Hypothesis Testing

L. Egidi

Fall 2021

University of Trieste

Fundamentals of hypothesis testing

Some commonly used tests

Relation between tests and confidence intervals

Nonparametric tests

Fundamentals of hypothesis testing

The idea of hypothesis testing

The basic aim of hypothesis testing within a *parametric statistical model* $f_{\boldsymbol{\theta}}(\mathbf{y})$ is **to establish whether the data could be reasonably be generated from** $f_{\boldsymbol{\theta}_0}(\mathbf{y})$, where $\boldsymbol{\theta}_0$ is a specific value of the parameter.

This is simply denoted by the succinct notation

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \,,$$

with H_0 being termed **null hypothesis**.

Complementary to the choice of H_0, it is required to select a complementary **alternative hypothesis** H_1, specifying the values of the parameter which become reasonable when H_0 does not hold.

Example: testing the mean of a normal sample

Assume the very simple model for independent observations $y_1, y_2, \ldots, y_n$ given by $Y_i \sim \mathcal{N}(\mu, 1)$. Then we may want to test

$$H_0 : \mu = 0$$

against

$$H_1 : \mu > 0$$

which amounts to testing the null hypothesis of data generated from a standard normal distribution, against the possibility that the true mean takes instead a positive value.

This choice of H_1 makes fully sense when we can rule out negative values of μ (**one-sided alternative**). If this is not the case, a better choice would be given by $H_1 : \mu \neq 0$ (**two-sided alternative**).

General formulation

In broad generality, hypothesis on a parameter $\boldsymbol{\theta}$ can be cast in the form

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$$

against

$$H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$$

where $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ form a bi-partition of the set containing all the possible values for the parameter $\boldsymbol{\theta}$, that is named the **parameter space $\boldsymbol{\Theta}$**.

The tools for addressing problems of such level of generality will be covered in the part of the course devoted to *likelihood methods*.

In what follows, instead, we will illustrate the main ideas by means of simple, yet important, instances.

Steps of hypothesis testing

The theory of hypothesis testing is rather articulated, so that it may help to go through the main parts of the theory in a systematic fashion.

Some noteworthy concepts are

- Test statistic
- Null and alternative distributions
- *p*-value
- Significance level, rejection and acceptance regions
- Errors and power

Test statistic

A **test statistic** is a statistic (namely, a function of the r.v. representing the available sample) which is used to carry out the test.

Large values (in absolute value) of the test statistic cast doubt on H_0 and on the theory underlying it.

Its choice depends on the problem under study. For the simple normal example mentioned above, a natural choice is to take as test statistic the (standardized) sample mean

$$Z = \frac{\overline{Y}}{\sqrt{\frac{1}{n}}} = \sqrt{n}\,\overline{Y}$$

Null and alternative distributions

The distribution of a test statistic will generally depend on the true value of the parameter under testing.

In the example, if H_0 is true (*under* H_0), then

$$Z \sim \mathcal{N}(0, 1)\,,$$

and this is called the **null distribution** of Z.

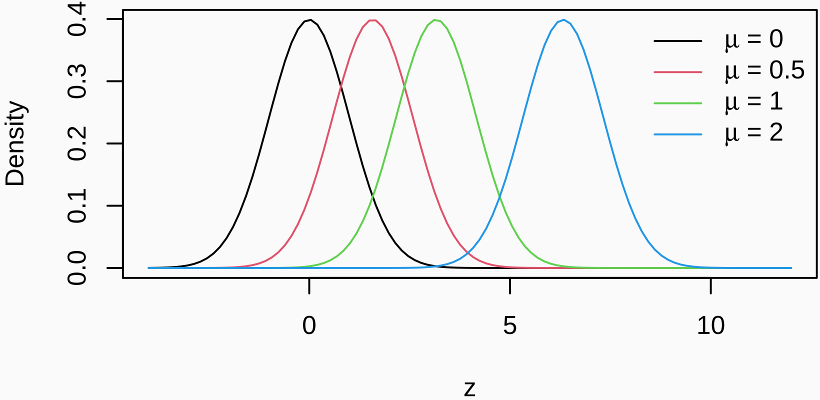
Instead, if H_1 holds (*under* H_1), it follows that

$$Z \sim \mathcal{N}(\Delta, 1)$$

where $\Delta = \sqrt{n}\,\mu > 0$ increases with the value of μ.

