known graphic test to check the normality is the *Normal Probability Plot*.

Anderson-Darling test and the *Normal Probability Plot* are afterwards described.

The following plot represents the cumulative histograms of two distributions (one normal and the other not normal) with the normal cumulative distribution (estimated) written above.



The Anderson-Darling test calculates the square mean difference between the "true" cumulative distribution (the one obtained by the observed data) and the theoretic (obtained by the hypothesis in which the data are distributed as a Gaussian). The approximation of this difference is calculated by the following formula:

$$AD = \sum_{i=1}^{n} \frac{1-2i}{n} \{\ln(F_0[Z_{(i)}]) + \ln(1 - F_0[Z_{(n+1-i)}])\} - n$$

where n is the sample dimension, $Z_{(i)}$ are the orderly and standardized values that has been observed, F_0 is the cumulative distribution of the standard Gaussian.

In the hypothesis in which data are effectively generated by a Gaussian (the picture on the left), it is expected a little mean quadratic difference.
In the hypothesis in which the data are generated by a different distribution from the Gaussian, it is expected a large mean quadratic difference.

The p-level (or the rejection cut-off value) depends on the analysed distribution and it is not a simple calculation.

When the normality is checked, the AD value for which rejecting the null hypothesis, at the level $\alpha = 0.05$ is given by the formula:

$$AD > \frac{0.752}{(1 + \frac{0.75}{n} + \frac{2.25}{n^2})}$$

The **normal probability plot** is a graphic tool. It is useful to evaluate the normality of the distribution of a sample of observed data.

It is supposed to have a sample of n independent observations: $x_1,\ x_2,\ \ldots,\ x_n$ come from the same unknown distribution

The data are arranged according to their value, in an ascending order. In this way an ordered sample is obtained: $x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)}$.

The ith observation of the ordered sample will correspond to the ith observed quantiles. Furthermore, it will be a reasonable estimation of the ith quantiles of the distribution (true and unknown) from which the data come.

The idea that stands at the basis of the normal probability plot is to compare the estimated percentiles values with the values of the correspondent theoretic percentiles that come from the standard normal distribution.

Example.

It is supposed to have ten observations (already ordered):

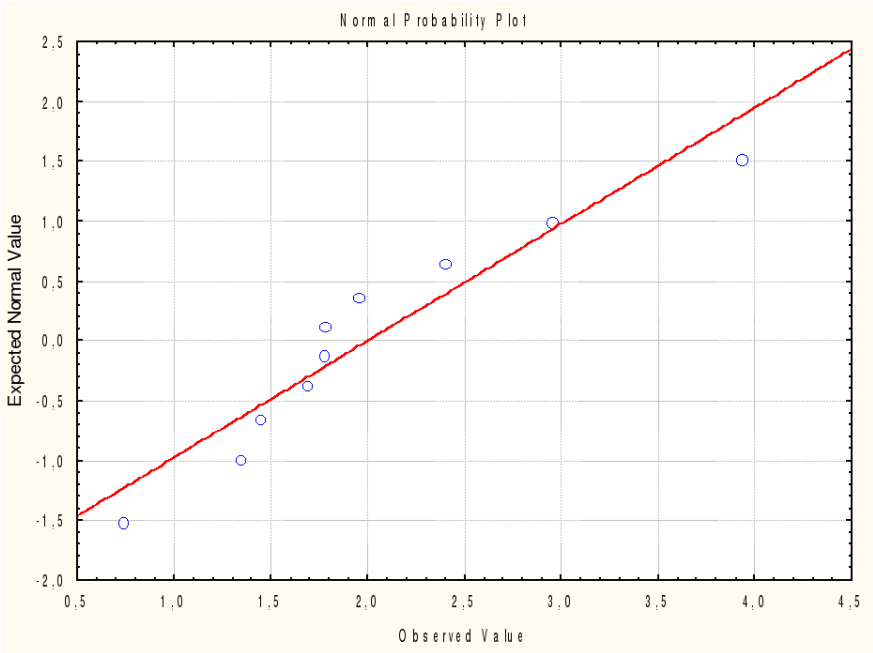
0.73 1.34 1.44 1.68 1.77 1.78 1.95 2.4 2.95 3.93

The first observation is the estimation of the first deciles (quantiles 1/10) for the data origin distribution. The value of the correspondent quantiles for the standard normal distribution is more or less -1.28.

The second observation is the estimation of the second deciles (quantiles 2/10) for the data origin distribution. The value of the correspondent quantiles for the standard normal distribution is more or less -0.84.

The third observation is the estimation of the third deciles, and so on.

