Machine Learning from Data

Laboratory Sessions

Verónica Vilaplana

veronica.vilaplana@upc.edu

TSC - Campus Nord – D5 - 118

Organization of the laboratory sessions

- Teams of 2 students

- 7 assignments + competition

- Material in Atenea
  - Assigment description
  - Matlab code
  - Report template
  - Slides / notes

- Each assignment is presented during the on-site session and (usually) the corresponding report should be handled through Atenea one week later

- Install Matlab following instructions in Atenea
  - "Install a MATLAB license in your computer"

Schedule

- Lab: Monday 10 – 11, Room D5-010

Date	Session
oct 1	Lab0: Exploratory data analysis
oct 8	Lab0: Exploratory data analysis
oct 15	Lab1: MAP for Gaussian data
oct 22	Lab2: Databases and feature selection, PCA, MDA
oct 29	Lab2: Databases and feature selection, PCA, MDA
nov 5	Lab3: Parzen , K Nearest neighbors
nov 12	Lab3: K Nearest neighbors / Lab4: Support Vector Machines
nov 19	Lab4: Support Vector Machines
nov 26	Lab5: Neural networks
dec 3	Lab6: Tree classifiers
dec 10	Lab7: Competition (Kaggle)
dec 17	Lab7: Competition (Kaggle)

Evaluation

1. Laboratory assignments (test of algorithms, evaluation of results, Matlab programming): **25%**
2. Delivery of proposed homework: **15%**
3. Participation in the competition: **15%**
4. Final exam: exercises related to theory, laboratory and homework: **45%**

It is compulsory to attend lectures, deliver homework, deliver assignments and do the final exam.

Lab 0

Exploratory data analysis

Departament de Teoria
del Senyal i Comunicacions

UPC

Outline

1. Introduction
2. Dataset: EnfX
3. Scatter plots
4. Histograms
5. Moments
6. Quantiles
7. Confidence intervals
8. Hypothesis test

Introduction

Exploratory data analysis helps to understand and sumarize a dataset before applying any machine learning model. It can be use to analyze the distribution of features and, in particular, to decide whether we may assume a normal distribution.

Gaussianity

Analysis:

- Histograms, comparison with a Gaussian distribution.
- Cumulative histogram, comparison with a cumulative Gaussian.
- Moments: mean, variance, skewness, kurtosis
- Distribution plots: normplot o qqplot
- Confidence intervals for distribution parameters. Given a confidence level, decide whether the sample distribution follows a Gaussian distribution.

Additionally, scater plots for pairs of classes are useful to analyze the separability of the classes

Separability

EnfeX dataset

Goal: Determine if one person is affected by a disease X

A dataset is created gathering information from N= 1000 subjects (healthy and sick).

For each subject there is a vector of **dimension d=**4 containing the following information:

- Age: 40 ≤ v(1) ≤ 70
- Average blood pressure: 6 ≤ v(2) ≤ 16
- Cholesterol level: 1 ≤ v(3) ≤ 3,5
- Weight (Kg) / height (mts): 30 ≤ v(4) ≤ 80

4 features

2 clases

For each subject there is a **label = 1 (sick) / 0 (healthy)**

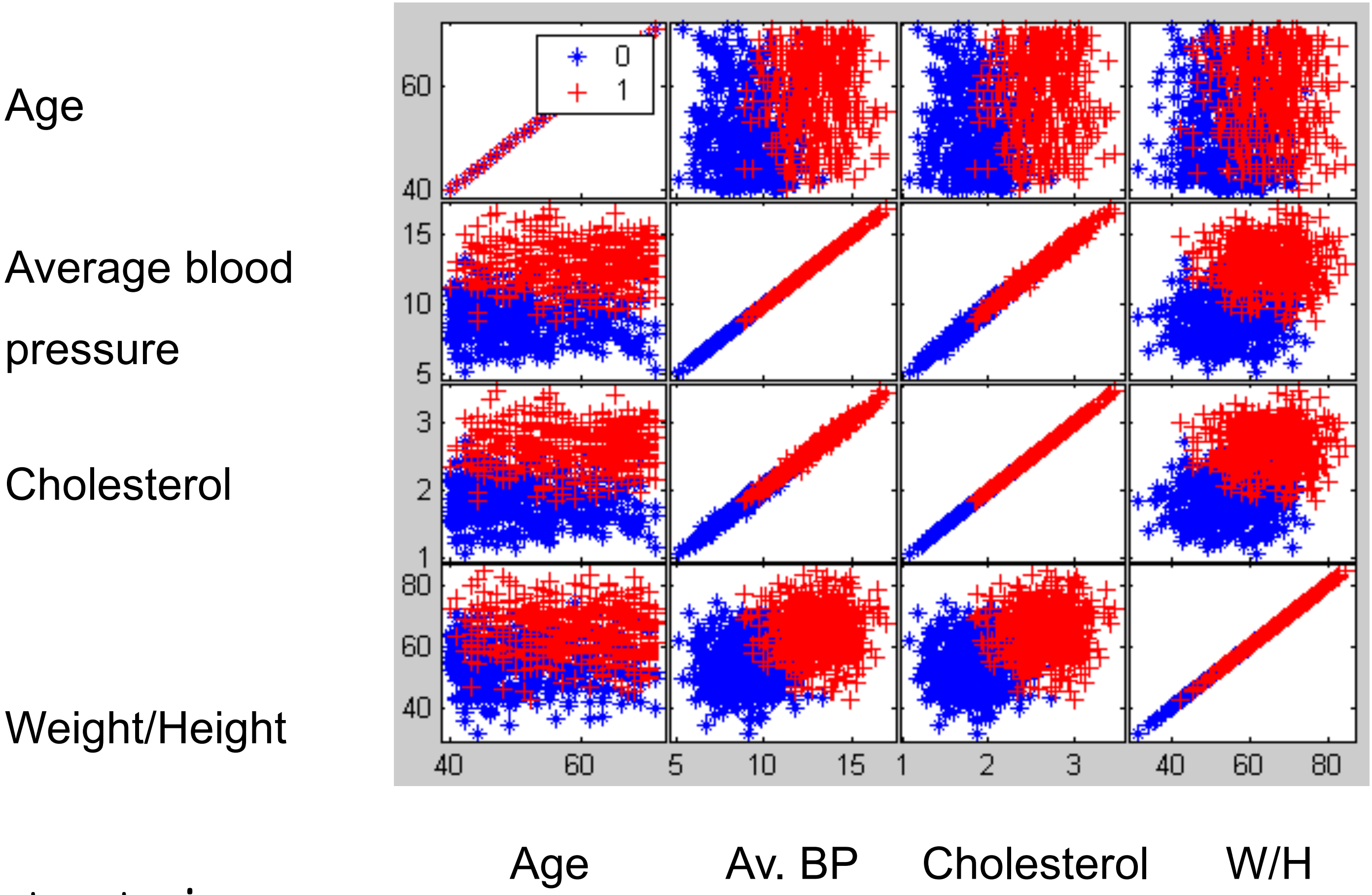
Example:

