DATA SCIENCE &
ARTIFICIAL INTELLIGENCE

SCIENTIFIC &
DATA-INTENSIVE COMPUTING

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistical Inference

Introductory course on Statistics and Probability

Nicholas A. Pearson

Università degli Studi di Trieste

September 12, 2025

Introduction

Statistics aims to **extract information from data**, and in particular on the process that generated the data.

Two intrinsic difficulties:

- It may be hard to infer what we wish to know from the available data;
- Most data contain some random variability: by replicating the data-gathering process several times we would obtain different data on each occasion.

We search for conclusions drawn from a single data set that are generally valid, and not the result of random peculiarities of that data set.

Terminology

Some key terminology:

- **Population**: the complete set of individuals, items, or observations of interest in a study;
- **Sample**: a subset of the population selected for analysis;
- **Parameter**: a numerical value that characterizes a population. Generally unknown and the primary quantity of interest in inference;
- **Statistic**: a numerical value computed from a sample. Used to estimate the corresponding population parameter;

Example:

- **Population**: students enrolled at the University of Trieste;
- **Sample**: students following this lecture;
- **Parameter**: mean height of the population μ;
- **Statistic**: sample mean $\bar{\mu}$

Statistical Models

Statistics is able to draw conclusions from random data mainly though the use of **statistical models**.

A statistical model can be thought as a mathematical cartoon describing how our data might have been generated, if the unknown features of the data-generating process were actually known.

If the unknowns were known, a good model would be able to generate data resembling the main features of observed data.

The **purpose of statistical inference** is to use the statistical model to go in the reverse direction: to **infer the model unknowns that are consistent with the observed data**.

Notation

Notation:

- y: random vector containing the **observed data**;
- θ: **vector of parameters of unknown value**;

We assume that knowing the parameters would answer the question of interest about the process generating the data.

The model specifies how data akin to y may be simulated, implicitly defining the distribution of y and how it depends on θ.

Moreover, a statistical model may depend on some known parameters γ and some further data x, treated as known and denoted as *covariates* or *predictor variables*.

Inferential questions

Given a statistical model for data y, with model parameters θ, there are some basic questions to ask:

1. **Point estimation**:
   what values of θ are most consistent with y?
2. **Interval estimation**:
   what range of values of θ are consistent with y?
3. **Hypothesis testing**:
   is y consistent with a hypothesis on the value of θ?

The central issue is the acknowledgment of the intrinsic uncertainty inherent in trying to learn about θ.

There are two classes of methods providing an answer to questions 1-3, namely the *frequentist* and *Bayesian* approach. We will focus on the former.

Point estimation

Given a model for the data y, with parameter θ, **point estimation** is concerned with finding a reasonable parameter estimate from the data.

The problem can be simply stated as finding the parameter value most consistent with the data, a definition that leads to the method of *maximum likelihood estimation*.

Sample Mean

We start with a simple model that assumes that the data are a random sample of i.i.d. random variables from a normal distribution:

$$y_1, y_2, \ldots, y_n \sim \text{i.i.d } \mathcal{N}(\mu, \sigma^2)$$

The parameters we may be interested in estimating are μ and σ^2.

Sample Mean

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Sample Variance

We start with a simple model that assumes that the data are a random sample of i.i.d. random variables from a normal distribution:

$$y_1, y_2, \ldots, y_n \sim \text{i.i.d } \mathcal{N}(\mu, \sigma^2)$$

The parameters we may be interested in estimating are μ and σ^2.

Sample Variance

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Sample Mean and Sample Variance

We start with a simple model that assumes that the data are a random sample of i.i.d. random variables from a normal distribution:

$$y_1, y_2, \ldots, y_n \sim \text{i.i.d } \mathcal{N}(\mu, \sigma^2)$$

The parameters we may be interested in estimating are μ and σ^2.

- **Sample Mean**: $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **Sample Variance**: $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Such estimates are actually sensible anytime we are interested in estimating the mean and variance of an i.i.d. sample.

Example

Assume that $n = 7$ people are driving on the highway and let the following be the kilometers they drove in one hour.

$$113, 109, 102, 105, 118, 102, 114$$

Compute the mean and standard deviation of this sample.