The distributions valid under H_1 are called the **alternative distributions** of Z.

R lab: visualizing the null and alternative distributions



The p-value

The p-value measures the distance between the data and H_0. Small values of it correspond to a test statistic unlikely to arise under H_0, and suggest that H_0 and the data are inconsistent.

In the example, the idea is that any value larger than the observed z_{obs} (the value of Z computed with the observed data) would cast even greater doubt on H_0.

The p-value is thus defined as *the probability (under H_0) of observing a value of the test statistic equal or larger than the observed one*

$$p = \Pr_{H_0}(Z \geq z_{obs})$$

Since under H_0 we have $Z \sim \mathcal{N}(0, 1)$, it follows that

$$p = 1 - \Phi(z_{obs})$$

R lab: computing the *p*-value for a sample

In case the null distribution is not known, it would be possible to compute the *p*-value by simulation whenever it is possible to generate data under H_0. In R:

```
set.seed(13); n <- 10; y_obs <- rnorm(n)
z_obs <- mean(y_obs) * sqrt(n)
print(z_obs)
```

```
## [1] 1.897537
```

```
M <- 100000; z_sim <- numeric(M)
for(i in 1:M) { y <- rnorm(n)
                z_sim[i] <- mean(y) * sqrt(n) }
c(mean(z_sim >= z_obs), 1 - pnorm(z_obs))
```

```
## [1] 0.02877000 0.02887856
```

Other alternative hypotheses: more details

For the simple example of test on μ and the same $H_0 : \mu = 0$, other two possibilities for H_1 could then be considered.

In either case, the same test statistic Z would still be used, but the computation of the p-value would change, due to the different direction of deviation from H_0.

For $H_1 : \mu < 0$, small values of Z would flag deviation from H_0 (that is, negative values with large absolute value), so that

$$p = \mathrm{Pr}_{H_0}(Z \leq z_{obs}) = \Phi(z_{obs})\,.$$

Instead, for $H_1 : \mu \neq 0$, both directions ought to be considered, and

$$p = \mathrm{Pr}_{H_0}(|Z| \geq |z_{obs}|) = 2\,\mathrm{Pr}_{H_0}(Z \geq |z_{obs}|) = 2\,\{1 - \Phi(|z_{obs}|)\}\,.$$

Significance level

We commonly say that a the result of a test is *significant at the 5% level* whenever the p-value is smaller or equal to 0.05. Other levels of some practical interest are 1% or 0.1%.

As stated in the CS book, an often-followed convention is

Range	Evidence against the null hypothesis
$0.05 < p \leq 0.1$	*marginal evidence*
$0.01 < p \leq 0.05$	*evidence*
$0.001 < p \leq 0.01$	*strong evidence*
$p \leq 0.001$	*very strong evidence*

A test *with fixed significance level* arises when the significance level is fixed in advance, and then it is just reported whether the p-value is smaller than the fixed level. If this happens, it may be reported that H_0 **is rejected**, otherwise we may say that H_0 **is not rejected** (or **accepted**).

Rejection and acceptance regions

If we define **the sample space** as the set of the values that our available sample may take, the **rejection region** of a test with fixed significance level is the subset of the sample space corresponding to the samples that would lead to a rejection of H_0.

The remaining part of the sample space forms instead the **acceptance region**.

Both these two regions are determined by means of a test statistic.

