Hypothesis Testing

N. Torelli, G. Di Credico, V. Gioia

Fall 2024

University of Trieste

Fundamentals of hypothesis testing[1]

Some commonly used tests[2]

Relation between tests and confidence intervals[3]

Nonparametric tests[4]

---

[1] Agresti, Kateri: sec 5.1-5.5
[2] Agresti, Kateri: sec 5.2-5.3-5.4
[3] Agresti, Kateri: sec 5.6
[4] Agresti, Kateri: sec 5.8

Fundamentals of hypothesis testing

The idea of hypothesis testing

THE HYPOTHESIS CAN BE SIMPLE $H_0: \theta = \theta_0$ OR COMPOSITE $H_0: \theta \leq \theta_0$

The basic aim of hypothesis testing within a *parametric statistical model* $f_\theta(\mathbf{y})$ is **to establish whether the data could be reasonably be generated from** $f_{\theta_0}(\mathbf{y})$, where θ_0 is a specific value of the parameter.

This is simply denoted by the succinct notation

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

$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases}$ TWO SIDES $\quad \theta \gtrless \theta_0$ ONE SIDES

with H_0 being termed **null hypothesis**. => NO EFFECT CASE

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.

WE NEED TO CREAT A SYSTEM WITH VALUE THAT DON'T OVERLAP:

THIS 2 ARE NO GOOD
$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{cases}$

THIS ONE YES
$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases}$

WE ASSUME A PARAMETRIC MODEL: $Y_i \overset{iid}{\sim} \mathcal{N}(\mu, 1)$

WE WANT TO PERFORM HYPOTHESIS TESTING ON μ:

$$\begin{cases} H_0: \mu = \mu_0 = 0 \\ H_1: \mu > 0 \end{cases}$$

WE WANT ALWAYSE TO MAKE HYPOTHESIS ON A PARAMETER, NOT AN ESTIMATOR

$(\mu \surd, \hat{\mu} \times)$

$\mu_0 = 0, \mu_0 > 0$

$\Rightarrow$ THE LOGIC IS:

- SPECIFY A MODEL WITH AN UNKNOWN PARAMETER
- MAKING ASSUMTION ON THE UNKNOWN PARAMETER

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 θ can be cast in the form

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

against

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

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

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 (A R.V. WITH HIS OWN KNOWN OR UNKNOWN DISTRIBUTION)
- Null and alternative distributions (LINK TO HYPOTHESIS TESTING WERE NULL DISTRIBUTION OR THE OTHER ONE IS ASSUMED TRUE)
- p-value
- Significance level, rejection and acceptance regions
- Errors and power (IF THE DECISIONS TAKEN ARE GOOD OR NOT)

(METHOD FOR PERFORM HYPOTHESIS TESTING AND FIND EVIDENCE FROM THE DATA IN FAVOUR OF AN HYPOTHESIS. THEY ARE MADE TO TAKE DECISIONS)

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.

IN SIMPLE CASES THEY ARE THE PIVOTAL QUANTITIES

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

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

$\overline{Y} \sim \mathcal{N}(\mu, 1/n)$ STARTING POINT TO COSTRUCT THE TEST STATISTIC => NEXT, WE HAVE TO STANDARDIZE IT

$\sqrt{n}(\overline{Y}-\mu) \sim \mathcal{N}(0,1)$ $\boxed{Z} = \sqrt{n}\,\overline{Y} \sim (\sqrt{n}\mu, 1)$

↳ TEST STATISTIC

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

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

ANOTHER NOTATION:
$Z|H_0 \sim \mathcal{N}(0,1)$
$Z|H_1 \sim \mathcal{N}(\sqrt{n}\mu, 1)$