Age	Average blood pressure	Cholesterol	Weight/Height
50	12	2	35

We will use this dataset to illustrate the different EDA tools

Scatter plots

A **scatter plot** shows all dataset elements by pairs of features and grouped by classes in a single figure



```
gplotmatrix
```

How to diagnose if a person is affected by a disease X or if has high risk of developing the disease (H1 hypothesis) ?

$$\Pr(H_1 | x(1), x(2), \ldots, x(d)) \underset{H_1}{\overset{H_2}{\substack{< \\ >}}} \Pr(H_2 | x(1), x(2), \ldots, x(d))$$

MAP criterion

$$\Pr(H_1 | x(1), x(2), \ldots, x(d)) =$$

$$= \frac{f(x(1), x(2), \ldots, x(d) | H_1) \Pr(H_1)}{f(x(1), x(2), \ldots, x(d) | H_1) \Pr(H_1) + f(x(1), x(2), \ldots, x(d) | H_2) \Pr(H_2)}$$

¿Are $f(.|H_0)$ and $f(.|H_1)$ Gausian functions?

If they are Gaussian, we can apply the classification techniques studied in Unit 2.1

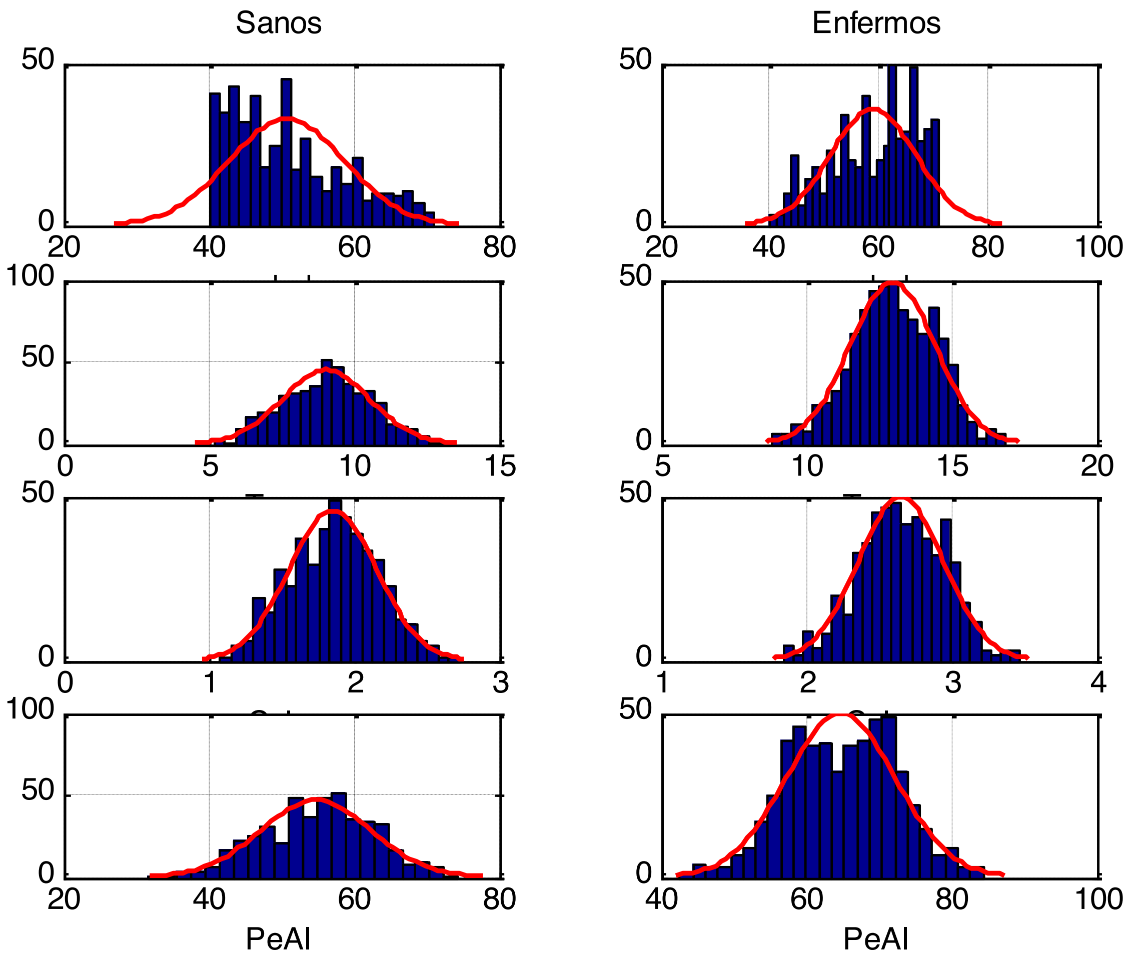
Gaussianity test for features using histograms

Histogram

It counts the number of observations that fall into each bin

It graphically represents the frequency distribution of the dataset

```
histfit
```