Rejection and acceptance regions for the example

In the simple normal example previously introduced, for $H_1 : \mu > 0$, it is simple to verify that a rejection region of level α is simply

$$\mathcal{R}_\alpha = \{\mathbf{y} : Z \geq z_{1-\alpha}\},$$

where $z_{1-\alpha}$ is the standard normal $(1-\alpha)$-quantile, i.e. 1.645 for $\alpha = 0.05$.

The acceptance region is just given by

$$\mathcal{A}_\alpha = \{\mathbf{y} : Z < z_{1-\alpha}\}.$$

(Note: the computation of the p-value, and of $\mathcal{R}_\alpha$ and $\mathcal{A}_\alpha$ would be exactly the same if the null hypothesis were of the form $H_0 : \mu \leq 0$, maintaining the same alternative hypothesis.)

Errors for a fixed-significance level test

When we adopt a test with fixed significance level, we move from using the *p*-value as a measure of evidence against H_0 to using a test to decide which of H_0 and H_1 is more supported by the data.

Two wrong decisions are possible. We commit a *Type I error* by rejecting H_0 when it is true, or a *Type II error* by accepting H_0 when it is false.

In the example, $\mathrm{Pr}_{H_0}(\mathbf{Y} \in \mathcal{R}_\alpha) = \alpha$, and in fact **the fixed significance level equals the probability of making a Type I error**.

Power of a test

For a test with fixed significance level, the power is the probability of (correctly) detecting that H_0 is false

$$\mathrm{Pr}_{H_1}(\mathbf{Y} \in \mathcal{R}_\alpha)\,.$$

The power of a test can be used for comparing alternative tests for the same problem, with tests with higher power being preferable.

The power is often used for designing studies, in particular for choosing the sample size in medical or industrial studies. Indeed, for fixed significance level, the power increases with the sample size.

Power of two tests for the example

For the simple example (with $H_1 : \mu > 0$), an alternative (but silly) test statistic may be given by taking the same Z as above computed by using only half of the sample (for n even).

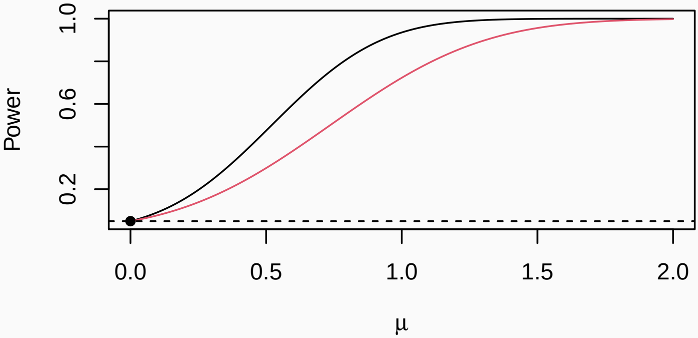
Fixing a significance level of 5%, the two tests have exactly the same probability of a Type I error, so for comparing them we must use their power.

The power is a function of the μ assumed under H_1, and for a certain $\mu \geq 0$ we obtain (since $z_{0.95} = 1.645$)

$$\mathrm{Pr}_\mu(Z \geq 1.645) = 1 - \Phi(1.645 - \sqrt{n}\,\mu)$$

R lab: power of two alternative tests

```r
mu <- seq(0, 2 , l = 1000); n <- 10; n1 <- 5
plot(mu, 1 - pnorm(1.645 - sqrt(n) * mu), type = "l",
     ylab="Power", xlab = expression(mu))
lines(mu, 1 - pnorm(1.645 - sqrt(n1) * mu), col = 2)
abline(h=0.05, lty = 2); points(0, 0.05, pch = 16)
```



Comments on the *p*-value

The usage of *p*-values is not free of controversies, and in ending the review of the general theory on testing some comments are in order.