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

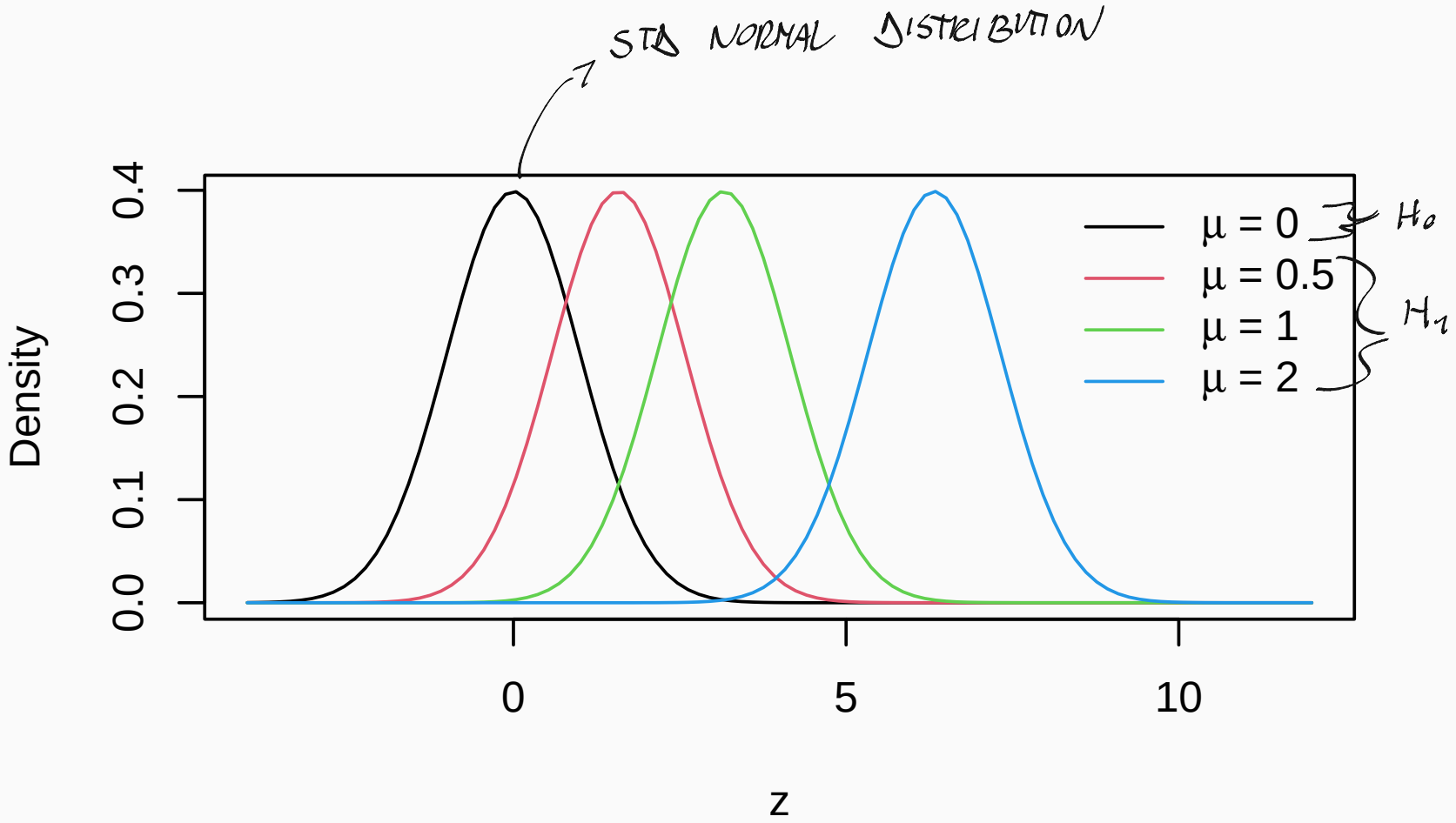
Instead, if H_1 holds (*under* H_1), it follows that

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

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

IT'S THE PROBABILITY OF OBSEARVING A VALUE FROM YOUR DATA THAT'S IS MORE EXREME THAN THE ONE YOU HAVE OBSERVED, ASSUMING THAT THE NULL HYPOTHESIS IS TRUE

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.

