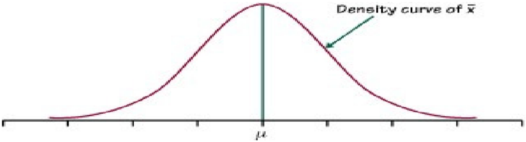
The Steps of a Significance Test

Example. We are making inference about the mean of the distribution of bank deposits of Italian families. Last year, we had $\mu = 3$ (tens of thousands of euro). We want to know if such mean this year is the same or more by knowing that $\sigma = 1$. To this end, we draw a random sample: 5,3,4,3,5,4,4,5,3. The estimated mean is $\bar{x} = 4$.



The difference between 3 and the estimate could be due to:

- sampling error;
- the fact that our idea is wrong.

How could we decide? ... We use a significance test

A *significance test* is a method of using data to summarize the evidence about a hypothesis. In other terms, it is a method to check whether or not the data support certain statements or predictions. These statements are hypotheses about population. A *significance test* about a hypothesis has *five steps*.

1. **Assumptions**
2. **Hypotheses**
3. **Test Statistic**
4. **p-value**
5. **Conclusion**

Step 1: Assumptions

Each significance test makes certain assumptions about some characteristics of the distribution of the population and the sampling distribution of a statistic.

In our example, we assume that the population has a normal distribution.

It implies that the sampling distribution of sample mean is normal as well (in real applications, it would be more realistic to assume the normality of the population on the log scale).

Step 2: Hypothesis

A *hypothesis* is a statement about a population, usually of the form that a certain parameter takes a particular numerical value or falls in a certain range of values.

Each significance test has *two hypotheses*:

- The *null hypothesis* (H_0) is a statement that the parameter takes a particular value. It has a single parameter value.
- The *alternative hypothesis* (H_a or H_1) states that the parameter falls in some alternative range of values. It specifies how the null can be false.

In our example we have $H_0 : \mu = 3$ vs $H_a : \mu > 3$.

In a significance test, the null hypothesis is presumed to be true unless the data give strong evidence against it.
Formally, we have to decide, on the basis of the information contained in the sample, if H_0 have to be rejected or not.

Step 3: Test Statistic

In order to decide if the null should be rejected or not, we summarize the evidence against the null by computing a (*test*) *statistic*.

In practice, a *test statistic* describes (measures) how far the point estimate falls from the parameter value given in the null hypothesis (usually in terms of the number of standard errors between the two).

The *test statistic* is zero when the estimate is equal to the value specified in the null. If the test statistic falls far from zero in the direction specified by the alternative hypothesis, it is evidence against the null hypothesis and in favor of the alternative hypothesis.

In our example a test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{4-3}{1/\sqrt{9}} = 3$$

Is it a large or small value? Should we reject or not?

Step 4: p-value

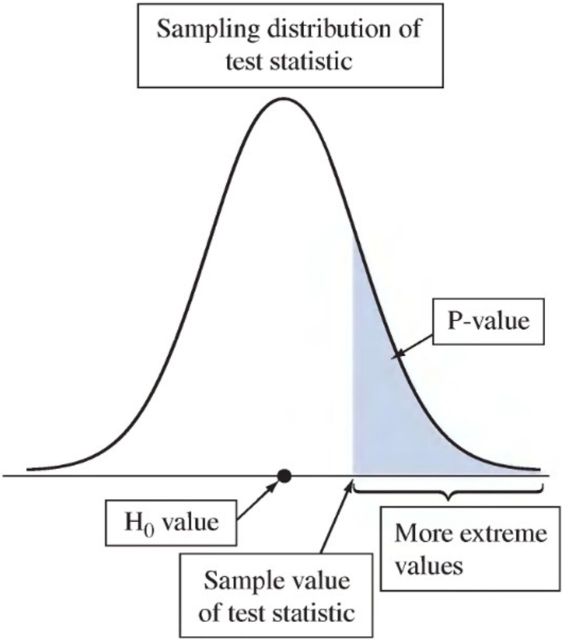
To interpret a test statistic value, we use a probability summary of the evidence *against* the null hypothesis, H_0.

- First, we presume that H_0 is true.
- Next, we consider the sampling distribution of the test statistic.

We summarize how far out in the tail of this sampling distribution the test statistic falls by computing, when H_0 is true, the probability of that value and values even more extreme. This probability is called a *p*-value.

The smaller the *p*-value, the stronger the evidence the data provide against the null hypothesis. That is, a small *p*-value indicates a small likelihood of observing the sampled results if the null hypothesis were true.

