Lecture 6
Statistical tests

Aim of this lecture is to describe:

- the general concept of statistical test
- the example of Chi square test
- a little bit of Kolmogorov-Smirnov test
- implementation by R.

Statistical tests

Example without theory

A railway company declares that the service has been improved and the average delay is now $\mu = 5$ Min.

For 10 days we measure the delays and observe:

$$5, 7, 4, 10, 6, 5, 8, 2, 8, 6.$$

Is the company right?

The empirical average is $\bar{x} = 6.1$.

Of course it is different from 5 (impossible to observe $\bar{x} = \mu$).

Is it too different, or could it be just a fluctuation?

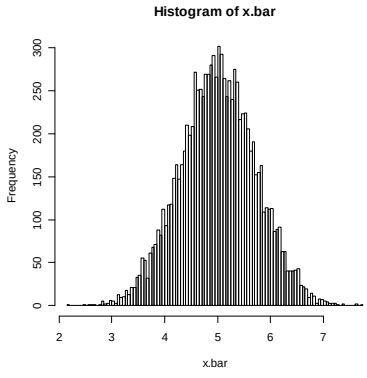
Monte Carlo idea: from a distribution having average $\mu = 5$, extract samples of cardinality 10 and compute $\bar{x}$. How typical or extreme is $\bar{x} = 6.1$?

Problem: which distribution, with average $\mu = 5$? Central Limit Theorem may help us: averages, like $\bar{x}$, tend to have (when the cardinality tends to infinity) the same type of distribution (gaussian), independently of the original distribution.

But the variance of $\bar{x}$ depends on the variance of the original distribution.

Thus, for simplicity, let us assume that the original distribution is gaussian. As variance, let us take the empirical one of the sample (it is clearly an approximation). We have $sd = 2.28$.

Now, generate 10000 samples of cardinality 10 from a $N(5, 2.28)$ and see how extreme is $\bar{x} = 6.1$.



We see that $\bar{x} = 6.1$ is quite extreme. We may also compute the *probability that a value of* $\bar{x}$ *is more extreme than* 6.1. This will be called the *p*-value. It is

$$p\text{-value} = 0.065.$$

It is quite small. However, it is larger than 0.05, one of the usual thresholds. It is up to us to

decide whether the sample is natural or not. Up to us, but with the help of this number, p-value$= 0.065$.

The components of this examples are:

- a sample
- an hypothesis ($\mu = 5$)
- a summary $\bar{x}$ of the sample, suitable for comparison with the hypothesis, called *test statistic*
- the distribution of the test statistic
- p-value, the probability that a value of the test statistic is more extreme than the observed value.

Null hypothesis

We have a sample. An assumption is made about it, like: it comes from a distribution with average 5, it comes from a Weibull with certain parameters, and so on.

Aim of a test: *reject an assumption*.

First element: the assumption, called *null hypothesis*, denoted by $\mathcal{H}_0$.

At the end of the test, either we reject $\mathcal{H}_0$, or we *cannot reject* $\mathcal{H}_0$ (we never *confirm* an assumption).

Example 1 of $\mathcal{H}_0$: the average delay is $\mu = 5$.

Example 2 of $\mathcal{H}_0$: the true density of `Piro` data is Weibull with parameters $a = 1.44$, $s = 28.25$.

Alternative hypothesis. Comparison with decision theory

The rigorous theory requires also the concept of *alternative hypothesis* $\mathcal{H}_1$. Since we do not give and prove theorems, its role will be hidden, sometimes.

Example 1 of $\mathcal{H}_1$: the true density of `Piro` data is not the previous Weibull.

Example 2, call it $\mathcal{H}_1'$: the true density of `Piro` data is the Weibull with parameters $a = 1$, $s = 30$.

The setting with both $\mathcal{H}_0$ and $\mathcal{H}_1$ looks similar to *decision theory*: based on certain observations, partial knowledge, we have to choose between two alternatives $\mathcal{H}_0$ and $\mathcal{H}_1$.