The first feature (age) does not follow a Gaussian distribution

Gaussianity test for features using the cdf

$$F_X(x) = \Pr\left\{X \le x\right\}$$

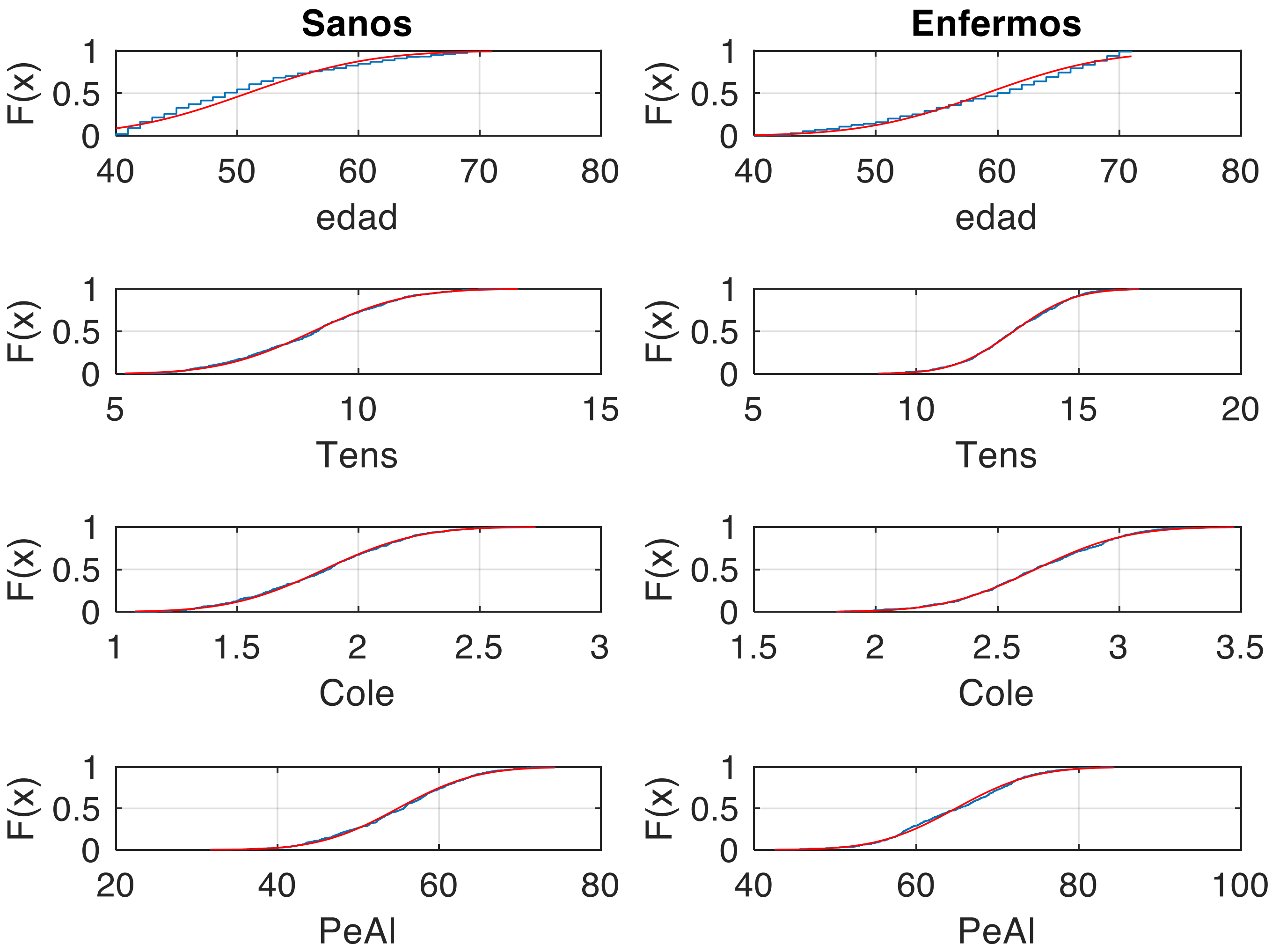
Cumulative Density Function

Cumulative histogram

Blue: feature cdf

Red: Gaussian fit

`cdfplot`



The first feature (age) does not follow a Gaussian distribution

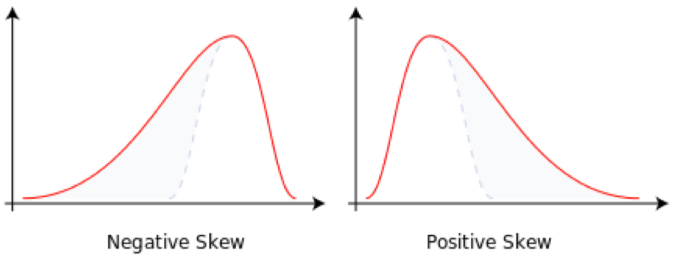
Moments

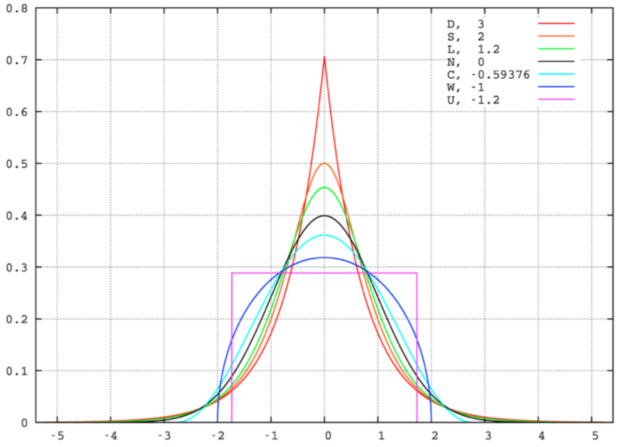
Mean $Mean(x) = \mu_1 = \mu = E\left[x\right]$

Variance $Var(x) = \sigma^2 = E\left[(x-\mu)^2\right] = \mu_2$

Skewness $Sk(x) = \dfrac{\mu_3}{\mu_2\sqrt{\mu_2}} = \dfrac{E\left[(x-\mu)^3\right]}{\sigma^3}$

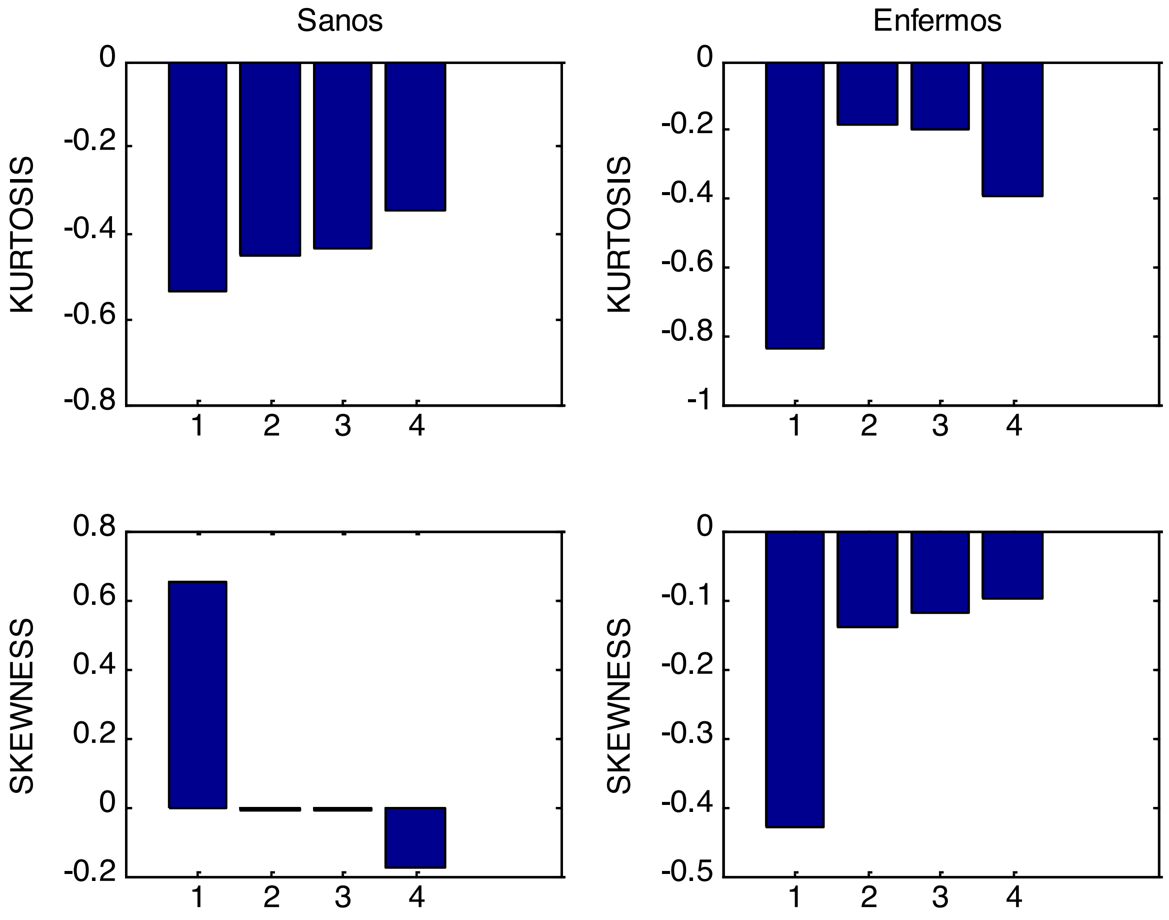
Kurtosis $K(x) = \dfrac{\mu_4}{(\mu_2)^2} = \dfrac{E\left[(x-\mu)^4\right]}{\sigma^4}$