Example

- **Sample mean**

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n} = \frac{113 + 109 + 102 + 105 + 118 + 102 + 114}{7}$$
$$= 109$$

- **Sample Variance**

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n-1} = \frac{(113 - 109)^2 + \cdots + (114 - 109)^2}{7-1}$$
$$= \frac{236}{6} \approx 39.33$$

- **Sample Standard Deviation**

$$s = \sqrt{s^2} = \sqrt{39.33} \approx 6.27$$

Unbiased Estimators

To figure out what could be a **good estimate**, we need to consider repeated estimation under **repeated replications of the data-generating process**.

The point is: what do we expect to find if we repeat the same analysis to many data sets generated from the same model?

Unbiased Estimators

If we replicate the random data and we repeat the estimation process, the result will be a different value of $\hat{\theta}$ for each replicate.

Since, the estimator is a r.v., it makes fully sense to compute its mean.

An estimator is said to be **unbiased** if:

$$E(\hat{\theta}) = \theta$$

Unbiasedness is a desirable property, and we would also like the estimator to have low variance.

Bias-variance Tradeoff

There is a tradeoff between unbiasedness and low variance, so we usually seek to get both (to some extent): ideally we would target a small Mean Squared Error (MSE);

$$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}$$

Which can be expressed as:

$$MSE(\hat{\theta}) = \{E(\hat{\theta} - \theta)\}^2 + var(\hat{\theta}) = \text{ Squared bias + Variance}$$

For a *normal random sample* it is easy to verify that both the *sample mean and sample variance are unbiased*.

The unbiasedness of the sample mean and variance is a general property, holding also for non-normal samples.

Other Properties

Consistency

A scalar estimator is said to be **weakly consistent** if, for any $\epsilon > 0$:

$$Pr(|\hat{\theta} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

Robustness

An estimator is said to be **robust** if it has good performance across a wide range of statistical models for the data.
The sample median is a robust estimation of location, not affected by possible outlying data.

Interval Estimation

Confidence intervals (C.I.) provide more satisfactory estimation results than point estimates alone, giving an entire set of values to estimate the model parameter.

They are built by considering a **single parameter at a time**.

C.I. are built using **pivots**, functions which have a *known distribution* and are *applied to both the data and the parameter of interest*.

Pivot Example

We start with a simple example where:

- We have a random sample $y_1, y_2, \ldots, y_n \sim$ i.i.d $\mathcal{N}(\mu, \sigma^2)$
- Our parameter of interest is the mean μ
- The variance σ^2 is assumed to be known

In this scenario the pivot is:

$$Z(\mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Obtaining a confidence interval

Given a value of α s.t. $0 < \alpha < 1$ it holds that:

$$Pr(z_{\alpha/2} \leq Z(\mu) \leq z_{1-\alpha/2}) = 1 - \alpha$$

Where a generic z_α is the α-th quantile of a standard normal distribution.

For symmetry $z_{\alpha/2} = -z_{1-\alpha/2}$.

This formulation is equivalent to:

$$Pr\left(\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Confidence Interval

Hence the random interval with endpoints:

$$\bar{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

contains the value μ with probability $(1-\alpha)$ and is called a $(1-\alpha)\times 100\%$ **confidence interval**.

Common choices for $(1-\alpha)$ are 0.95 or 0.99.