However, in decision theory, $\mathcal{H}_0$ and $\mathcal{H}_1$ are considered at the same level, the decision ends up with a choice.

In test theory the role of $\mathcal{H}_0$ and $\mathcal{H}_1$ is *not symmetric*. $\mathcal{H}_0$ can only be rejected or not rejected, but we cannot end up with a statement like: "$\mathcal{H}_0$ is true".

Example of decision theory: `Piro` density can be only of two types, Weibull with parameters $a = 1.44$, $s = 28.25$, Weibull with parameters $a = 1$, $s = 30$. We do not have a prejudice. Choose between them. This is the comparison between the fit of two specified densities, an important problem, but different from the test of a given density.

However, **test → decision**: at the end of the story of test theory, we shall compute p-values. Thus, we could compute the p-value under hypothesis $\mathcal{H}_0$ and the p-value under hypothesis $\mathcal{H}_1'$. The hypothesis with the smallest p-value is the less realistic. Therefore, a test procedure (a priori non symmetric) can be transformed in a decision procedure (symmetric).

Decisions and Bayes

We have a universe Ω, a partition (B_k) and we have to take a decision: which element of B_k is the right one?

Assume the elements B_k influence something we can observe. Assume we observe an

event A. Think to B_i as the possible causes, the input, A as the consequence, the output, the observation.

Bayesian decision rule is simply: *decide the most probable cause B_i, conditioned to the observation A.*

Bayes rule

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_k P(A|B_k)P(B_k)}$$

allows to compute the probabilities of interest. Since the denominator is the same for all i, we get the most probable B_i, conditioned to A, by maximizing the numerator:

$$B_i^{opt} := \arg\max_{B_i} P(A|B_i)P(B_i).$$

Very often we decide that a priori all the possibilities B_i are equally probable (no prejudice). Hence simply

$$B_i^{opt} := \arg\max_{B_i} P(A|B_i).$$

The probabilities $P(A|B_i)$ are precisely similar to p-values! Take as A the event "the test statistic is more extreme than the observed value", and take as B_i the different hypothesis. Therefore, deciding between alternative hypothesis by means of the corresponding p-values, is a way to perform decision theory. **Test → decision**.

Test = algorithm. Test statistics

A tests is an algorithm.

Input: the sample and some detail of $\mathcal{H}_0$.

Output the value of *test statistic*. Denote it generically by z (it was $\bar{x}$ in the example above).

Example: a politician tells us that 65% of people prefers alternative A to B. We suspect he is wrong. We ask 100 people and observe that only 47 of them prefer A to B. We need to compare the assumption H_0= "65% prefers A to B" with the sample. We need an algorithm which takes the numbers 65, 47, 100 and gives us a result. The result is a number, call it z, the test statistic. Crude example of test statistic: the relative error

$$z = \left| \frac{65 - 47}{65} \right|$$

(it is very poor: it does not take into account the cardinality of the sample; if we ask 10000 people and get 47%, there must be a difference with the case of only 100 people!).

We may think that z is random: repeat sampling. The random variable Z, the test statistic, has a probability distribution. Assume it is described by some density $f(z)$.

More precisely, *if $\mathcal{H}_0$ is true*, then Z has a certain density $f_{\mathcal{H}_0}(z)$. If an alternative $\mathcal{H}_1'$ is true, Z has another density $f_{\mathcal{H}_1'}(z)$.

How to conclude, from the value of z, whether H_0 is false or not?

Either we compute (by means of $f_{\mathcal{H}_0}(z)$) the probability that Z takes values more extreme than the observed z (this probability is the p-value).

Or we prescribe a priori a value of probability, like 5%, namely we identify a tail or two tails from the distribution $f_{\mathcal{H}_0}(z)$ with such a priori probability, and see whether the

observed z lies in such tail(s) or not.