1. The *p*-value **is NOT the probability that H_0 is true**, since the latter is not even an event.
2. It is a **continuous measure** that is usually dichotomized (lower or greater than a fixed treshold) by human subjectivity.
3. The results of statistical tests, and *p*-values in particular, should never be taken without considering context-specific knowledge. Even a small *p*-value may not be particularly meaningful if the alternative hypothesis is logically implausible.
4. Hypothesis testing is useful in certain contexts, but it has some important limitations. For (very) large sample sizes, even tiny deviations from the null hypothesis will lead to small *p*-values. For large sample sizes, there are alternative approaches which are more fruitful, and techniques based on **model selection** are often preferable to statistical tests.

Some commonly used tests

One-sample t test

Given a normal random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, a classical testing problem on μ is of the form (for two-sided alternative, say)

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

The test statistic is given by

$$T = \frac{\overline{Y} - \mu_0}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1}\,, \qquad \text{when } H_0 \text{ is true}$$

with the p-value given by

$$p = \Pr_{H_0}(|T| \geq |t_{obs}|)$$

which can be computed as $p = 2\Pr_{H_0}(T \geq |t_{obs}|) = 2\,\{1 - F_{t_{n-1}}(|t_{obs}|)\}$, since the t distribution is symmetric around 0.

Example

The DAAG book introduces the simple dataset `pair65`, about an experiment on the effect of heat on the stretchiness of elastic bands: a small sample of differences between two different conditions for 9 bands.

heated	ambient	difference
244	225	19
255	247	8
253	249	4
254	253	1
251	245	6
269	259	10
248	242	6
252	255	-3
292	286	6

Example (cont'd)

Focusing on the 9 differences on the amount of stretch, we test

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

by means of the `t.test` function, resulting in significance at 5% level

```
##
##  One Sample t-test
##
## data:  difference
## t = 3.1131, df = 8, p-value = 0.01438
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   1.641939 11.024728
## sample estimates:
## mean of x
##  6.333333
```

Approximate tests

For large random samples, the Central Limit Theorem ensures that $\overline{Y} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, being $\mu = E(Yi)$ and $\sigma^2 = \mathrm{var}(Y_i)$.

A test statistic for $H_0 : \mu = \mu_0$ is therefore

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \dot{\sim} \mathcal{N}(0,1)\,, \qquad \text{when } H_0 \text{ is true}$$

The estimator of the variance S^2 can be replaced by a more suitable one. For example, for binary data, $Y_i \sim \mathcal{B}_i(1, \pi)$, commonly used test statistics are $Z = \frac{\widehat{\pi} - \pi_0}{\sqrt{\frac{\widehat{\pi}\,(1-\widehat{\pi})}{n}}}$ or $Z = \frac{\widehat{\pi} - \pi_0}{\sqrt{\frac{\pi_0\,(1-\pi_0)}{n}}}$, the latter being preferable.

Tests based on the CLT are instances of **approximate tests**, for which the property concerning the Type I error level holds only approximately.

Two sample t-test

Given two **independent normal samples**, represented by $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)\,, i = 1, \dots, n_X$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)\,, i = 1, \dots, n_Y$, the test statistic for testing the equality between the two means is

$$T = \frac{\overline{X} - \overline{Y}}{\mathrm{SE}(\overline{X} - \overline{Y})}$$

with $\mathrm{SE}(\overline{X} - \overline{Y})$ estimated by $\sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}$.

A different formula is instead adopted is if it is possible to assume that $\sigma_X^2 = \sigma_Y^2$.

The distribution of T when H_0 is true is given by a suitable t distribution.

Like for the one-sample case, there are general formulas for large samples, employing the normal distribution.

Paired t-test

Paired observations arise whenever each unit of a random sample of size n is observed twice, under different conditions, so that we end up again with two set of variables $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)\,, i = 1, \ldots, n$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)\,, i = 1, \ldots, n$.

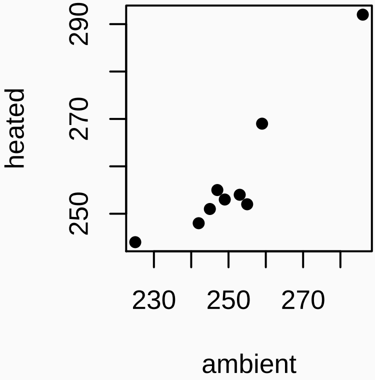
However, now the pair (X_i, Y_i) refers to the same unit, so that the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ **are no longer independent**.