Confidence Interval Interpretation

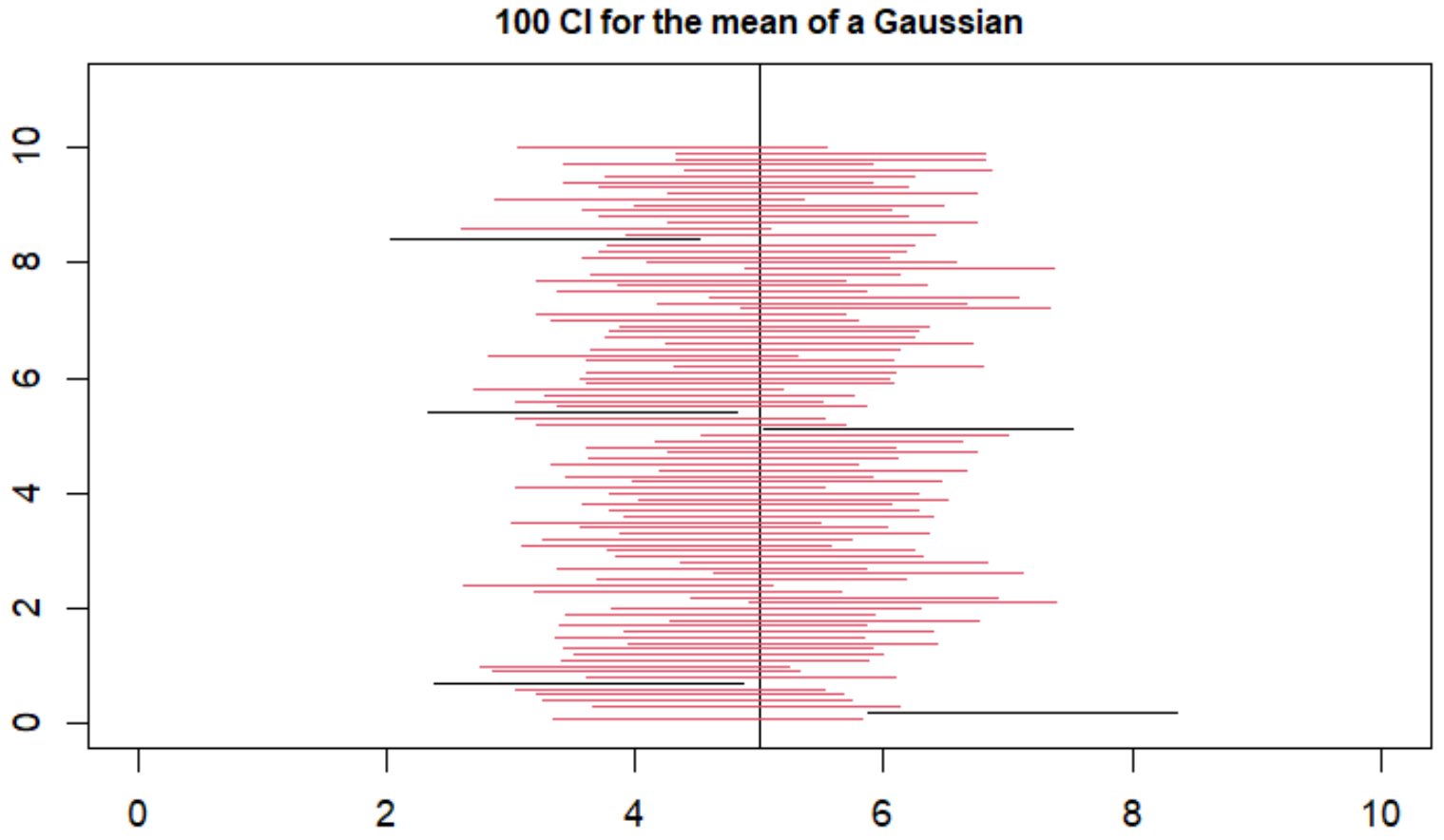
Given a specific set of data $y_1, \ldots, y_n$ we compute the confidence interval by replacing $\bar{Y}$ with the **observed sample mean** $\bar{y}$.

$$\bar{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \bar{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

This interval either does or does not contain the true value of μ.

Given a hypothetical sequence of *100 sets of data generated from the statistical model*, if we compute a confidence interval for each of these sets we expect that on average $(1-\alpha) \times 100\%$ of these intervals will include the true μ.

Confidence Interval Interpretation



Example

A gunpowder manufacturer developed a new formula that was tested on eight bullets. The resultant initial velocities (in m/s) are:

$$916, 892, 895, 904, 913, 916, 895, 885$$

Assuming that the initial velocities have normal distribution with $\sigma = 12$ m/s, compute the confidence interval for the significance level $\alpha = 0.05$ for the initial mean velocity of the bullets.

Example

- We know thar $X \sim \mathcal{N}(\mu, 12^2)$ with μ unknown
- From the sample we can derive that:

$$n = 8 \qquad \bar{x} = 902$$

and we know that $\alpha/2 = 0.025$ and $z_{1-\alpha/2} \approx 1.96$

- Then the confidence interval can be computed as:

$$CI_{0.95}(\mu) = \left[\bar{x} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[902 \pm 1.96 \cdot \frac{12}{\sqrt{8}}\right]$$
$$\approx [893.68, 910.32]$$

Pivot Example

Let's modify just one condition from the previous example:

- We have a random sample $y_1, y_2, \ldots, y_n \sim$ i.i.d $\mathcal{N}(\mu, \sigma^2)$
- Our parameter of interest is the mean μ
- **The variance σ^2 is assumed to be not known**

In this scenario the pivot is:

$$T(\mu) = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where s^2 is the estimated sample variance.