If H_0 is true, the typical values of Z are described by $f_{\mathcal{H}_0}(z)$: they lie in the regions where $f_{\mathcal{H}_0}(z)$ is not too small. Cut one or both tails of $f_{\mathcal{H}_0}(z)$, depending on the intuitive meaning of Z. If Z belongs to such tails, we observe an unusual event, under the assumption H_0. Since it is not reasonable to observe an unusual event, we reject H_0.

To summarize: a test is based on an algorithm which computes a number z and check whether it falls into a tail of $f_{\mathcal{H}_0}(z)$. If this happens, $\mathcal{H}_0$ is rejected.

We may **either** compute the probability of the tail identified by the experimental value of z, **or** choose a priori a value like 5%, identify the corresponding tail, and see whether the experimental z falls into such tail.

Error probabilities

Assume we have developed an algorithm to make the test.

Assume we have developed an algorithm to take a decision.

In test theory we compute an *asymmetric* error probability: the probability that the algorithm rejects $\mathcal{H}_0$, when it was true. With the notations above, it is the integral of the tails of $f_{\mathcal{H}_0}(z)$.

It is called the *significance level* of the test, usually denoted by α. It must be small, like 0.05 (sometimes we say that the significance is 95%, instead of 5%).

The event "the algorithm rejects $\mathcal{H}_0$ when it is true" is called *error of first kind*.

In decision theory, on the contrary, we compute a symmetric probability: the probability that the choice made by the algorithm between $\mathcal{H}_0$ and $\mathcal{H}_1$ is wrong.

In test theory we may also compute a second type of asymmetric error probability: the probability that $\mathcal{H}_0$ is not rejected by the algorithm, when instead a particular alternative $\mathcal{H}_1'$ was true. This event is called *error of second kind*. Denote by β its probability (it must be small). The quantity

$$1 - \beta$$

is called the *power of the test*, corresponding to the particular alternative $\mathcal{H}_1'$.

Example: for our example of `Piro` data, it is not possible to compute β for $\mathcal{H}_1$, since $\mathcal{H}_1$ is too vague. One can hope to compute β for $\mathcal{H}_1'$.

Direct structure of a test procedure

- Specify $\mathcal{H}_0$
- specify $\mathcal{H}_1$ (sometimes this is hidden)
- choose α
- compute the test statistic z. The conclusion is: reject or do not reject $\mathcal{H}_0$.

We asked above: How large do we choose the tail? We choose α, and compute the corresponding tail, the corresponding quantile. Then we may implement the algorithm (we may check whether z lies in the tail or not).

Hidden in this meta-structure is the choice of the number of elements of the sample we want to use for the test: statistical tests are applied to a sample. Sometimes the sample is given a priori (like for `Piro` data), sometime else we still have to make the experiments.

More elaborate structure:

- specify $\mathcal{H}_0$ and choose α
- specify also a particular $\mathcal{H}_1'$
- choose β, compute the number of elements of the sample which may guarantee the given α and β

- perform the test.

This is the logical structure of Design Of Experiments (DOE): before sampling, we plan experiments in order to have a priori prescribed performaces (here α and β). Here we may only act on the cardinality of the sample.

p-value

Most software do not ask you α (and β). They ask you only the sample, and some detail of $\mathcal{H}_0$.

The output is a number, called *p-value*.

The p-value is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed.

Relation between *p*-value and a priori chosen significance α?

First notice that for some $\alpha \in (0,1)$ the test will end up with rejection, for other values $\alpha \in (0,1)$ it will not reject $\mathcal{H}_0$. Second, realize that $(0,1)$ is split in two parts

$$\underset{\text{no rejection}}{(0,p)} \quad \text{——} \quad \underset{\text{rejection}}{(p,1)}$$

such that for all $\alpha \in (p,1)$ the test will end up with rejection, for all $\alpha \in (0,p)$ it will not reject $\mathcal{H}_0$. The number p is the *p*-value.

Thus p is the best level of significance which produces rejection.