If x es Gaussian, then *Sk*(*x*)=0 and *K*(*x*)-3 =0 (note that these are not sufficient conditions for gaussianity)

Example

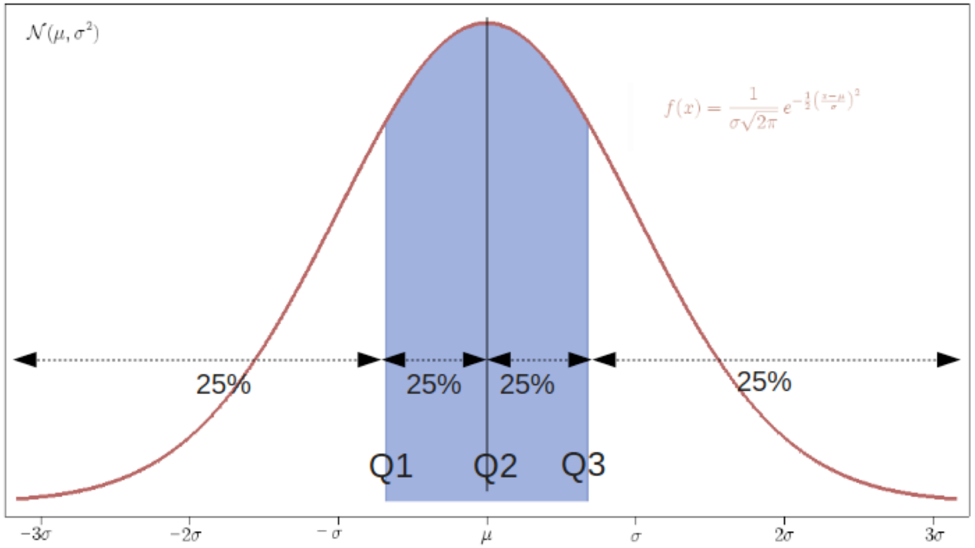

Sanos
Enfermos
KURTOSIS
0
-0.2
-0.4
-0.6
-0.8
1
2
3
4
KURTOSIS
0
-0.2
-0.4
-0.6
-0.8
-1
1
2
3
4
SKEWNESS
0.8
0.6
0.4
0.2
0
-0.2
1
2
3
4
SKEWNESS
0
-0.1
-0.2
-0.3
-0.4
-0.5
1
2
3
4

Quantiles

- **Quantiles** are cut points dividing the range of a probability distribution into continuous intervals with equal probability or dividing the observations in a sample in the same way.
- For a sample: q-quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes. There are q-1 of the q-quantiles, for each integer k, 0 < k < q.

Examples:

- **quartiles** divide the distribution in four parts (quantiles 0.25, 0.5 and 0.75)
- **percentiles** divide the distribution in one hundred parts



Gaussianity test for features

- A quantile-quantile plot (q-q plot) is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- First, the set of intervals for the quantiles is chosen. A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).
- If the two distributions are similar, the points in the q-q plot will approximately lie on the line y=x. If the distributions are linearly related, the plots will approximately lie on a line, but not necessarily on the line y=x.

- To compare the sample distribution with a Gaussian distribution

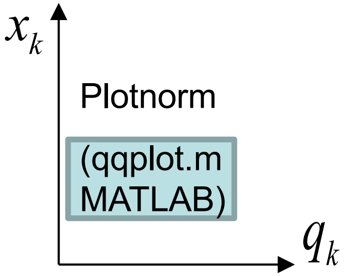
- sort the n samples in ascending order x_1, x_2, …x_n

- compute quantiles of a normal distribution for k=1, 2, …, n

$$q_k = F^{-1}\left(\frac{k-0.5}{n}\right)$$

- plot values of x_1,…, x_n against q_1,…,q_n

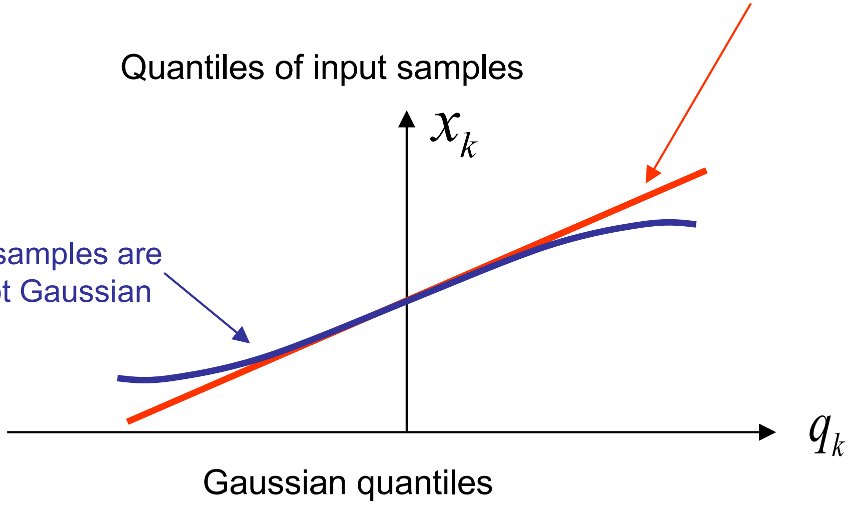
(or the other way)



Plotnorm

(qqplot.m MATLAB)