In our example *p*-value=P(Z $>$ 3)=1-P(Z$<$3)=1-0.9987=0.0013.


Sampling distribution of test statistic
P-value
H_0 value
Sample value of test statistic
More extreme values

Step 5: Conclusion see p. 405 of book for summary

Report and interpret the p-value in the context of the study. Based on the p-value, make a decision about H_0 (either reject or do not reject H_0 if the p-value is small or not).

Before seeing the data, we decide how small the p-value would need to be to reject H_0. This cutoff point is called the *significance level*.

In practice, the most common significance level is 0.05 (or 0.01 if the sample size is very large).

When we reject H_0 we say the results are *statistically significant*.

In our example we have p-value$=0.0013<0.05$ then we reject the null H_0.

Significance Tests About Proportions

Example: Are Astrologers' Predictions Better Than Guessing?

An astrologer prepares horoscopes for 116 adults.

Each subject also filled out a California Personality Index (CPI) survey.

For a given adult, his or her horoscope is shown to the astrologer along with their CPI survey as well as the CPI surveys for two other randomly selected adults. The astrologer is asked which survey is the correct one for that adult.

In the actual experiment, the astrologers were correct with 40 of their 116 predictions (a success rate of $0.345 > 0.333$).

Step 1: Assumptions

- The variable is categorical in 0-1 coding
- The sample size is sufficiently large that the sampling distribution of the sample proportion is approximately normal.

In our example we have 1 if the astrologer gives a correct answer and 0 otherwise.

Step 2: Hypotheses

- The **null hypothesis** has the form $H_0 : p = p_0$.
- The **alternative hypothesis** has the form:
  - $H_1 : p < p_0$ (one-sided, left tail, lower tailed)
  - $H_1 : p > p_0$ (one-sided, right tail, upper tailed)
  - $H_1 : p \neq p_0$ (two-sided, two tails, two tailed)

In our example $H_0 : p = 1/3$ (random guessing) and $H_1 : p > 1/3$ (the predictive power of the astrologer is effective).

Step 3: Test Statistic

The test statistic is:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

(Sample proportion - Null hypothesis proportion) / Standard error when null hypothesis is true

It has a standard normal distribution when the null is true.

In our example

$$Z_{obs} = \frac{0.345 - 0.333}{\sqrt{0.333(1-0.333)}}\sqrt{116} = 0.27$$

This z -score measures the number of standard errors between the sample proportion p_hat and the null hypothesis value p0

Step 4: p-value

The p-value summarizes the evidence. It describes how unusual the observed data would be if H_0 were true.
There are three ways to compute the p-value:

- $H_1 : p < p_0$ (one-sided, left tail)
  $\Rightarrow P(Z < z_{obs})$
- $H_1 : p > p_0$ (one-sided, right tail)
  $\Rightarrow P(Z > z_{obs})$
- $H_1 : p \neq p_0$ (two-sided, two tails)
  $\Rightarrow P(|Z| > |z_{obs}|)$
  $= P(Z < -|z_{obs}|) + P(Z > |z_{obs}|) = 2 \times P(Z > |z_{obs}|)$

In our example p-value$=P(Z > 0.27)=1\text{-}0.6064=0.3936$

1 - left = right

Step 5: Conclusion

Compare the p-value with the pre-specified significance level (usually 0.05).

In our example the p-value of 0.39 is greater than the significance level. We do not reject the null. There is *not* strong evidence that astrologers have special predictive powers.

Example. We want to test if a coin is fair or not. To this end, we toss the coin 121 times obtaining 55 heads. What can we conclude?

- **Step 1: Assumptions** the coin is fair = 0.5 H 0.5 T

  The assumptions are satisfied:

  $121 \times 0.5 = 121 \times (1 - 0.5) = 60.5 > 15.$
- **Step 2: Hypotheses**

  $H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$
- **Step 3: Test Statistic**

  $$z_{obs} = \frac{55/121 - 0.5}{\sqrt{0.5(1-0.5)}}\sqrt{121} = -1$$
- **Step 4: p-value**

  p-value$=2 \times P(Z > 1) = 2 \times (1 - 0.8413) = 0.3174$
- **Step 5: Conclusion**

  We do not reject the null because p-value>0.05.

Significance Tests About Means

Step 1: Assumptions

1. The variable is quantitative
2. The population distribution is normal.
3. Assumption 2. The sample size is sufficiently large that the sampling distribution of the sample mean is approximately normal (population variance is unknown); the exact distribution is a T-distribution.