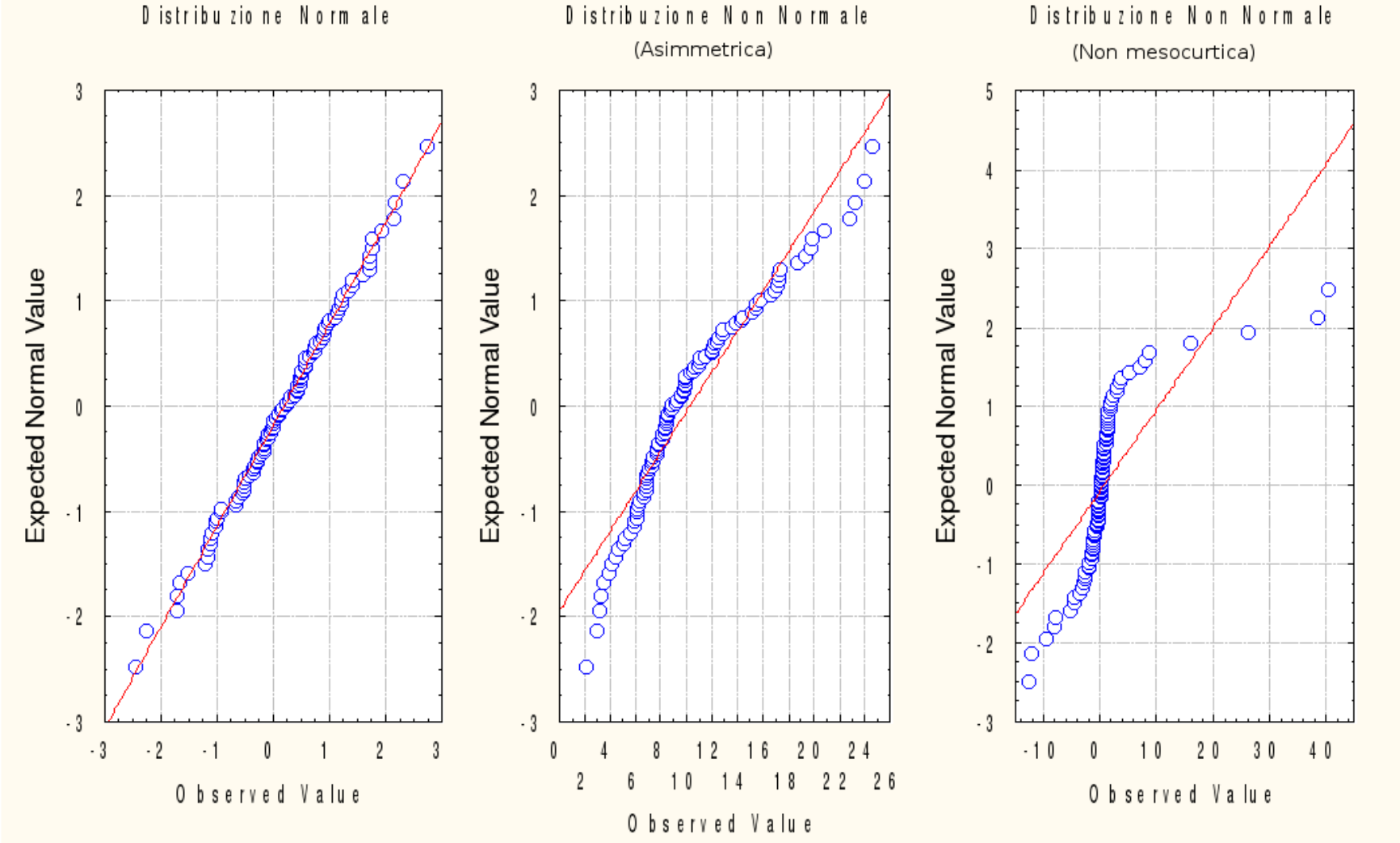
The result of the elaboration of the data is a table with $n = 10$ couples of values. If these couples are drawn in a scatter plot, the result is the following plot:



If the points tend to place around a line, then it is possible to accept the hypothesis that the origin distribution of our data is normal.

Some notes.

- ⋆ The line that has been drawn in the previous plot is usually added in order to facilitate the reading of the "linearity" of the points disposition. Particular meanings of the slope or of the intercept of this line do not have to be searched.
- ⋆ The percentiles estimation usually follows some rules a little bit more complex than the previous one. As a result, it will be obtained more considerable estimations. Anyway, the logic of the plot creation is the one we talked about.
- ⋆ In the normal probability plot the data in the tails of the plot are usually less reliable. This happens mainly when there is a small sample size.
- ⋆ The previous plot is concerning a sample that comes from the normal distribution. Take in consideration the sample size!
- ⋆ Afterwards, some examples of normal probability plot for large samples that come from different distribution will be explained.


Distribuzione Normale
Expected Normal Value
Observed Value
Distribuzione Non Normale
(Asimmetrica)
Expected Normal Value
Observed Value
Distribuzione Non Normale
(Non mesocurtica)
Expected Normal Value
Observed Value

For qualitative type phenomenon, the interest is often in the (percentage) ratio of units which present the interesting characteristic.

For exampleit can be useful to know: in the case of a productive process, the percentage of the defective pieces; in a clinical reasearch, the proportion of people who positively respond to a medicine; in a bank, the proportion of insolvent customers.

All of these phenomenon can be modeled with a binomial random variable of parameters n and p. n indicates the number of proofs (trials) and p indicates the probability of success in every single trial, which is supposed to be equal in each trial.

It wants to be checked the hypothesis that the true (and unknown) proportion of interest events (happened) is p_0 starting from sample data. This is against the alternative hypothesis which is different.

The null hypothesis is formulated: $H_0 : p = p_0$,
Against the alternative hypothesis: $H_A : p \neq p_0$.

x is the number of "successes" obtained after having done n trials. $\hat{p} = \frac{x}{n}$ is the proportion of "successes" in the sample.
In its easiest formulation, the p-test for a proportion is like follows:

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

That, if H_0 is valid and n is enough large, it is distributed like a normal standard.

Example.

A factory that produce DVDs wants to check the hypothesis that the proportion of defective pieces is 0.01 ($=1\%$), to a level of statistical significance $\alpha = 0.05$.

With this aim, it selects, by chance, 100 pieces and it tests them. The result of the test shows that there are always two defective pieces.

In order to check the null hypothesis $H_0 : p = 0.01$ against the alternative hypothesis $H_A : p \neq 0.01$, it is necessary to use the previously seen statistics:

$$\frac{(2/100) - 0.01}{\sqrt{0.01(1-0.01)/100}} = 0.01/\sqrt{0.000099} = 1.005$$

It is possible to conclude that **it does not exist empirical evidence that allows to refuse the null hypothesis**. This conclusion is obtained because the obtained quantity (1.005), in absolute value, is smaller than the cut-off value of a normal standard for $\alpha = 0.05$, which is 1.96.

It is possible to reach the same conclusion observing the p-value, that can be calculated as the probability to have, in a standard normal distribution, a value smaller than -1.005 or the probability to have a greater than 1.005 one. This probability is 0.315. Since the p-value is greater than the fixed α value, the null hypothesis is not refused.