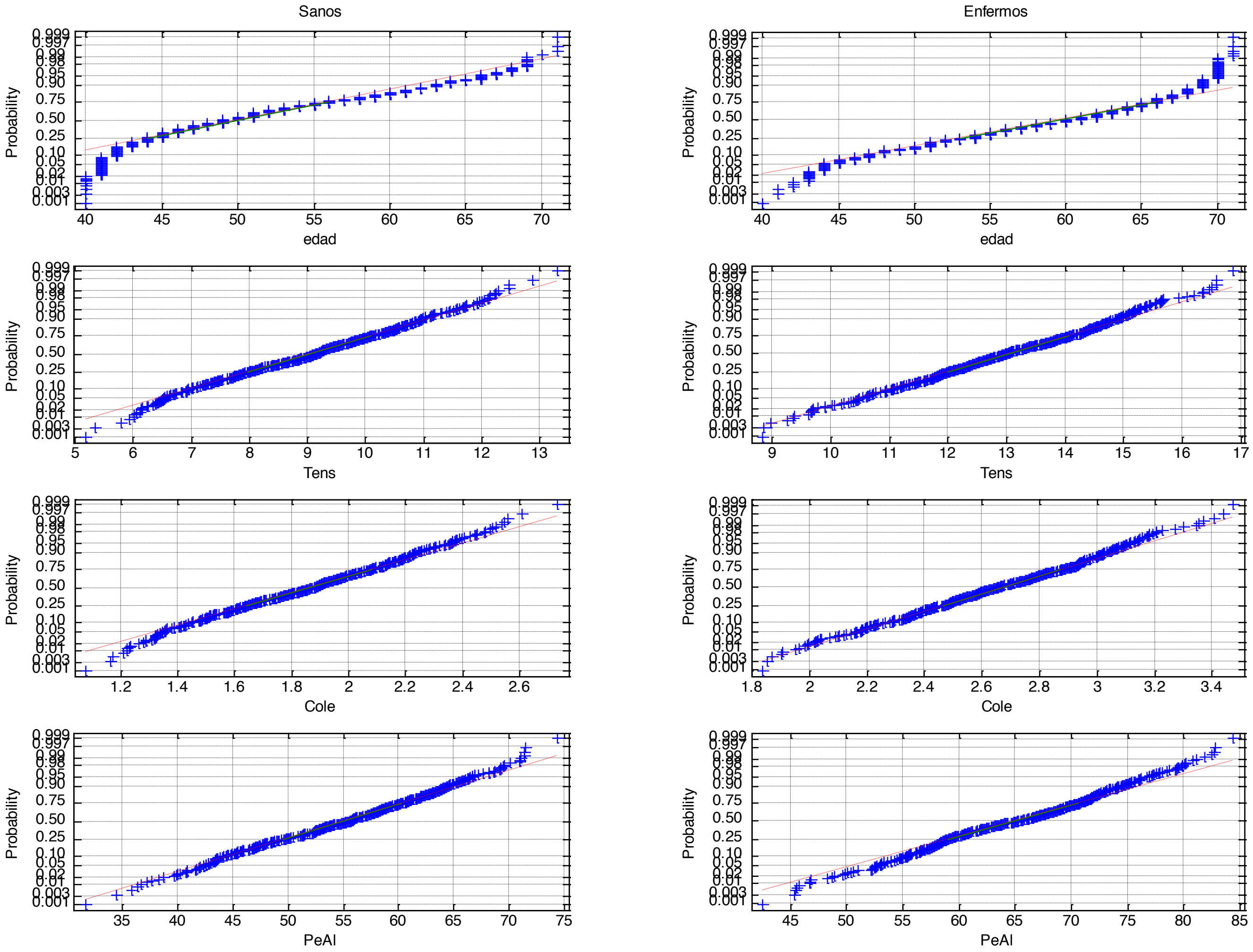
If samples follow a Gaussian disstribution we should get a line



if samples are not Gaussian

Example

Norm Plot:
representation of quantiles



Confidence interval for the mean

Let us assume the samples have been randomly selected from a random process with unknown parameters (mean and variance).

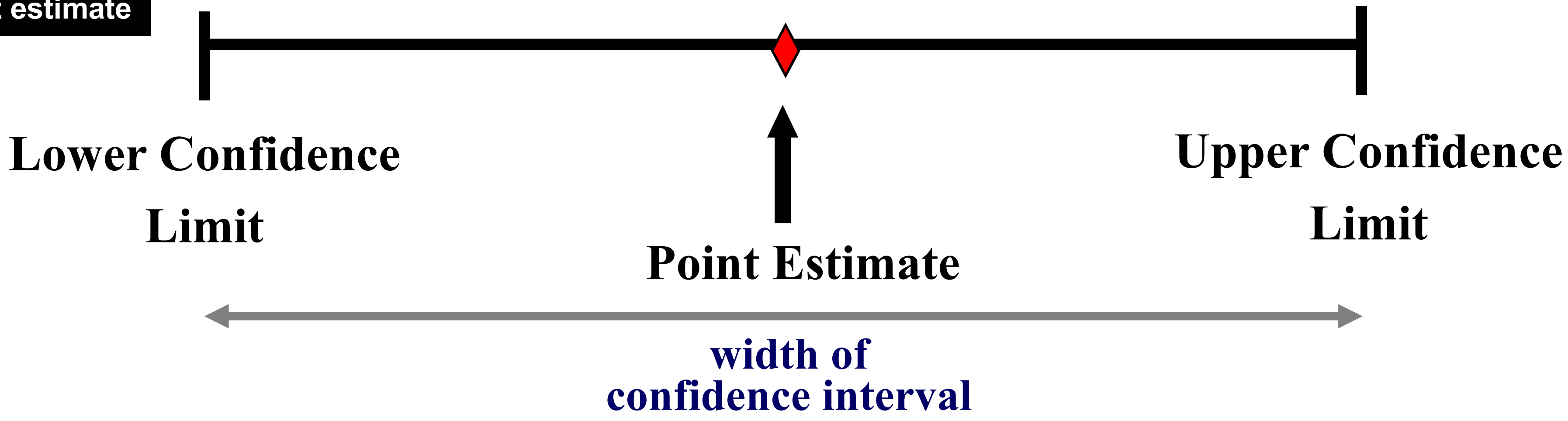
We want to estimate the actual population mean μ_0 but we can only get the sample mean $\overline{x}$: find a range of values that we can be really confident contains μ_0

The value of a single sample statistic: mean or proportion

A **point estimate** is a single number. For the population mean and population standard deviation, a point estimate is the sample mean and sample standard deviation.