If p is very small, like 0.01, it means that even with such a priori prescribed significance α, the test would reject $\mathcal{H}_0$. It is a strong indication that $\mathcal{H}_0$ is false.

Chi square fit test (test di "adattamento")

Chi square distribution with *k* degrees of freedom

Definition *If $Z_1, ..., Z_k$ are independent standard Gaussians, then*

$$X^{(k)} := Z_1^2 + ... + Z_k^2$$

is Chi square with k degrees of freedom.

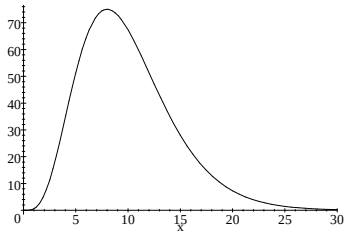
One can prove:

Theorem *The probability density of a Chi square with k degrees of freedom is*

$$f(x) = Cx^{\frac{k}{2}-1} exp\left(-\frac{x}{2}\right) \quad for\ x > 0$$

and zero for $x < 0$; $C^{-1} = 2^{k/2}\Gamma(k/2)$. Thus it is a Gamma density with

$$shape = \frac{k}{2}, \quad scale = 2.$$



$k = 10$

From the rules of mean value and variance we have

$$E[X^{(k)}] = kE[Z_1^2] = k$$
$$Var[X^{(k)}] = kVar[Z_1^2] = k\left(E[Z_1^4] - E[Z_1^2]^2\right) = 2k$$

since $E[Z_1^4] = 3$ (simple exercise). Thus the r.v.

$$\frac{X^{(k)}}{k} \text{ has mean 1 and s.d. } \frac{\sqrt{2}}{\sqrt{k}}.$$

The values of $\frac{X^{(k)}}{k}$ lie around 1, with high probability, if k is high. For instance,

$$P\left(\frac{X^{(10)}}{10} > 1.83\right) = 0.05$$
$$P\left(\frac{X^{(100)}}{100} > 1.24\right) = 0.05.$$

This fact will be the basis of Chi square test.

Asymptotically Chi square

Assume X is a discrete r.v. taking values $1, 2, ..., n_{class}$ with probabilities

$$P(X = k) = p_k$$

where $p_k \in [0,1]$, $\sum_{i=1}^{n_{class}} p_k = 1$.

Assume we have a sample of values in $\{1, 2, ..., n_{class}\}$. For every $k \in \{1, 2, ..., n_{class}\}$, denote by $\hat{p}_k$ the relative frequency of occurrence of k in the sample.

Example of the preference between A and B: code A=1, B=2,

- $n_{class} = 2$
- $p_1 = 65\%$, $p_2 = 35\%$
- $\hat{p}_1 = 47/100$, $\hat{p}_2 = 53/100$.

Compute

$$\chi^2 = \sum_{k=1}^{n_{class}} \frac{\left(\hat{p}_k - p_k\right)^2}{p_k}.$$

Theorem *The probability distribution of χ^2 converges, as $n \to \infty$, to a Chi square with $n_{class} - 1$ degrees of freedom.*

It means

$$P(\chi^2 \in [a,b]) \sim \int_a^b f_{n_{class}-1}(x)dx$$

where $f_{n_{class}-1}(x)$ is the Chi square density with $n_{class} - 1$ degrees of freedom

Chi square test

Idea: if the *empirical* distribution $(\hat{p}_k)$ differs too much from the *theoretical* distribution (p_k), χ^2 takes values which are too large with respect to the *average value* $n_{class} - 1$.

Formalization:

- $\mathcal{H}_0$ = "the sample comes from the discrete distribution (p_k)". Under this assumption, the r.v. χ^2 is roughly a Chi square with $n_{class} - 1$ degrees of freedom;

- let us choose, for instance, $\alpha = 0.05$;
- let us identify the right tail of $f_{n_{class}-1}(x)$ (we choose the right tail because of the idea above) having area α; this gives us a quantile $\lambda_{\alpha,n_{class}}$;
- compute the *test statistic* χ^2 from the sample and compare it with $\lambda_{\alpha,n_{class}}$; if $\chi^2 > \lambda_{\alpha,n_{class}}$, reject $\mathcal{H}_0$.