It is possible to build a confidence interval for the "true" proportion of successes. This is, similarly to what has been seen for the Z-test, the interval:

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}; \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

Example.

The true proportion of defective DVDs at a confidence level of 0.95, with the data we have seen in the previous example, is:

$$\left(\frac{2}{100} - 1.96 \cdot \sqrt{\frac{0.01(1-0.01)}{100}}; \frac{2}{100} + 1.96 \cdot \sqrt{\frac{0.01(1-0.01)}{100}}\right)$$

That is the interval $(0.0005; 0.0395)$.

As regards both the statistical hypothesis test and the building of the confidence intervals, various methodologies exist. Those that have been already shown, are the easiest versions.

As regards the statistical hypothesis tests, it also exists an exact version of the test, based on the Beta distribution.

In the case of two means coming from different samples, it has been seen how it is possible to check the hypothesis that the two populations true (and unknown) means, are equal. In the same way, in the case of two samples, it is possible to check, starting from sample data, the hypothesis that the true proportions (p_A and p_B) of two different populations are equal.
The null hypothesis is formulated: $H_0 : p_A = p_B$,
Against the alternative hypothesis: $H_A : p_A \neq p_B$.

In its easiest formulation, the p-test for two proportions has the following form:

$$\frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_A(1-\hat{p}_A)/n_A + \hat{p}_B(1-\hat{p}_B)/n_B}}$$

That, if H_0 is true and with n_A and n_B enough large, it is distributed as normal standard.

Example.

It is supposed that two different machineries, A and B, have produced respectively 200 and 500 pieces within a working day. 3 of the pieces produced by the machinery A and 4 of the pieces produced by the machinery B were defective.

It wants to be checked the hypothesis that the two machineries produce the same proportion of defective pieces, against the hypothesis that the proportion is different, at a level of statistical significance $\alpha = 0.05$.

By the application of the previously seen formula, it is obtained:

$$\frac{(3/200) - (4/500)}{\sqrt{(3/200)(1-(3/200))/200 + (4/500)(1-(4/500))/500}} = 0.73$$

It is possible to say that **it does not exist empirical evidence that allows to refuse the null hypothesis**. This conclusion comes from the fact that the obtained quantity (0.73), in absolute value, is smaller than the cut-off value of a normal standard for $\alpha = 0.05$, which is 1.96.

It is possible to reach the same conclusion observing the p-value that can be calculated as the probability to have, in a standard normal distribution, a value smaller than -0.73 or the probability to have one greater than 0.73. This probability is 0.465. Since the p-value is greater than the fixed α value, the null hypothesis is not refused.

A large number of versions exist also for the p-test for two proportions. Here it has been presented the easiest version (for the convenience of the exposition).

The statistical hypothesis test for two proportions can also be generalized to verify that a difference equal to δ exists between the two proportions:

$$H_0 : p_A - p_B = \delta$$

In conclusion, both for the test for one proportion and the test for two proportions, it is possible to check the null hypothesis against the unilateral alternative hypothesis. The cut-off value, with which it is possible to compare the obtained statistics, and the p-value, can be similarly calculated to what has been previously seen.

The χ^2 (Chi-Squared) test was originally projected to try out the concordance between two distributions, one theoretic and the other observed.

This test does not formulate distributive arguments for its own applicability. **It is often used to check the association between the various modalities of two or more qualitative characters**.

It is supposed to have some statistics samples on the number of defective pieces. These pieces have been found in the factory and are associated to the used machinery. The machineries are 3: A, B and C. It wants to be checked if the number (to better say, the percentage) of the defective pieces depends or not on the machinery which produces them.

		Conditions of the pieces	
		Flawless	Defective
Machinery	A	132	13
	B	99	18
	C	157	11

Two alternative and totally equally hypothesis can be formulated supposing that the data that have been collected are illustrated in the table. These data have also to be coded in percentages.

The percentage of defective pieces for the three machineries are more or less the same.

		Conditions of the pieces	
		Flawless	Defective
Machinery	A	91.0%	9%
	B	84.6%	15.4%
	C	93.5%	6.5%

The breakdowns are distributed almost uniformly in the three machineries.

		Conditions of the pieces	
		Flawless	Defective
Machinery	A	34.0%	31.0%
	B	25.5%	42.9%
	C	40.5%	26.2%

Each of the two proportions of previous hypothesis can be expressed as:
H_0: it does not exist a link between Machinery and Breakdowns.

Against the alternative hypothesis:
H_A: it does exist a link between Machinery and Breakdowns (or, the percentage of defective pieces is greater for one or more machineries or the breakdowns tend to be greater in one or more machineries).

In a general form: it is supposed to note down on n objects, two qualitative characteristics. The first one (Q1) composed by r modalities, and the second one (Q2) composed by c modalities. It is necessary to valuate if the distribution of these two characteristics **is not associated** within the whole population.

A frequencies table with this structure can be built:

	Q2			
Q1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
	$\cdots$	n_{ij}	$\cdots$	$\cdots$
	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
	$n_{.1}$	$\cdots$	$n_{.c}$	n

where n_{ij} represents the number of objects which present the ith characteristic of Q1 and the jth of Q2. $n_{i.}$ represents the total for the ith row and $n_{.j}$ represent the total of the j column.

p_{ij} are the relative frequencies calculated on the columns ($p_{ij} = n_{ij}/n_{.j}$).
The case of total lack of links can be expressed like:
$p_{i1} = p_{i2} = \cdots = p_{ic} = p_{i.}$, $i = 1, \ldots, r$, where $p_{i.} = n_{i.}/n$.
In other words, the relative frequencies calculated on the columns (almost all of them) are equal. They are also equal to the marginal relative frequencies.

The same thing can be expressed calculating the relative frequencies on the rows: $p_{1j} = p_{2j} = \cdots = p_{rj} = p_{.c}$, $j = 1, \ldots, c$, where $p_{.j} = n_{.j}/n$.