The `pair65` data set is exactly of this nature. Like in that example, the resolution is to focus on the random sample of the n differences $D_i = X_i - Y_i$, for which $E(D_i) = \mu_X - \mu_Y$: for testing the equality of the two means μ_X and μ_Y we just apply the theory for the one-sample t-test, with $\mu_0 = 0$.

For the `pair65` data set, the p-value of about 0.014 suggests that heat may indeed have an effect on stretchiness.

Example

Even though the `pair65` data is very small, the fact that the two groups of observations are not independent is readily suggested by a scatterplot



By (blindly) applying the test for independent data we would get a *p*-value of about 0.40, hinting at a quite different conclusion.

Relation between tests and confidence intervals

Main result

As displayed for the `pair65` data testing, the `t.test` R function returns also the confidence interval for the parameter under testing, in that case the true mean of the differences in stretchiness.

This is not by chance, since there is a close connection between hypothesis testing on the value of a certain parameter and confidence intervals for that parameter.

For the case of a mean, for example, the basic idea is that

> *If the confidence interval for μ does not contain zero, this is equivalent to rejection of the hypothesis that the true mean is zero.*

Important: the connection is between two-sided confidence intervals and two-sided alternative hypotheses. For one-sided alternative hypotheses, the connection is with one-sided confidence intervals.

More precisely

The general result is as follows, and states a perfect equivalence between the two methods:

1. Given a method to find a confidence interval of level $(1-\alpha)\%$ for a certain scalar parameter θ, we can establish whether the *p*-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is smaller than the significance level α by checking if θ_0 is included in the interval
2. Given a method to find a *p*-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can obtain a confidence interval of level $1-\alpha$ by selecting all the θ_0 values that will lead to a *p*-value larger than α

Example: `pair65` data

The 95% and 99% confidence intervals for the mean of the differences are, respectively

95%	1.6419	11.0247
99%	-0.4930	13.1596
98.56217%	0.0000	12.6667

The 95% confidence interval does not contain zero, while the wider 99% does, implying that the hypothesis $\mu = 0$ is rejected for $\alpha = 0.05$, but not for $\alpha = 0.01$.

Note that for a confidence interval of level $1 - p = 0.9856217$, we obtain a lower limit exactly equal to 0: the p-value, in fact, corresponds to a significance level which is borderline between rejection and non-rejection of H_0.

Nonparametric tests

Main idea behind nonparametric tests

Nonparametric tests specify only partially a statistical model for the data, so that they may provide more robust inferences than parametric tests with contaminated data, outliers or, more generally, in settings where model specification is hard.

This is sometimes useful, especially when only certain aspects of the data are of interest, or for checking the results obtained with a full model specification.

The details of such tests, and more generally the theory supporting their validity, would require a substantial amount of space. Here we just mention such solutions in passing, as a tool in the statistician's reservoir that at times may be a useful complement to parametric tests.

Wilcoxon rank sum and signed rank tests

The main idea of nonparametric tests is illustrated by the Wilcoxon rank sum test, which can be used to replace the *t* test when normality is doubtful, due to outliers or excessive rounding, for example.

The test uses the **ranks**, which are the index of each observation in the sample sorted in ascending order. For instance, for the `pair65` set of differences

difference	rank
19	9
8	7
4	3
1	2
6	5
10	8
6	5
-3	1
6	5

Wilcoxon rank sum and signed rank tests

In the example, the R function `wilcox.test` returns a *p*-value of 0.017, which is very similar to what returned by the parametric test, thus reinforcing the conclusion.

There are also two-sample extensions, for both independent data or paired data (though the latter can be performed by considering the differences, as done here). The two-sample version (for independent samples) is known as *signed rank test* or *Mann-Whitney test*.