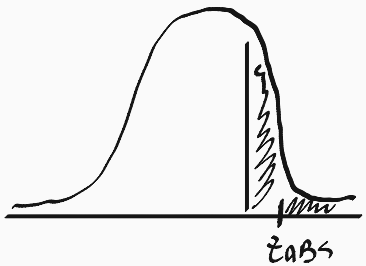
The p-value is thus defined as *the probability (under H_0) of observing a value of the test statistic equal or larger than the observed one*

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

$$z_{obs} = \frac{\bar{y} - \mu_0^{\,0}}{\sigma/\sqrt{n}} \to 1$$

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

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



THIS CAN NOT BE μ

z FOLLOW THE STD NORMAL DISTRIBUTION WE COMPERE THE VALUE WITH THE PLOT

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
```

THE SIMULATED IS VERY CLOSE

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

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

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

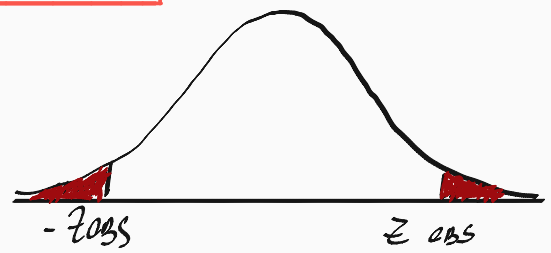
$$p = \Pr_{H_0}(|Z| \geq |z_{obs}|) = 2 \Pr_{H_0}(Z \geq |z_{obs}|) = 2 \{1 - \Phi(|z_{obs}|)\} \, .$$

$2(1 - P(Z < |z_{obs}|))$

z VALUE

QUANTILE

SIMMETRY



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*

THINK A BIT DON'T JUMP TO CONCLUSION

THE P-VALUE IS NOT MEANT FOR TAKE COICNE BUT SUPPORT THE CONCLUSION. IS JUST A MEASURE

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.

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

$Y_i \sim \mathcal{N}(\mu, 1)$

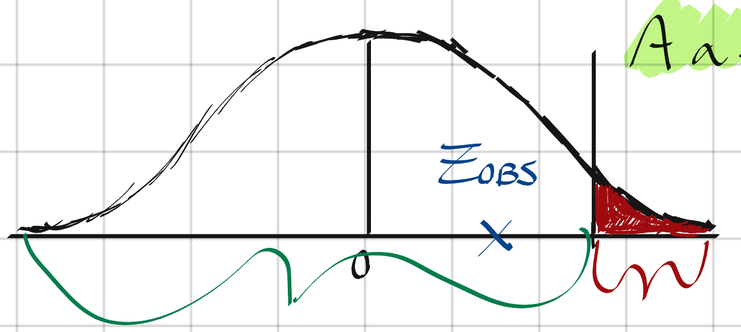
$Z = \sqrt{n}\,\bar{Y} \sim \mathcal{N}(\sqrt{n}\,\mu, 1)$

- $Z_{H_0} \sim \mathcal{N}(0,1)$

- $\alpha = \begin{cases} 0.01 \\ 0.05 \\ 0.1 \end{cases}$

$R_\alpha = \{ y : z > z_{1-\alpha} \}$

$A_\alpha = \{ y : z < z_{1-\alpha} \}$



ACC. SPACE REG. SPACE