As previously said, these two formulations are alternative and equivalent.

In the case of the first of the two formulations, if there are not any links between the two qualitative characteristics, the result is:

$$p_{ij} = p_{i.} \iff n^*_{ij}/n_{.j} = n_{i.}/n \iff n^*_{ij} = (n_{i.} \cdot n_{.j})/n$$

The last equality determines the mathematic condition in order not to have any links between the studied characteristics.
With a frequency table with two entries, the formula above allows us to establish which would be the frequencies $n^*_{ij}\ (= n_{i.} \cdot n_{.j}/n)$ expected in the case of perfect independence (that is, lack of link) between the two qualitative characteristics that are studied.
The Chi-Squared test uses the formula that follows to check if the independence exists or not:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

If this quantity is large, the null hypothesis is rejected. It is possible to say that it exists a link between the two qualitative characteristics that have been analysed. In order to accept or to refuse the null hypothesis, the cut-off value is obtained by the quantiles of level α (usually equal to 0.05) of the distribution χ^2 with $(r-1)\cdot(c-1)$ degrees of freedom.

Notes.

- ⋆ This result is asymptotic, this means that it is valid for large sample size $(n > 100)$.
- ⋆ In particular, in literature it is suggested not to have expected sample sizes (n^*_{ij}) smaller than 5.
- ⋆ Certain authors accept at maximum 2 or 3 cells with expected sample sizes smaller than 5. This happens because the calculated quantity for the test, is a weighted mean with the inverse of the expected sample sizes within the cells. The smaller are that values, the smaller will be the variations that will highly contribute to the whole result of the test.

Going back to the example, the observed quantities:

		Condition of the prices		
		Flawless	Defective	Tot
Machinery	A	132	13	145
	B	99	18	117
	C	157	11	168
	Tot	388	42	430

The expected quantities:

		Condition of the prices	
		Flawless	Defective
Machinery	A	130.84	14.16
	B	105.57	11.43
	C	151.59	16.41

Test value: 6.271 $Pr\{\chi^2_2 > 6.271\} = 0.043$

When an experiment is planned, the first question is often: "**How many data do I have to collect?**"

After the test, if it results not significant, it is fundamental to answer to the question: "**With the collected sample, which probability do I have that the test results important according to certain hypothesis?**"

The first question is also called **prior power**, and the second **one posterior power**.

The analysis of the **power** and the **analysis of the sample size** are important instruments. They are able to check the ability of a statistical hypothesis test to notice when the null hypothesis is false, and to decide which dimension of the sample is required in order to have a good probability to refuse the null hypothesis when it is false.

In a generic statistical hypothesis test, the value of an interest quantity in the distribution of a population, is typically checked specifying a null hypothesis H_0 that contrasts with an alternative hypothesis H_A.

Once that the hypothesis system has been established, the statistics procedure follows the following steps:

✓ **BEFORE** Collecting the data (the sample) of the population:

1. To look for an appropriate **statistics** called **test statistic**. It is a function of data which sums up certain characteristics of the sample and which gives information on the hypothesis system. The test statistic is a random value as the selection procedure is random. Moreover, as a result it can give two different sets of values, if in the reality H_0 or H_A are true. For example, if H_0 is true, the statistics value should result small. On the contrary, if H_A is true, the statistics value should result great;
2. To evaluate **the distribution** of the test statistic when H_0 is true;

3. To set a **cut-off** value on the possible values of the test statistic. It will be able to discriminate if H_0 or H_A are true. The cut-off should be chosen so that it exists a little probability to refuse H_0 when it is correct. This is true because the test statistic represents a random value. This probability is called α.

✓ **AFTER** Having collected the data (the sample) of the population:

1. To calculate the statistics value on the selection data.
2. To determine if H_0 or H_A are true using the statistics value obtained by the data and the cut-off value.

INDEPENDENTLY of how the experimentation is conducted, it is possible to present four situations:

	The population is H_0	The population is H_A
The test says H_0	OK	**Type II Error (β)**
The test says H_A	**Type I Error (α)**	OK

If the tests refuse H_0 while H_0 is effectively true for the population, a **Type I error** is produced. The value of the probability to commit this error is called α and it is checked by the tester.

If the test accepts H_0, while H_A is the "true status" for the population, a **Type II error** is produced. The probability to commit this error is called β, and it is not directly verifiable when the test is on progress. The value $(1 - \beta)$ is called **test power** and represents the probability to correctly refuse the hypothesis H_0 when it is false.

The fundamental parameters that result from an analysis of the power are:

- ⋆ The dimension of the sample (n) or
- ⋆ The measure of the power of the test ($1 - \beta$)

And the values can be modulated one in function of the other.

These values can be estimated on the basis of the relations that exist between 5 quantities:

- ⋆ The **probability** α to commit a Type I Error. It has to be specified also **the direction of the hypothesis** H_A, i.e. if it is **unilateral** or **bilateral**;

- **The probability β** to say, the probability to erroneously refuse the alternative hypothesis H_A (this means to accept H_0) when it is true;

- The dimension of **the difference δ** between the hypothesized value and the true value of the interests quantity (or the difference **d** between the hypothesized value for the statistics and the true value of the statistics). In the case of a statistical hypothesis test on a sample mean, δ is equal to the difference between μ_0, the hypothesized mean, and μ, the true mean of the two populations. When the test is applied to two independent samples, δ is equal to the difference between μ_1 and μ_2, the two populations means.

- The **variance σ^2**, if known (only in the case of the Z-test), or an **estimation s^2** of the population variance. It is measured by a pilot study or a preliminary sample when the **true variance σ^2 is unknown**;
- **The dimension n** of the sample, seen as dimension of each subgroups of the whole sample (to say, the whole sample in the Z-test and t-tests for one sample, and each of the two subsamples in the t-tests for two dependent and independent sample).