A range of numbers constructed around the point estimate

A **confidence interval** provides additional information about variability



Lower Confidence Limit

Point Estimate

Upper Confidence Limit

width of confidence interval

Confidence interval for the mean

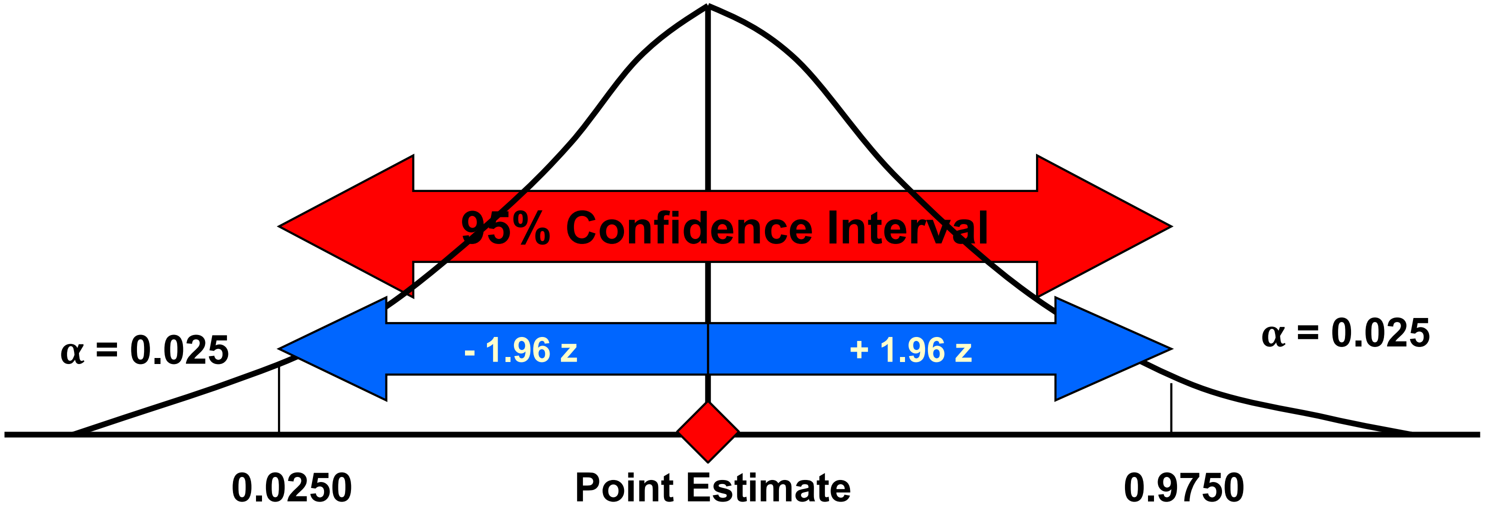
A confidence interval gives a range estimate of values:

Takes into consideration variation in sample statistics from sample to sample
Based on all the observations from one sample
Gives information about closeness to unknown population parameters
Stated in terms of level of confidence

Example: **95%** confidence, **99%** confidence
Can **never** be **100%** confident

Confidence Level confidence in which the interval will contain the unknown parameter ($\boldsymbol{\mu_o}$) a percentage (less than 100%).

The ***level of significance***, or "**α**" is the chance we take that the true population parameter is not contained in the confidence interval. Therefore, a 95% confidence interval would have an "**α**" of 5%



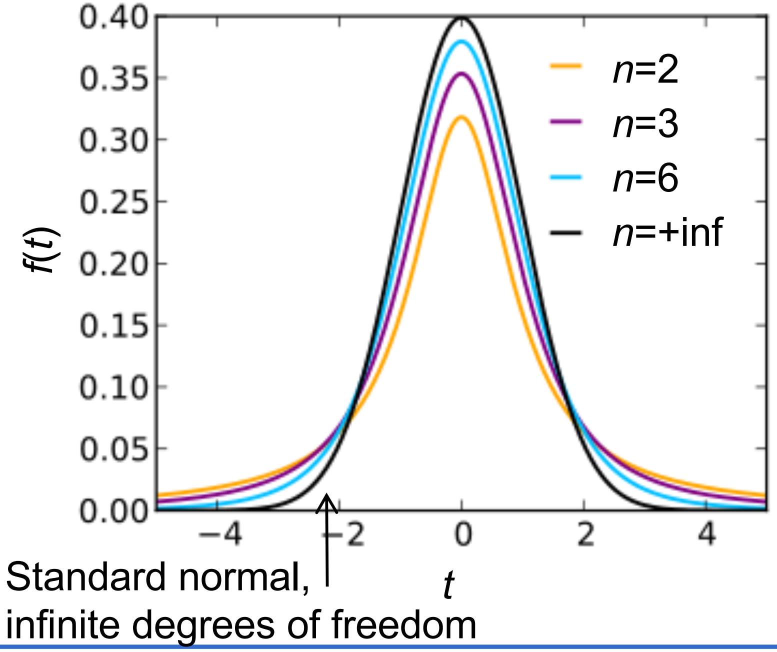
Confidence interval for the mean

Let us assume the samples have been randomly selected **from a normal random process with unknown parameters (mean and variance)**.

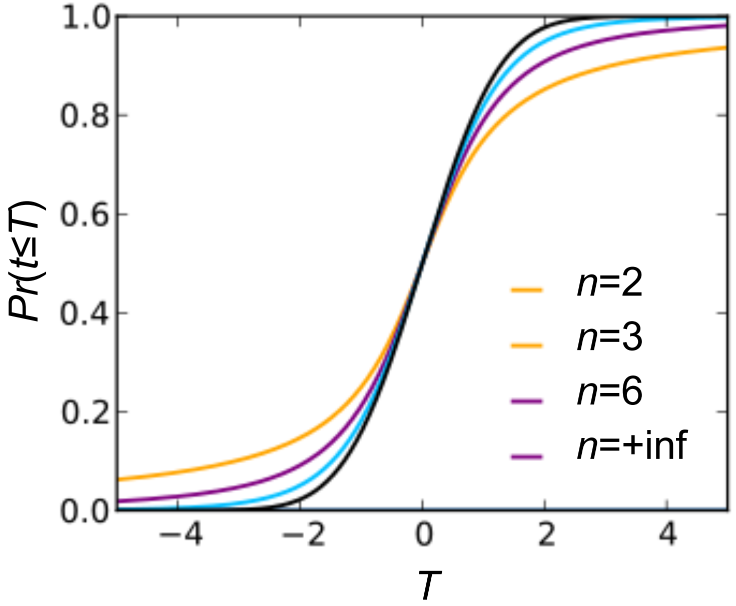
If the population standard deviation **σ** is unknown, we can substitute the **sample standard deviation** (**s**). This introduces extra uncertainty, since ‘**s**’ is variable from sample to sample.