Step 2: Hypotheses

- The **null hypothesis** has the form $H_0 : \mu = \mu_0$.
- The **alternative hypothesis** has the form:
  - $H_1 : \mu < \mu_0$ (one-sided, left tail, lower tailed)
  - $H_1 : \mu > \mu_0$ (one-sided, right tail, upper tailed)
  - $H_1 : \mu \neq \mu_0$ (two-sided, two tails, two tailed)

Step 3: Test Statistic The test statistic measures how far the sample mean falls from the null hypothesis value μ_0, as measured by the number of standard errors between them. The test statistic is

S is standard error as before (sample std dev / sqrt(n))

$$T_{obs} = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$$

omit sqrt(n) from the formula on the side in case

Under the null it is distributed as a Student t with $n-1$ degree of freedom (if assumption 2. is true). If n is large enough, it has a standard normal distribution even if the population is not approximately normal.

Step 4: p-value

The p-value describes how unusual the test statistics would be if H_0 were true. There are three ways to compute the p-value:

- $H_1 : \mu < \mu_0$ (one-sided, left tail)$\Rightarrow$P$(Z < t_{obs})$
- $H_1 : \mu > \mu_0$ (one-sided, right tail)$\Rightarrow$P$(Z > t_{obs})$
- $H_1 : \mu \neq \mu_0$ (two-sided, two tails) $\Rightarrow$P$(|Z| > |t_{obs}|)$=$2\times$P$(Z > |t_{obs}|)$

Step 5: Conclusion

Compare the p-value with a pre-specified significance level (usually 0.05). Reject the null if the p-value is less than the significance level.

Example. In a supermarket the mean of the amount of spending per client is 50 Euro. After a restoration, the management wants to test if the new organization induces the clients to increase the amount of spending. In a sample of $n = 196$ of buyers, drawn at random, they recorded $\bar{x} = 52$ and $s = 28$. What can we conclude?

- **Step 1: Assumptions**
  - The variable (amount of spending) is quantitative.
  - The data are obtained using randomization.
  - The sample size is large enough.
- **Step 2: Hypotheses**

  $H_0 : \mu = 50$ vs $H_1 : \mu > 50$.

REM: t and z statistics have the same form

- **Step 3: Test Statistic**

  The test statistic is

$$t_{obs} = \frac{52 - 50}{28}\sqrt{196} = 1$$

- **Step 4: p-value**

  p-value$=$P$(Z > 1)$=1-0.8413=0.1587

Z > 1 = Z < -1 = 1 - Z < 1

- **Step 5: Conclusion**

  We do not reject the null.

Decisions and Types of Errors in Significance Tests About Means

Because of sampling variability, decisions in significance tests always have some uncertainty. A decision can be in error.
Two are the possible errors:

	H_0 true	H_0 false
Accept H_0	Correct Decision	Type error II
Reject H_0	Type error I	Correct Decision

For every test it is possible to compute the probability of a type I or II error.

The significance level is the probability of a type I error

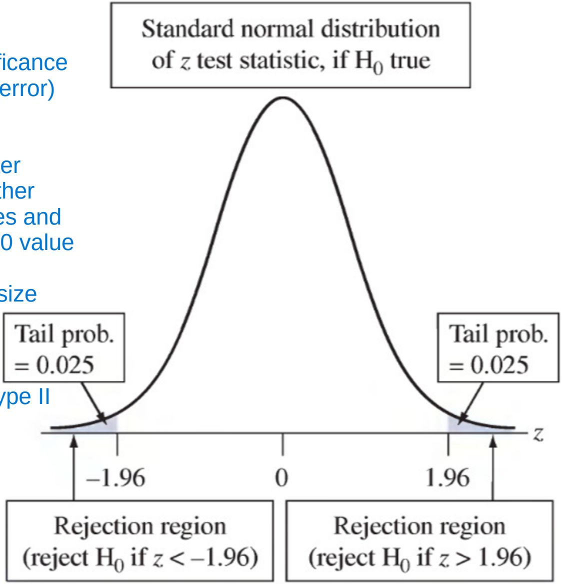
On type 2 error

For a fixed significance level , P(Type II error) decreases:

- as the parameter value moves farther into the Ha values and away from the H0 value

- as the sample size increases

Power = 1 - P(Type II error)



Optimal test

Summary of chapter p.453

The two error probabilities are closely (inversely) related. It follows that it is not possible to minimize both simultaneously.

How do we choose the best test?

Two approaches:

1. the best test minimizes a weighted mean of the two error probabilities;
2. (Neyman-Pearson approach) given an upper bound for the probability of a type I error, the best test minimizes the probability of a type II error for each possible value of the parameter in the region of the alternative hypothesis.

The tests that we have discussed before follows the second approach.