The easiest case of the power calculation is to use the Z-distribution to determine if the population mean is equal to an hypothesized value μ_0. It is also used to see if the population mean distances itself from the hypothesized value, in an unilateral meaning, in the hypothesis where the standard deviation σ of the population is known.

In this case the statistic used is the one that follows:

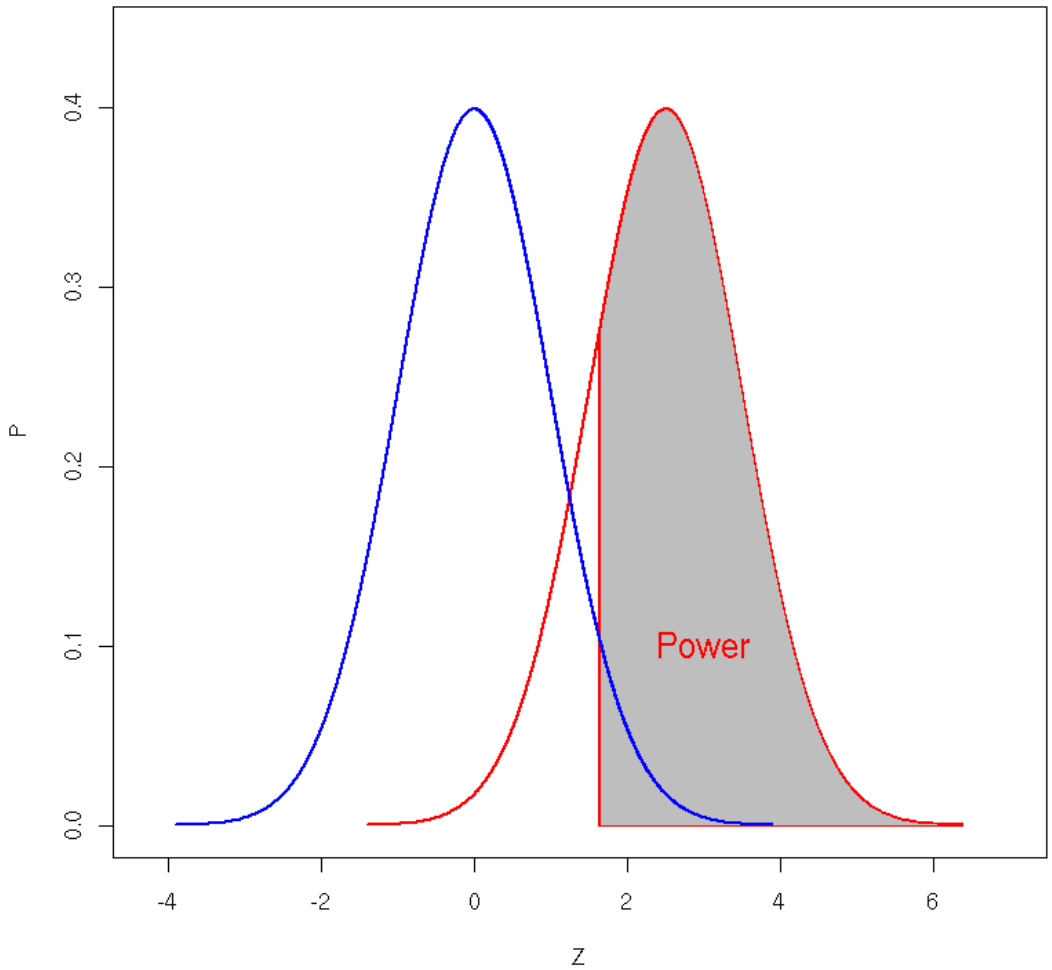
$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

The Z distribution, when the null hypothesis is true, is a normal standard $N(0, 1)$. The cut-off for the refuse of the null hypothesis is, therefore, given by the quantiles $Z_{1-\alpha}$ of the standard normal distribution.

When $\mu \neq \mu_0$, the alternative hypothesis H_A is valid and the Z-distribution is not a normal standardized as happens when the null hypothesis H_0 is true. On the contrary, it is distributed according to a normal distribution with the following mean different from zero:

$$Z \sim N\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$$

Z-distribution under the null hypothesis (on the left) and the alternative hypothesis (on the right) for the calculation of the power of the test.

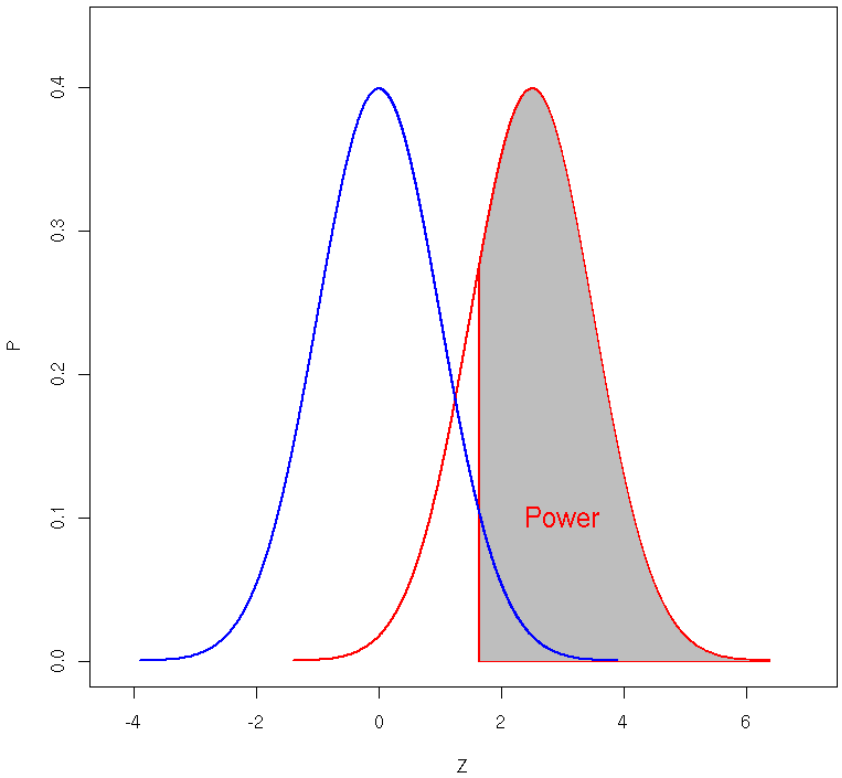


The estimation of the Z-test power happens in this way:

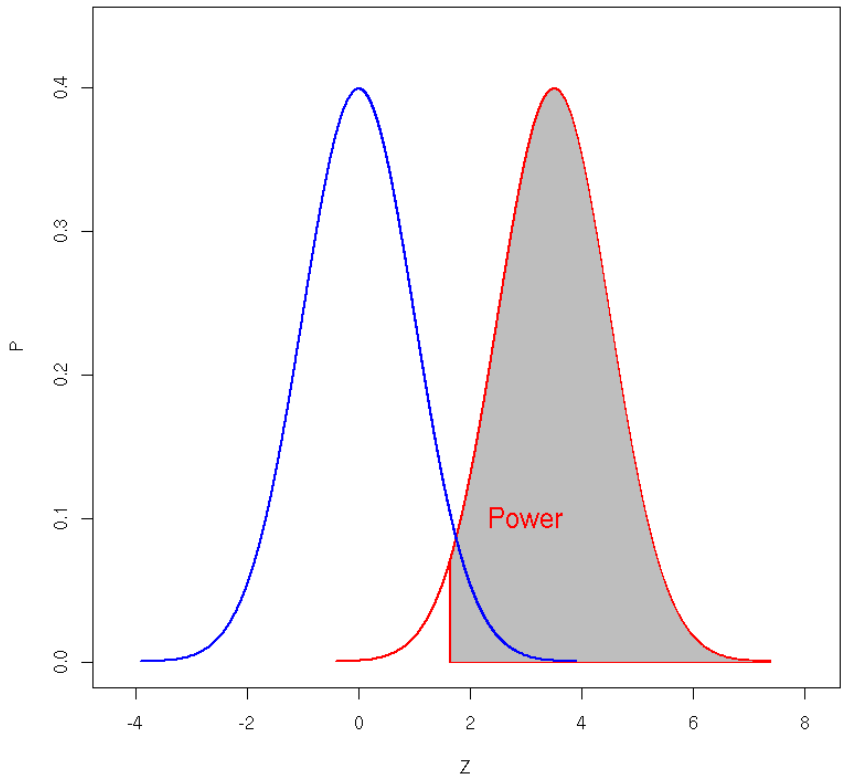
- ⋆ The decision of the cut-off value $Z_{1-\alpha}$ or the standard normal distribution, valid for the null hypothesis;
- ⋆ The calculation of the Z-distribution power valid when H_A is true, as area of this distribution not central at the right of the cut-off value $Z_{1-\alpha}$.

From the expression of Z, it is deducible that Z increases. For this reason, the test should refuse the null hypothesis when the **difference** between the true mean μ and the hypothesized mean μ_0 increases.