Confidence Interval

Hence the random interval with endpoints:

$$\bar{Y} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \qquad \bar{Y} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

contains the value μ with probability $(1-\alpha)$.

The notation $t_{n-1,\alpha}$ indicates the α-th quantile of a Student't distribution with $n-1$ degrees of freedom.

Example

Let's consider the same example as before, but assume now that the variance is unknown.

- We know thar $X \sim \mathcal{N}(\mu, \sigma^2)$ with μ, σ^2 unknown
- From the sample we can derive that:

$$n = 8 \qquad \bar{x} = 902 \qquad s^2 = 143.43$$

and we know that $\alpha/2 = 0.025$ and $t_{n-1,1-\alpha/2} = t_{7,0.975} \approx 2.365$

- Then the confidence interval can be computed as:

$$CI_{0.95}(\mu) = \left[\bar{x} \pm t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] = \left[902 \pm 2.365 \cdot \sqrt{\frac{143.43}{8}}\right]$$

$$\approx \left[891.99, 912.01\right]$$

CI for a proportion

Is it possible to define a confidence interval on the proportion π, the success probability of a random sample of n binary variables.

- We have a random sample $y_1, y_2, \ldots, y_n \sim$ i.i.d $\mathcal{B}(1, \pi)$
- Our parameter of interest is the proportion π

In this scenario the pivot is:

$$Z(\pi) = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \sim \mathcal{N}(0,1)$$

and $\hat{\pi} = \bar{Y}$ and $SE(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$, which is estimated by plugging in $\hat{\pi}$ in place of π.

CI for a proportion

Hence the random interval with endpoints:

$$\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \qquad \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

contains the value π with probability $(1-\alpha)$

Exercise

In the Friuli Venezia Giulia Region, there were $n = 10337$ newborns, of which $x = 5286$ were born males. What is the confidence interval at level 95% of the proportion of male newborns ?

Exercise

From the sample we can compute that

$$\hat{\pi} = \frac{5286}{10337} = 0.511$$

Then the interval can be defined as:

$$\left[\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Plugging in $\hat{\pi} = 0.511$ and $n = 10337$:

$$\left[0.511 \pm 1.96\sqrt{\frac{0.511(1-0.511)}{10337}}\right] = [0.501, 0.521]$$

Hypothesis Testing

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

This can be denoted as:

$$H_0 : \theta = \theta_0$$

with H_0 being the **null hypothesis**.

In is also necessary to select a complementary **alternative hypothesis** H_1, which specifies the values of the parameter which become reasonable when H_0 does not hold.

Hypothesis Testing

We have a random sample $y_1, y_2, \ldots, y_n \sim$ i.i.d.$\mathcal{N}(\mu, \sigma^2 = 1)$.
We may want to test the following **set of hypothesis**.

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

The null hypothesis implies that the test is on $Y_i \sim \mathcal{N}(0, 1)$.

What does the alternative hypothesis suggest?

This formulation of the alternative hypothesis is called **one-sided alternative**, while a **two-sided alternative** takes the form $H_1 : \mu \neq 0$.

Key Concepts in Hypothesis Testing

We will discuss the following concepts, which are central (and somewhat interchangeable points of view), to hypothesis testing.

- Test statistic
- Null and alternative distributions
- p-value
- Significance level
- Rejection and Acceptance regions

Test Statistic

Test statistics are similar to the pivot introduce for the confidence intervals. A **test statistic** is a statistic (a function applied to the sample) which has known distribution.

An example of a test statistic, for the previous hypothesis test, is:

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

In the previous scenario we know that $\sigma^2 = 1$ and the null hypothesis is that $H_0 : \mu = 0$, therefore:

$$Z = \frac{\bar{Y}}{\sqrt{\frac{1}{n}}} = \sqrt{n}\bar{Y}$$

Null and alternative distributions

Once the test statistic has been appropriately selected, the choice of the parameters in the null hypothesis and alternative hypothesis identify the **null** and **alternative distributions**.