THE INGREDIENTS WE NEED ARE:

- THE HYPOTHESIS
- THE TEST STATISTIC
- THE LEVEL α

SUPPOSE WE HAVE A MEAN $\mu = 0.336$ AND WE MULTIPLY IT WITH $\sqrt{28}$ ($\sqrt{n}$) AND OBTAIN AN OBSERVED VALUE OF $z_{obs} = \mu\sqrt{28} = 1.78$.
THE QUANTILE OF $1-\alpha$, WITH $\alpha = 0.05$ GIVE US 1.64.

WHAT CAN YOU SAY? YOU REJECT THE NULL HYPOTHESIS.

IF WE SELECT A SMALLER VALUE FOR α : $1-\alpha = 0.99$.
THEN THE QUANTILE BECAME 2.326 AND WE CAN'T REJECT THE NULL HYPOTHESIS.

SO, α IS THE PROBABILITY, ASSUMING H_0 TRUE, THAT Y IS IN THE REJECTION REGION.

WE ARE SAYING THAT WE WANT TO FIX THE PROBABILITY OF REJECTING H_0. SO, IF WE REJECT α WE ARE MAKING AN ERROR.

α IS CALLED TYPE I ERROR PROBABILITY: WE REJECT H_0 EVEN IF IS TRUE.

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

IT MEANS THAT IF YOU REPEAT THE SAMPLE A VERY LARGE NUMBER OF TIME, $1-\alpha$ % WILL BE OK, BUT THE α% WILL HAVE AN ERROR.

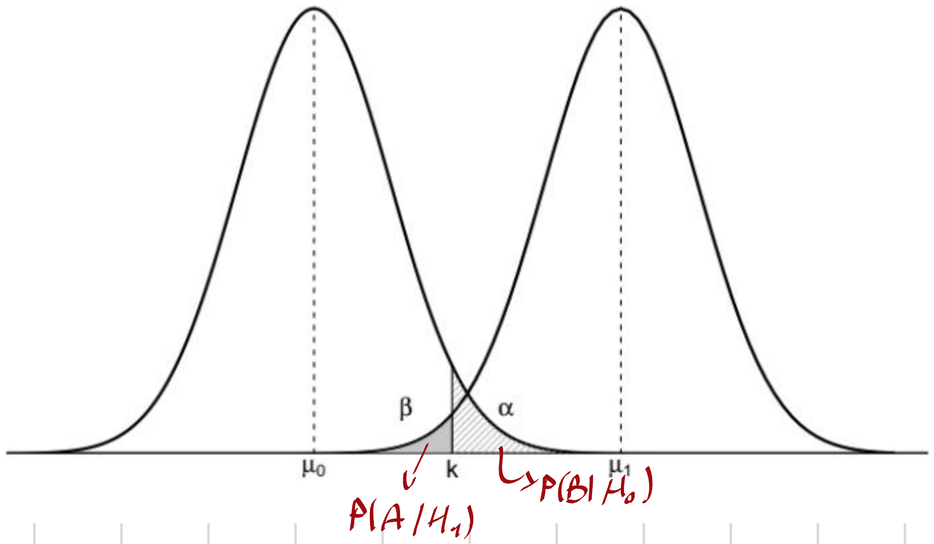
	H_0 TRUE	H_0 FALSE
A	$1-\alpha$	β
R	α	$1-\beta$

TYPE I ERROR: REJECTING A TRUE H_0
TYPE II ERROR: ACCEPT A FALSE H_0
POWER OF THE TEST: IT IDENTIFIES WHAT THE PROBABILITY OF CORRECTLY REJECTING H_0

LET'S MAKE α AND β INTERACT

WE ALWAYSE ASSUME THE SAME MODEL BUT A SIMPLE HYPOTHESIS:

OUR GUESS FALL UNDER THE NULL HYPOTHESIS UNDER THE STD NORMA DISTRIBUTION



IF $\alpha \to 0$, THE QUINTILE GOES $+\infty$ AN β BECOME THE ENTIRE AREA (1) AND FOR SURE HAVE TYPE II ERROR.
BUT, IF $\beta \to$, IT HAPPENS THAT THE QUANTILE GOES $-\infty$ AND $\alpha = 1$: TYPE I ERROR

TRYING TO REDUCE THE PROBABILITY OF I TYPE INCREASE THE PROBABILITY OF II TYPE.