Given a sample, we compute the statistic *t* , which follows a *t*-student distribution with *n*-1 degrees of freedom…

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}; \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$



Standard normal, infinite degrees of freedom



t follows a *t*-student distribution with *n*-1 degrees of freedom $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

For a level of significance $\alpha = 0.05$

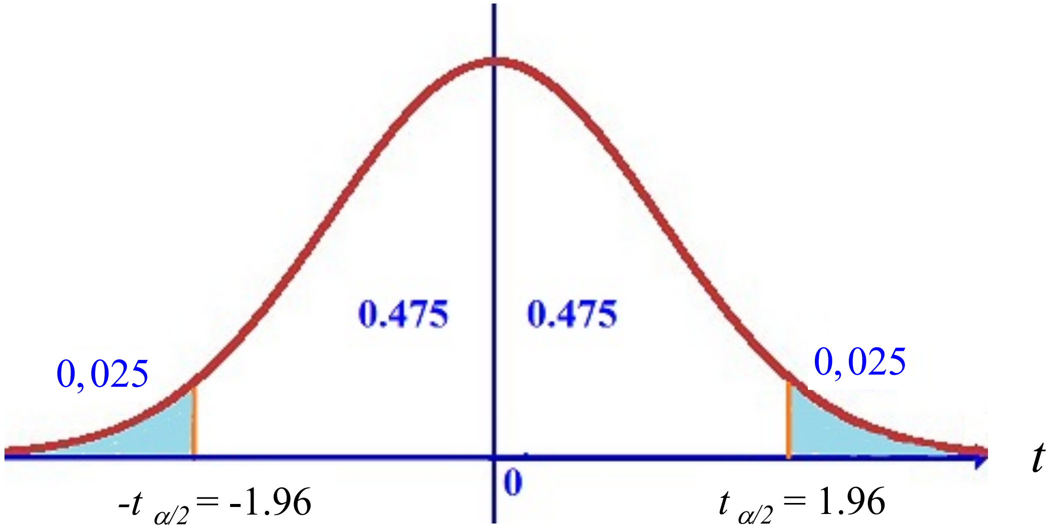
Let $t_{\alpha/2}$ be the value such that

$$\frac{\alpha}{2} = \Pr\{t \le -t_{\alpha/2}\} = \Pr\{+t_{\alpha/2} \le t\}$$

$$P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$$

Then, $$P\left(-t_{\alpha/2} \le \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \le t_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu_0 \le \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$



Then we can state that the true mean value μ_0 is within this interval

$$\left(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

with (1-α)%=95% confidence.

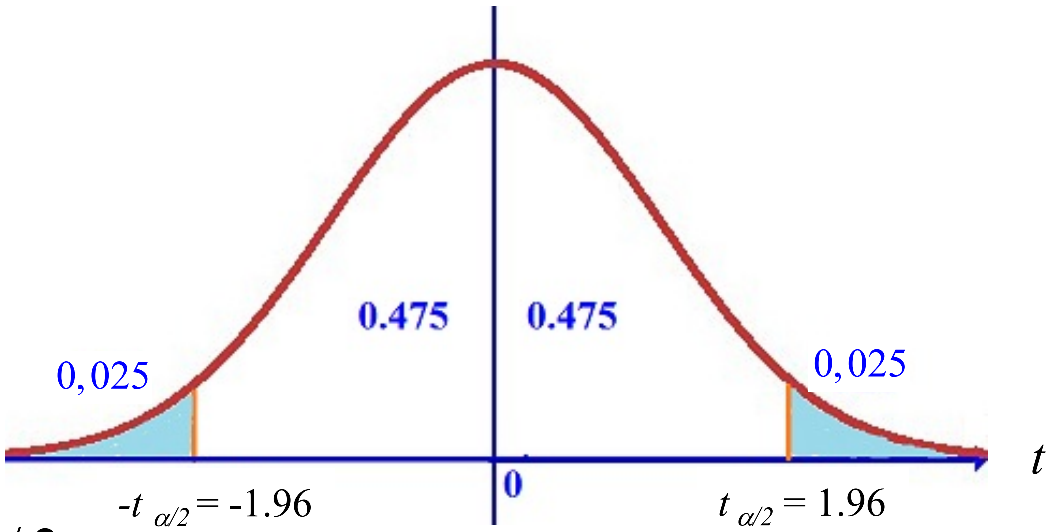
How do we compute it in practice?

For a confidence level $100 \times (1-\alpha)$

1. Define the level of significance α

2. Compute

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$



3. Find $t_{\alpha/2}$ such that $P\left(t \le -t_{\alpha/2}\right) = \alpha/2$

```
Matlab: tinv(p,df)
where p=1-α/2
      df = n-1  (degrees of freedom)
```

4. The confidence interval is $\left(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$

Hypothesis testing

- A **statistical hypothesis** is an assertion or conjecture concerning one or more populations.
- To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population.
- Instead, **hypothesis testing** concerns on how to use a random sample to judge if there is evidence that supports or not the hypothesis.

- Hypothesis testing is formulated in terms of two hypotheses:

  H_0: the null hypothesis (initial assumption)

  H_1: the alternate hypothesis

- So, there are two possible outcomes:

  - Reject H_0 (and accept H_1) because of insufficient evidence in favor of H_0

  - Do not reject H_0 because of insufficient evidence to support H_1

- **Important!** Note that failure to reject H_0 does not mean the null hypothesis is true. It only means that we do not have sufficient evidence to support H_1.

Example:

- In a jury trial the hypotheses are:
  - H_0 (defendant is innocent);
  - H_1 (defendant is guilty)

- H_0 (innocent) is rejected if H_1 (guilty) is supported by evidence beyond "reasonable doubt".
- Failure to prove H_1 (guilt) does not imply innocence, only that the evidence is insufficient to reject H_0.

- Two methods are defined for hypotesis testing:
  1. **Critical value:** determine a threshold over a statistics of our data
  2. **p-value:** probability of having more extreme values for our statistics than the observed one

- Because we are making a decision based on a finite sample, there is a possibility that we will make mistakes. The possible outcomes are:

	H0 is true	**H1 is true**
Do not reject H0	Correct decision	Type II error (β) **(minor mistake)**
Reject H0	Type I error (α) **(big mistake)**	Correct decision

- The acceptance of H_1 when H_0 is true is called a **Type I error**. Failure to reject H_0 when H_1 is true is called a **Type II error**.

$$\underbrace{\alpha = \Pr\left\{\text{Decide } H_1 \middle| H_0\right\}}_{\text{Significance level}} \qquad \beta = \Pr\left\{\text{Decide } H_0 \middle| H_1\right\}$$

Example: Type I error - convicting the defendant when he is innocent!

- The lower the significance level is α , the less likely we are to commit a type I error. **Generally, we would like small values of** α , typically 0.05 or less.

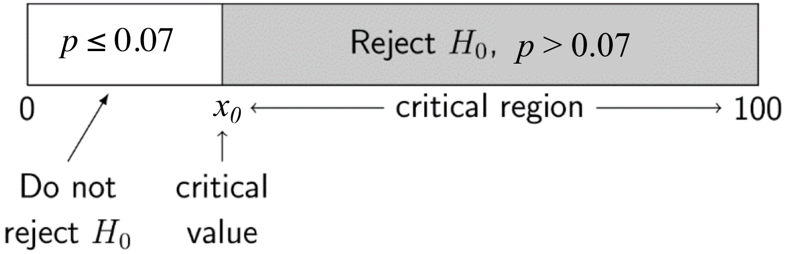
Hypothesis testing based on critical value

Case study…

A company manufacturing RAM chips claims the defective rate of the population is lower than 7%. Let *p* denote the true defective probability. We want to test if:

$$H_0: p \leq 0.07 \quad H_1: p > 0.07$$

- We are going to use a sample of 100 chips from the production to test the claim.