- if H_0 is true (*under* H_0), then

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

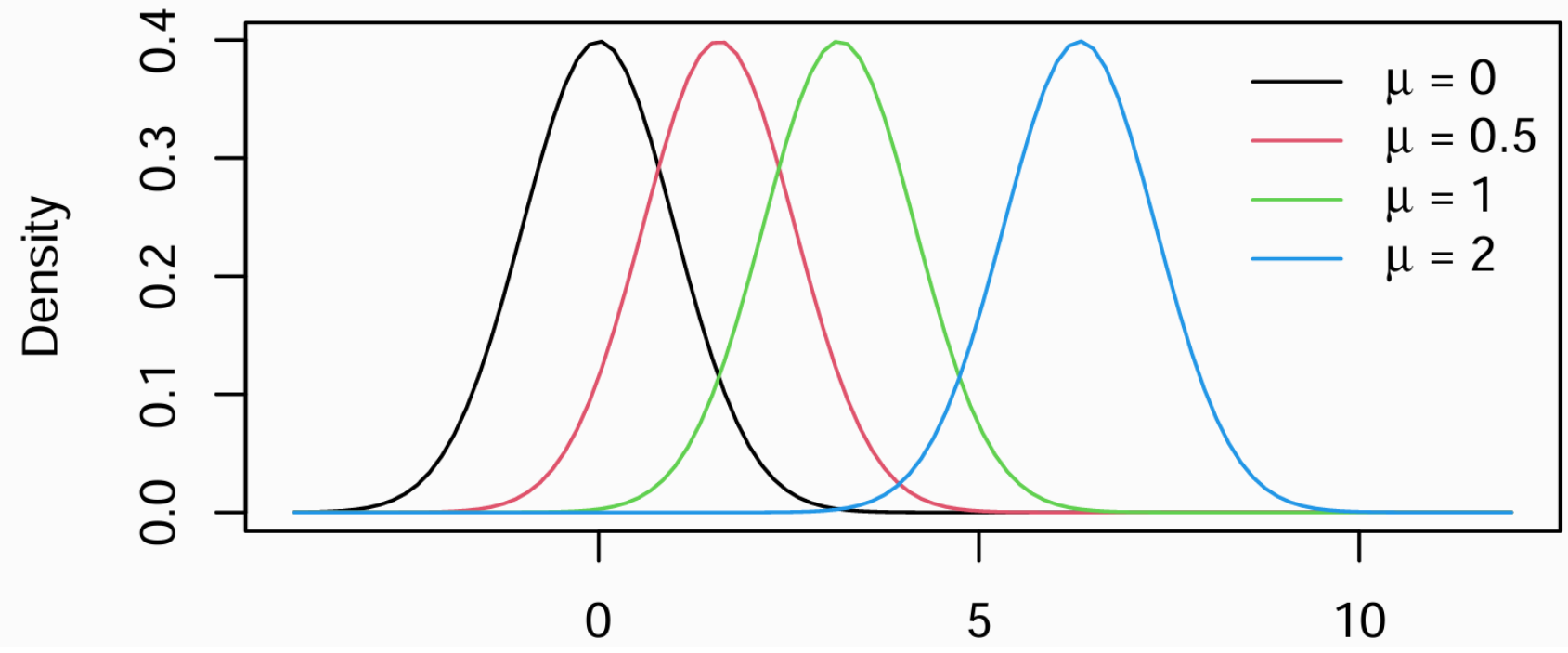
  This is called the null distribution of Z.

- if H_1 holds (*under* H_1), it generally holds that

$$Z \sim \mathcal{N}(\mu_1, 1)$$

  with $\mu_1 > 0$. This is known as the alternative distribution of Z.

Null and alternative distributions



Hypothesis testing

If we have a random sample $y_1, y_2, \ldots, y_n \sim$ i.i.d.$\mathcal{N}(\mu, \sigma^2)$, with σ^2 assumed to be unknown we have similar results to those found for confidence intervals.