  For instance,

$$\text{if } n_{class} = 11\text{: reject when } \frac{\chi^2}{n_{class}-1} > 1.83$$

$$\text{if } n_{class} = 111\text{: reject when } \frac{\chi^2}{n_{class}-1} > 1.24$$

at level $\alpha = 0.05$.

This test has many applications, for instance to the so called contingency tables. One of the applications is to density fit.

Chi square test for density fit

- $\mathcal{H}_0$ = "the sample comes from the density $f(x)$".
- Split the "essential" range of $f(x)$ in intervals (a partition). This step is subjective, the result of the test may depend on it. In this step we may include our preference for an interval over another: we may give more importance to tails, for instance.
- Call $I_1, ..., I_{n_{class}}$ the number of such intervals, the *classes*. Compute

$$p_k = \int_{I_k} f(x)dx, \quad k = 1, 2, ..., n_{class}.$$

  These are the theoretical probabilities.
- Compute the empirical probabilities $\widehat{p}_k$ as it is done in an histogram: $\widehat{p}_k$ is the relative occurrence in I_k of the sample.
- Apply the test (either by hands, choosing α, computing $\lambda_{\alpha,n_{class}}$, χ^2 and comparing them; or by `R`, as below, getting the p-values as outcome).

Example

Let us apply Chi square test to `Piro` data, using R. The procedure is a little bit tricky, because we have to:

- define the intervals (this requires some work)
- compute empirical frequencies
- compute theoretical frequencies
- test.

If we analyze `Piro` data without outlier, choose 5 classes, assume ML Weibull, we get

$$p - value = 0.938.$$

This means that we cannot reject the assumption, even very strongly. Strong indication that the fit is very good.

We may check that the result depends on the number of classes. However, it is always good.

We may check that the result is still good with little change of parameters, like $a = 1$, $s = 30$. This shows that the end of a test cannot be: *we confirm* $\mathcal{H}_0$. Indeed, several slightly different $\mathcal{H}_0$ are not rejected.

However, if we try $a = 3$, $s = 40$, we get $p = 1.794e - 05$. This Weibull is rejected.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test makes a comparison between theoretical and empirical cdf. The test statistic is usually denoted by *D*.

For details, we suggest to read texts or the notes “generalità sui test statistici; test di Kolmogorov-Smirnov”.

The idea is to compute the supremum of the difference between the two cdf. In this sense, it is similar to the “distance” between cdf introduced in Lecture 3, but with the supremum norm instead of the integral norm.

With R, simply write

`ks.test(Dati, "pweibull", a, s)`.

The output is the value *D* of the Kolmogorov-Smirnov test statistics, and the *p*-value. The null hypothesis $\mathcal{H}_0$ is that Weibull with parameters *a*, *s* is the correct distribution. A very small *p*-value thus means that we strongly reject this assumption. A moderate or large *p*-value means that such Weibull is reasonable for that sample.

It does not like repeated values. `Piro` data have repeated values. Perturb them with infinitesimal numbers and try again.

Applied to `Piro` data without outlier, and assumption Weibull with parameters $a = 1.44$, $s = 28.25$, it gives

$$p = 0.87$$

so the assumption looks very reasonable.

If we try with the outlier, we get

$$p = 0.57$$

which is still good. Weibull ML fit including the outlier is still reasonable. But if we try, with outlier, ML log-normal, we get

$$p = 0.97$$

which is better. Again a confirmation that log-normal fit should be better than Weibull, for the full `Piro` set. This is an example of application of *test theory as a decision theory*.

If we try with different assumptions, like $a = 1$, $s = 30$, we still get good values. However, if we try $a = 3$, $s = 40$, we get $p = 0.007$. This Weibull is rejected.