WE HAVE TO FIND A BALANCE BY SETING α AND COMPARE THE TESTS TO THEIR POWER. THE HIGHER THE POWER, THE BETER. BECAUSE IT MEANS THAT YOU ARE GOING TO TAKE THE RIGHT DECISION TO REJECT H_0 ASSUMING IT FALSE AMONGES ALL THE POSSIBLE TESTS.

SO, THE FUNCTION DEPENDS ON: . α
. THE HYPOTESIS
. N

} 3 QUANTITIE THAT HAVE AN IMPACT ON TYPE I ERROR

CONCLUSION: . MINIMIZE THE PROBABILITY OF COMMITING A TYPE I ERROR IS EQUAL TO INCREASE THE INTERVALL, REDUCING R_α
. MAXIMIZE THE POWER OF THE TEST IS EQUAL TO MINIMIZE THE ERROR OF THE TYPE II, INCREASING R_α

SO, IT'S NOT POSSIBLE TO MINIMIZE α AND MAXIMIZE THE POWER AT THE SAME TIME.
FIX AN α, MINIMIZE β AND MASSIMIZE π $(1-\beta)$

Power of a test

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

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

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).

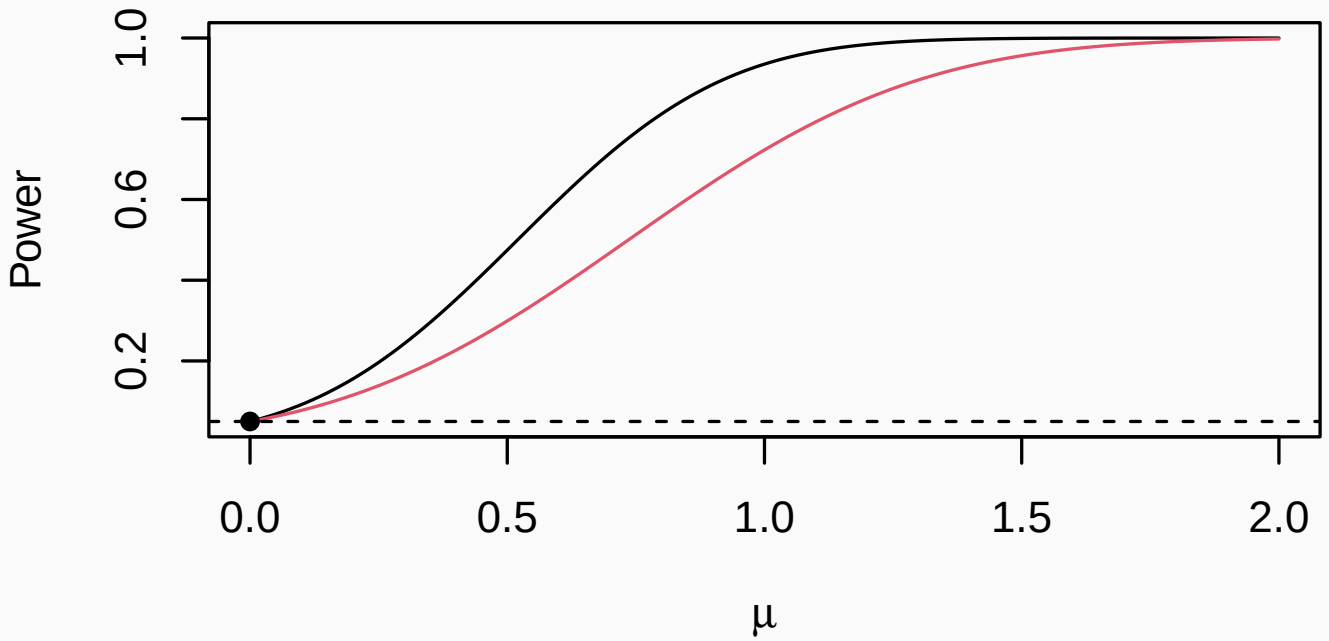
Fixing a significance level of 5%, the two tests have exactly the same probability of a Type I error, so for comparing them we must use their power.

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

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

R lab: power of two alternative tests

```
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. The results of statistical tests, and p-values in particular, should never taken without considering context-specific knowledge. Even a small p-value may not be particularly meaningful if the alternative hypothesis is logically implausible.

3. 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.