In particular:

$$T = \frac{\bar{Y} - \mu}{\sqrt{\frac{s^2}{n}}} \overset{H_0}{\sim} t_{n-1}$$

The symbol $\overset{H_0}{\sim}$ indicates the distribution of the test statistic under the null hypothesis.

p-value

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

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

It measures "how likely" is the observed value with respect to the null distribution.

In the previous example, under the null hypothesis we know that $Z \sim \mathcal{N}(0, 1)$ then the p-value is computed as :

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

Note that, with $H_1 : \mu < 0$ or $H_1 : \mu \neq 0$ the p-value is computed in a different manner.

Significance level

How to know if the result of a test was in line with the null hypothesis or not? We use a **significance level** to compare the p-value to.

The most common choice is selecting a significance level equal to 5% and it is said that: *the result of a test is significant at the* 5% *level when the p-value is smaller or equal to 0.05*

In this case we say that *we reject the null hypothesis*. If the p-value is larger than the significance level then *we accept the null hypothesis*.
Other choices for the significance level are 0.1% or 0.01%.

Example

The heights of male students in a college are thought to be Normally distributed with mean 170cm and standard deviation 7.

The height of 10 male students are measured and the sample mean is computed at 174 cm.

Determine, at the 5% significance level, if there is evidence that the height of male students of this college is higher than 170cm.

Example

From the text of the exercise we know that $Y_i \sim \mathcal{N}(\mu = 170, \sigma^2 = 7^2)$. From a sample $n = 10$, it was found that $\bar{Y} = 174$cm.
The set of hypothesis to test is:

$$H_0 : \mu = 170$$
$$H_1 : \mu > 170$$

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{174 - 170}{\frac{7}{\sqrt{10}}} \approx 1.807$$

$p = 1 - \Phi(1.807) = 0.035 < 0.05$ means that the *null hypothesis is rejected*, that is, there is enough significant evidence to state that the average height of the male students is greater than 170cm.

Example

An group of physicians is studying the relationship between triglycerides and obesity. Known results from the CNR suggest that, for non overweight people. the distribution of triglycerides follows a normal distribution with $\mu = 145$mg/dl and $\sigma = 80$ml/dl.

For a sample of $n = 20$ overweight patients, the mean triglycerides count was 165 mg/dl.

Is it possible to say that clinically obese patients have a higher triglycerides count with respect to the rest of the population? Use a significance level at 5%.

Exercise

The set of hypothesis to test is:

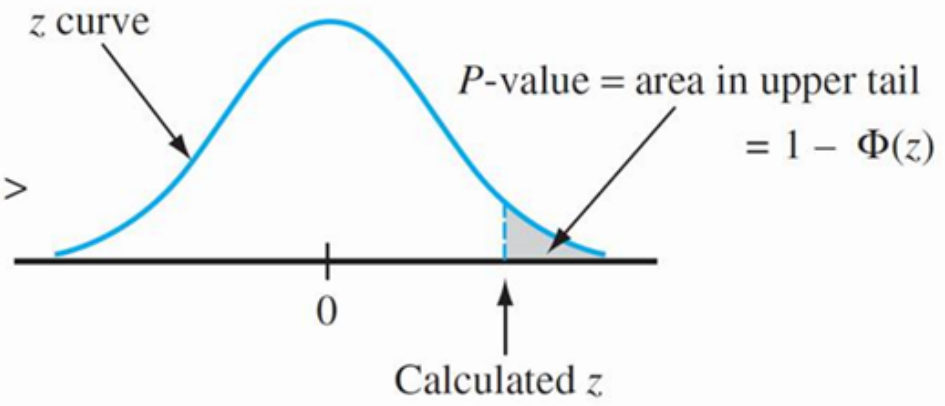
$$H_0 : \mu = 145$$
$$H_1 : \mu > 145$$

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{165 - 145}{\frac{80}{\sqrt{20}}} \approx 1.118$$