- Let *X* denote the number of defective chips in the sample of 100.
- Reject H_0 if $X \geq x_0$ (chosen "arbitrarily" in this case). *X* is called the ***test statistic***.

- **How to find a critical value to compare *X* for a desired level of significance?**

In this example, the density function is binomial: $\Pr\left\{X = k \middle| H_0\right\} = \begin{pmatrix} 100 \\ k \end{pmatrix} p^k (1-p)^{100-k}$

We have to evaluate one of the two equivalent expressions:

$$1-\alpha = \Pr\left\{X \le x_0 \middle| H_0\right\} = \sum_{X=0}^{x_0} \begin{pmatrix} 100 \\ X \end{pmatrix} p^X (1-p)^{100-X}$$

$$\alpha = \Pr\left\{X > x_0 \middle| H_0\right\} = \sum_{X=x_0+1}^{100} \begin{pmatrix} 100 \\ X \end{pmatrix} p^X (1-p)^{100-X}$$

If the level of significance is α = 0.05, for p = 0.07,

$$\Pr\left\{X > 10 \middle| H_0\right\} = 0.0908$$

$$\Pr\left\{X > 11 \middle| H_0\right\} = 0.0469$$

x_o=11, and hence X > 11 implies rejection of H_0 with 95,31% of certainty (or 4,69% of error).

For N=500, $X \geq$ 45 implies rejection of H_0

For N=1000, $X \geq$ 84 implies rejection of H_0

Hypothesis testing through the p-value

- The **p-value** is the probability (calculated assuming H_0 is true) of obtaining a test statistic value at least as contradictory to H_0 as the value obtained for the sample

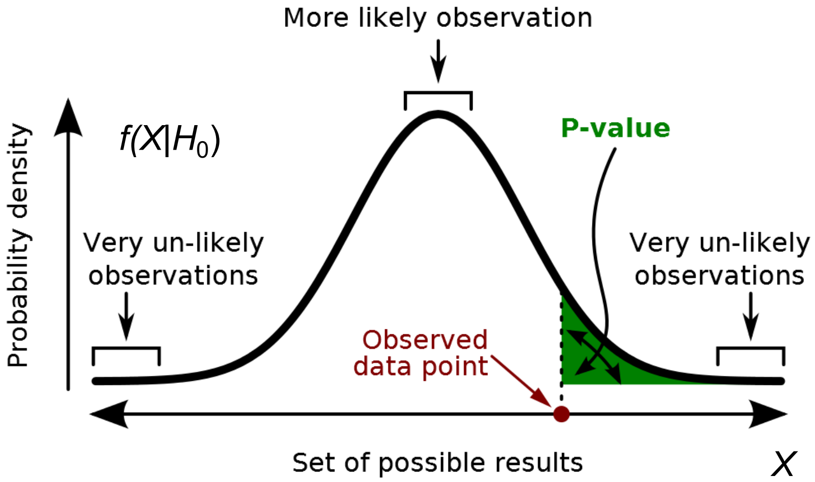
- That is: **the probability, assuming H_0, of obtaining a result equal to or more extreme than what was actually observed.** We want to compare it to the probability of rejecting H_0 if H_0 were true.

- Define a test statistic X. For the given data, the **value of the test is *d***. Assume H_0 is true. Then calculate the probability of observing values of X at least as extreme as d, given that H_0 is true

Thus,
$$\text{p-value} = \Pr(X \geq d \mid H_0)$$
$$(\text{or } \Pr(X \leq d \mid H_0))$$

If $\text{p-value} \leq \alpha$ **then reject H_0 ,**

else, do not reject H_0



The p-value can be interpreted as the smallest level α at which the observed data are significant.

Example

- In the previous example, at a significance level $\alpha = 0.05$, if the number of defective pieces is x_0=11, then the p-value is $P(X>x_0)=0.0469 < 0.05$ Therefore, we can reject H_0

- Suppose that, for a given hypothesis test, the p-value is 0.09. Can H_0 be rejected?
- It depends!
  - At a significance level $\alpha = 0.05$, we cannot reject H_0 because p-value = 0.09 > 0.05
  - However, for significance levels greater or equal to 0.09, we can reject H_0

Case study: fitness for Gaussian distribution

Chi-squared test computed from a sample of size *n*

Hypothesis: **H_0: the distribution is Gaussian** **H_1: not Gaussian**

1. Full range of *n* sample values divided into *k*-bins
2. Assume H_0: n_i probability of samples falling in bin *i*, π_i prob. for a Gaussian distribution

$$H_0 : \{n_1 = n\pi_1, n_2 = n\pi_2, \ldots, n_k = n\pi_k\}$$

3. Test statistic definition…

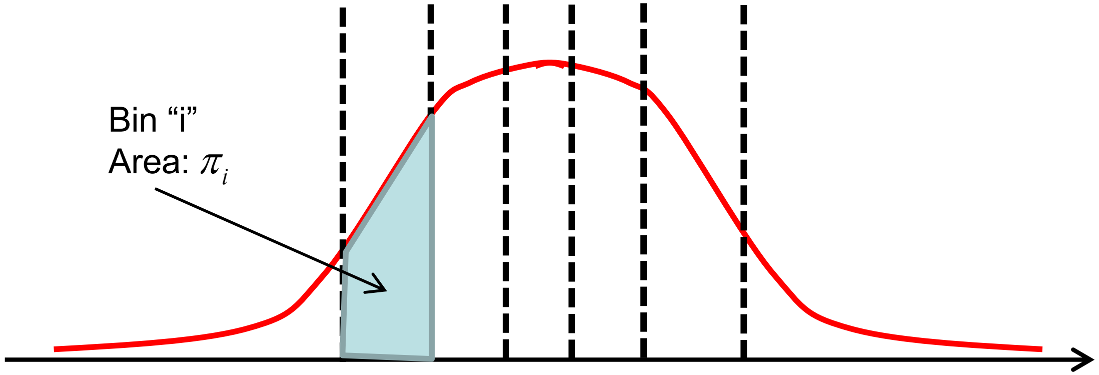
$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

n_i number of samples in bin *i*

it has a chi-squared distribution of *k*-3 degrees of freedom

4. Compute p-value $P(X^2 > d \mid H_0)$

_d_ is the value of the test statistic computed from the data



5. Reject H_0 if $p - value \leq \alpha$ for a significance value α . In other words: if samples are Gaussian, we should expect X^2 to be small. If it is large, H_0 is rejected.

`Matlab: chi2gof`

Additional bibliography

- Confidence intervals and hypothesis tests
  - https://newonlinecourses.science.psu.edu/statprogram/reviews/statistical-concepts
  - https://cnx.org/contents/MBiUQmmY@22.8:IWGQ0U41@8/Introduction