$\delta = 2.5$

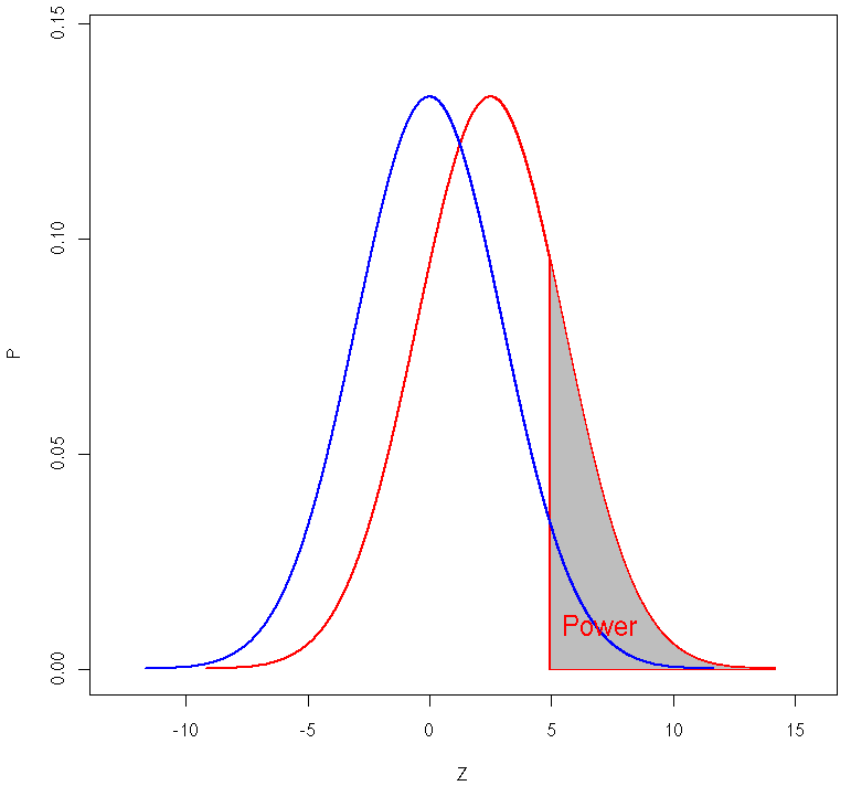


$\delta = 3.5$

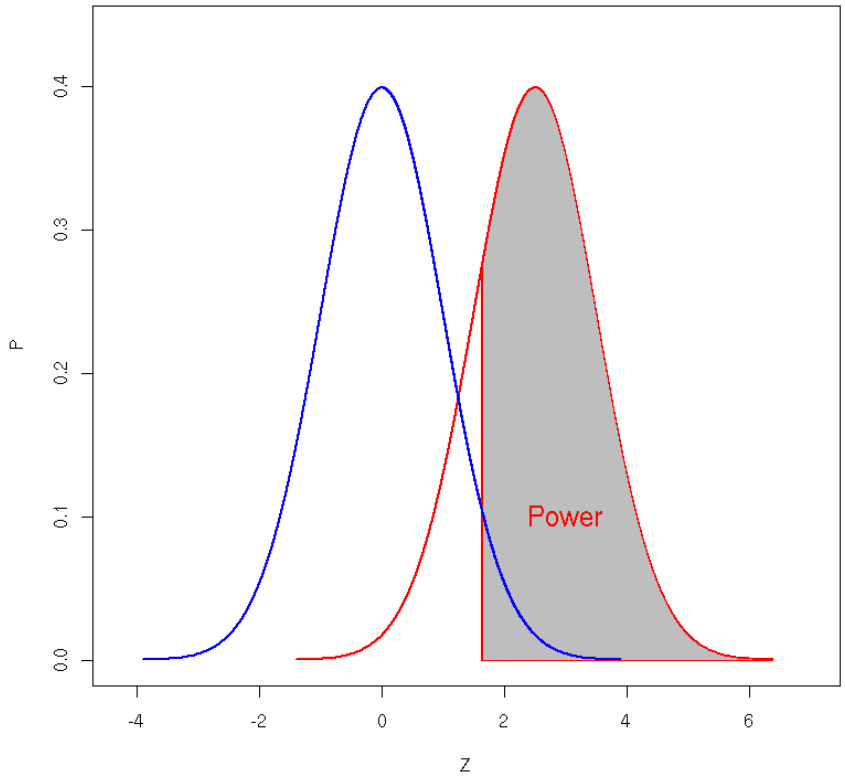


At the same time, when the value of σ decreases, or the sample size n increases, the Z-value increases. For this reason, the power of the test to highlight a certain effect of dimension δ increases.

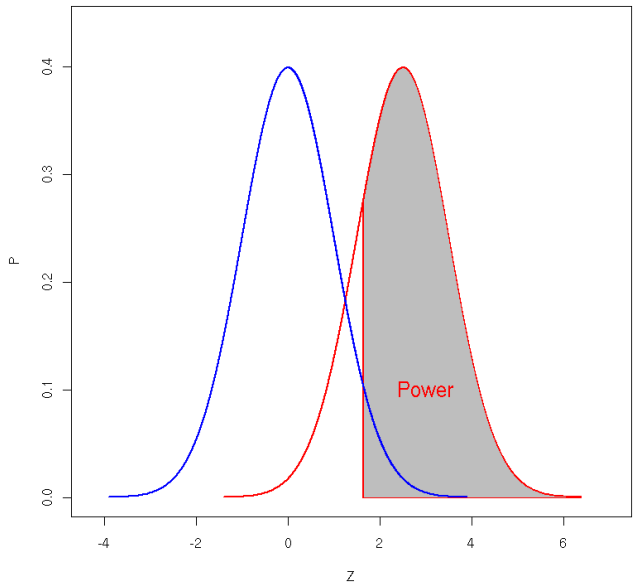
$\sigma = 3$

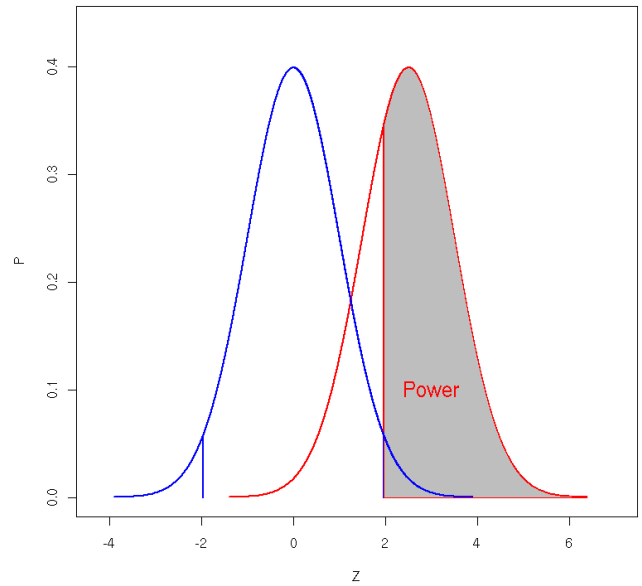


$\sigma = 1$



Another aspect that influences the power of the test is the fact that the alternative hypothesis is **unilateral** or **bilateral**. This choice modifies the refusing area of the null hypothesis. With the same level of statistical significance α, an unilateral test (picture on the left) is always more powerful of the correspondent bilateral test (picture on the right) as the critic value, upon which H_0 is refused, is smaller. The unilateral test is, without any doubts, preferable but it requires a larger quantity of preliminary information on the possible result of the test.





In the majority of the real cases, the **population variance is unknown** and the **use of s^2 instead of σ^2** requires the use of the t distribution in substitution of the Z-distribution.

When n is enough large, the Z-normal distribution and the t-distribution **are almost equal**. They provide similar answers. The differences are considered important when the sample has less than 30 observations and therefore the degrees of freedom of the t are a limited quantity.

Unlike the Z standard normal distribution, which is the unique, the **Student's t** distribution is a whole family of distributions, one for each value of degrees of freedom.

In these cases, the calculation of the three main parameters ($\boldsymbol{n}$, $\boldsymbol{\delta}$, $\boldsymbol{1-\beta}$) is based on the same concepts, but with different formulas.