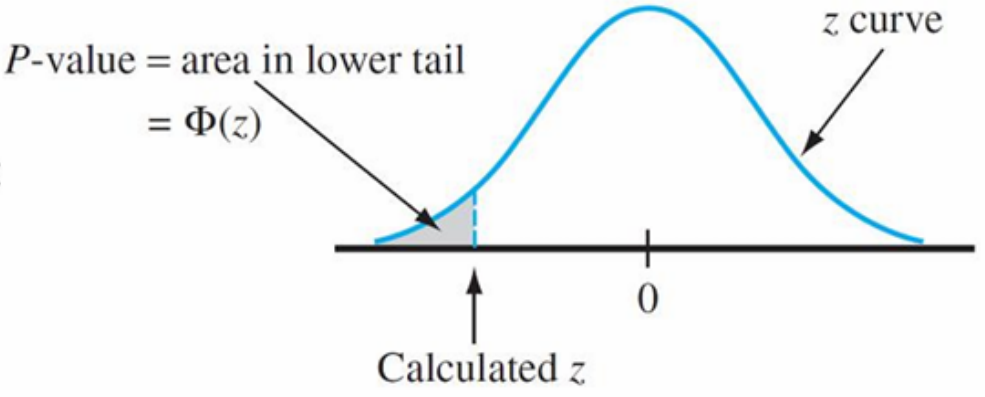
$p = 1 - \Phi(1.118) = 0.13 > 0.05$ means that the *null hypothesis is accepted*.

One-sided test

1. **Upper-tailed test**
   H_a contains the inequality $>$

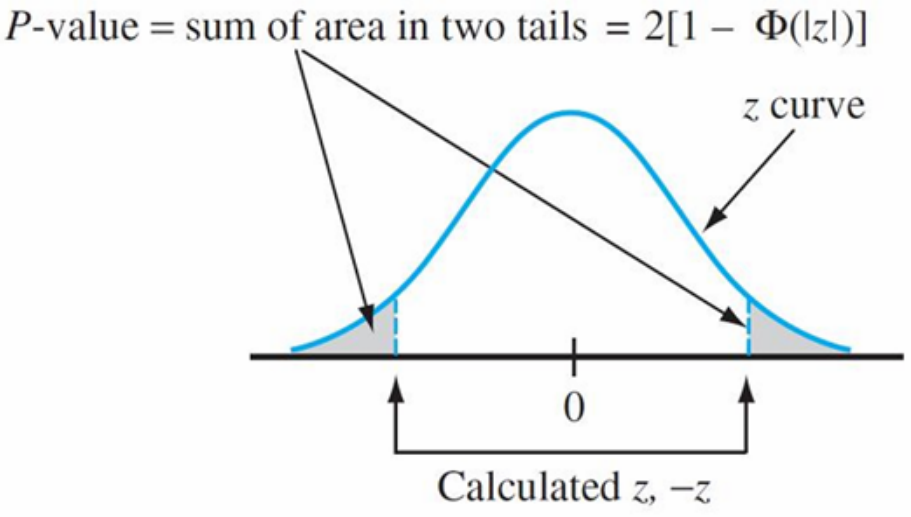


2. **Lower-tailed test**
   H_a contains the inequality $<$



Two-sided test

3. **Two-tailed test**
   H_a contains the inequality $\neq$



Rejection and Acceptance regions

Starting from the **sample space** (the set of possible values that the sample may take), it is possible to divide the space into a **rejection region** and an **acceptance region**.
Both are defined starting from the test statistic.
For the set of hypothesis

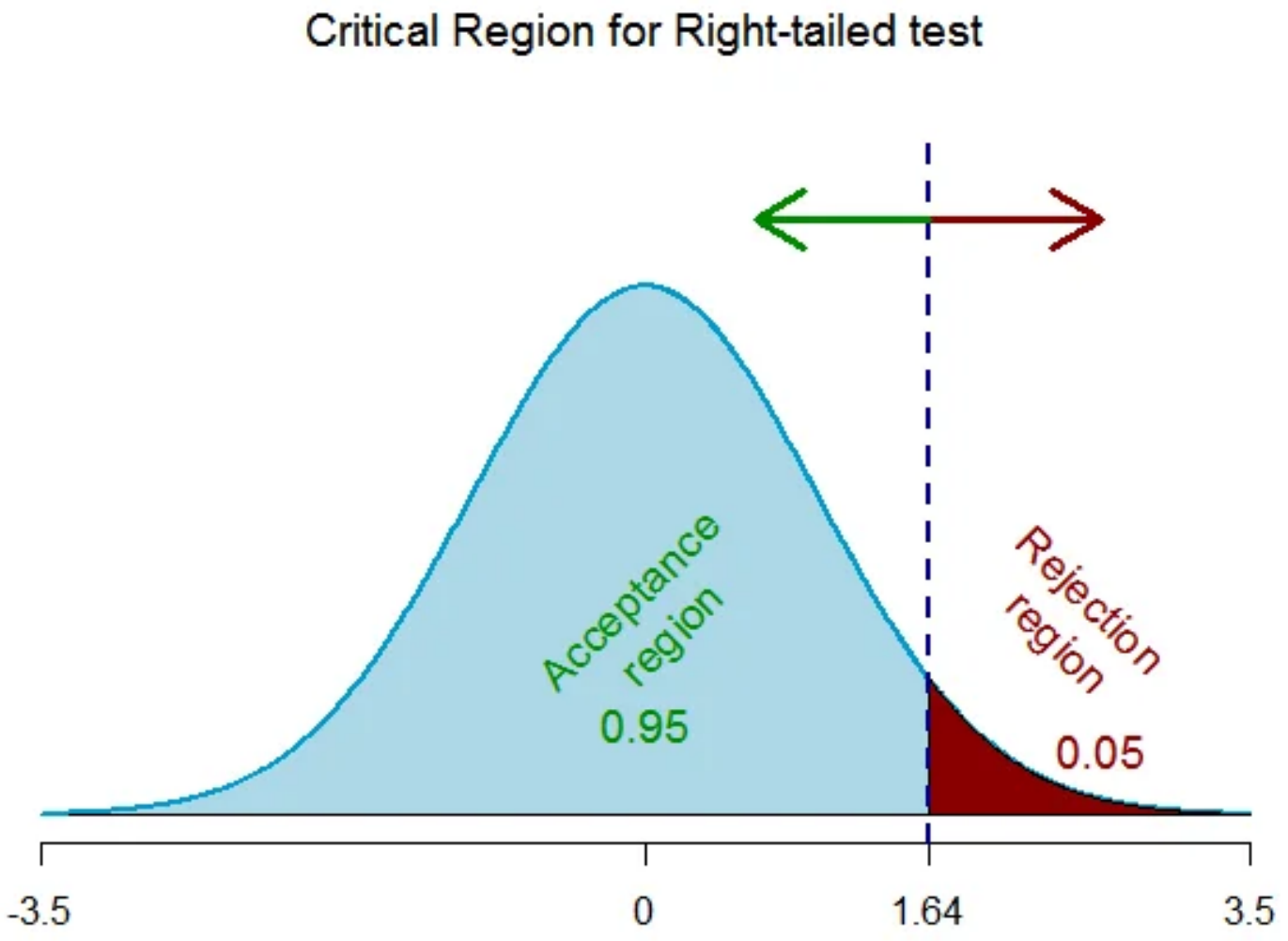
$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu > 0 \end{cases}$$

then the regions are defined by:

- $\mathcal{R}_\alpha = \{y \text{ s.t. } Z \geq z_{1-\alpha}\}$
- $\mathcal{A}_\alpha = \{y \text{ s.t. } Z < z_{1-\alpha}\}$

where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of a standard normal distribution. For $\alpha = 0.05$, $z_{0.95} = 1.645$.

Rejection and Acceptance regions



Example - heights

$$H_0 : \mu = 170$$
$$H_1 : \mu > 170$$

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{174 - 170}{\frac{7}{\sqrt{10}}} \approx 1.807$$

Knowing that $z_{1-\alpha} = 1.645$:

$$\mathcal{R}_\alpha = \{z_{obs} > 1.645\} \qquad \mathcal{A}_\alpha = \{z_{obs} \leq 1.645\}$$

Then $1.807 > 1.645 \rightarrow$ null hypothesis is *rejected* ($1.807 \in \mathcal{R}_\alpha$).

Example - triglycerides

$$H_0 : \mu = 145$$
$$H_1 : \mu > 145$$

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{165 - 145}{\frac{80}{\sqrt{20}}} \approx 1.118$$

Knowing that $z_{1-\alpha} = 1.645$:

$$\mathcal{R}_\alpha = \{z_{obs} > 1.645\} \qquad \mathcal{A}_\alpha = \{z_{obs} \leq 1.645\}$$

Then $1.118 < 1.645 \rightarrow$ null hypothesis is *accepted* $(1.118 \in \mathcal{A}_\alpha)$