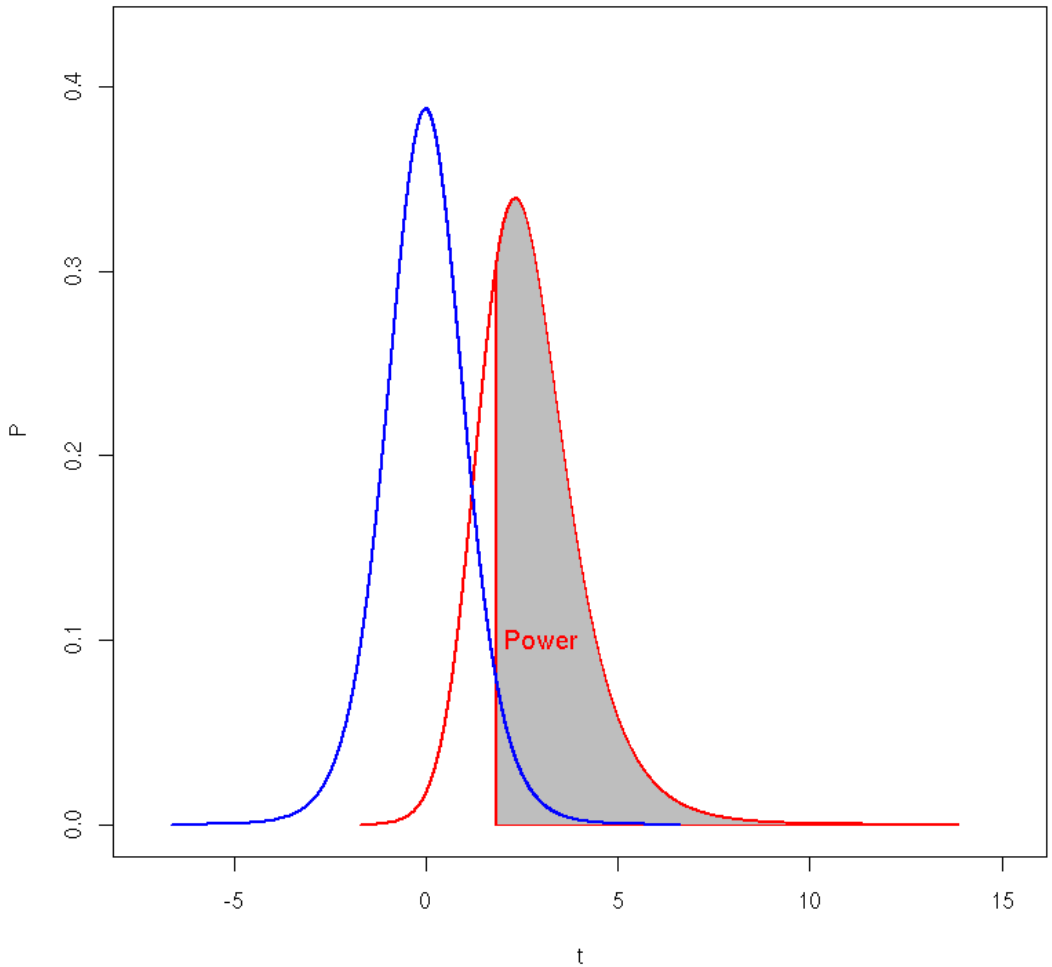
The analysis of the power and the calculation of the sample size is the same when the t-test is referred both to a **sample** and to two **dependent samples (paired)** because in this case it uses the column of the differences.

The t-test **power** is estimated:

- ⋆ Calculating the cut-off value $t_{1-\alpha,\nu}$ where ν are the degrees of freedom of the t distribution, equal to $\boldsymbol{n-1}$,
- ⋆ And then establishing the area of t distribution valid with H_A at the right of the cut-off value:

$$t_{\beta,\nu} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,\nu}$$

t-distribution under the null hypothesis (at the left) and the alternative hypothesis (at the right) for the calculation of the power of the test.



If it is necessary to establish, before collecting the data, how many observations are necessary to obtain or how many measures have to be done, for the estimation of $\boldsymbol{n}$ it is used the following formula:

$$n \geq (t_{1-\alpha,\nu} + t_{\beta,\nu})^2 \cdot \frac{s^2}{\delta^2}$$

This formula depends on the two parameters $t_{1-\alpha,\nu}$ and $t_{\beta,\nu}$, which change according to the number of degrees of freedom $\boldsymbol{\nu}$ and then (according to) the value of $\boldsymbol{n}$.

The calculation of $\boldsymbol{n}$ is done in an iterative way, starting from a first approximation (for example $n = 30$ and then $\nu = 29$). This statement makes the calculation of the quantiles $t_{1-\alpha,\nu}$ and $t_{\beta,\nu}$ possible. They are used in order to obtain a new estimation of $\boldsymbol{n}$. If the calculated value is different from the initial value, new values of t, that correspond to the new estimation of $\boldsymbol{n}$, are calculated. The value of $\boldsymbol{n}$ is then calculated again, until it is equal to the previous value (it is obtained by half up rounding $\boldsymbol{n}$ to its nearest integer because the number of observation cannot be, because of its nature, a decimal number).

In a **t-test** for two samples **not matched**, or **independent**, the difference between the means of two samples is tested.

If the two samples are **balanced**, the test power is calculated with the same procedure of the t-test for one sample. The difference is that the number of df of the t-distribution has to be multiplied for 2 ($\nu = 2n - 2$):

$$1 - \beta = P\left(t_{(2n-2,\delta)} \geq t_{(1-\alpha,2n-2}\right)$$

In this expression $t_{(1-\alpha,\nu)}$ is the quantiles of the central t distribution which corresponds to the cut-off value. $t_{(\nu,\delta)}$ is the non-central t distribution (with a non-centrality δ parameter), which is valid when the alternative hypothesis H_A is true. The probability that, with this distribution, the calculated statistics is at the right of the cut-off value, is the **power** $1 - \beta$ of the test.

When the two **samples are not balanced**, the calculation of the power happens in terms of the total sample size **N** and of the **weight** of the two independent samples (ω_1 and ω_2 proportions of the two different samples on the total sample size, with $\omega_1 + \omega_2 = 1$):

$$1 - \beta = P\left(t_{(N-2,\delta)} \geq t_{(1-\alpha,N-2)}\right)$$