Practica 0

Test de gaussianidad

Departament de Teoria
del Senyal i Comunicacions

UPC

INDICE

0.1 Observaciones preliminares

0.2 Base de datos de ejemplo: EnfX

0.3 Gaussianidad a partir del histograma

0.4 Gaussianidad a partir de los momentos

0.5 Gaussianidad a partir de los cuantiles

0.6 Test de hipótesis

0.7 Test de hipótesis mediante el p-value

0.8 La base de datos PRBB_Brain

1 Observaciones preliminares

Al disponer de una base de datos (BD) en general es conveniente conocer la distribución de sus componentes, y en particular si dichas distribuciones son o no gaussianas.

Gaussianidad

Análisis a realizar:

- Obtención de Histograma y comparación con la f.d.p de una gaussiana.
- Obtención del Histograma acumulado y comparación con la función de distribución de una gaussiana.
- Obtención de momentos: Media, Varianza, Skewness y Kurtosis
- Obtención de gráficas de tipo normplot o qqplot.
- Cálculo de Intervalos de Confianza al estimar determinados parámetros de la distribución. Con un determinado nivel de confianza decidir si la distribución de una determinada muestra (o característica) es o no es gaussiana.

Adicionalmente el Scatter Plot es útil para obtener información gráfica sobre la "separabilidad" de las clases en función de las características.

Separabilidad

2 Bdatos de ejemplo: EnfeX

Objetivo: Diagnosticar si una persona posee la enfermedad X

Se forma una **Base de Datos** analizando N=1000 pacientes (sanos y enfermos).

Para cada paciente se registra un **vector de dimensión d=4:**

- Edad: $40 \leq v(1) \leq 70$
- Tensión Arterial Media: $6 \leq v(2) \leq 16$
- Nivel de Colesterol: $1 \leq v(3) \leq 3{,}5$
- Peso(Kg)/Altura(mts): $30 \leq v(4) \leq 80$

Para cada paciente se registra una **etiqueta o Label = 1(Enfermo)**
0(Sano)

Ejemplo:

Edad	Presión Arterial	Colesterol	Peso/Altura
50	12	2	35

Esta BD se toma como ejemplo para ilustrar las estrategias propuestas de análisis de Gaussianidad y de análisis de separabilidad.

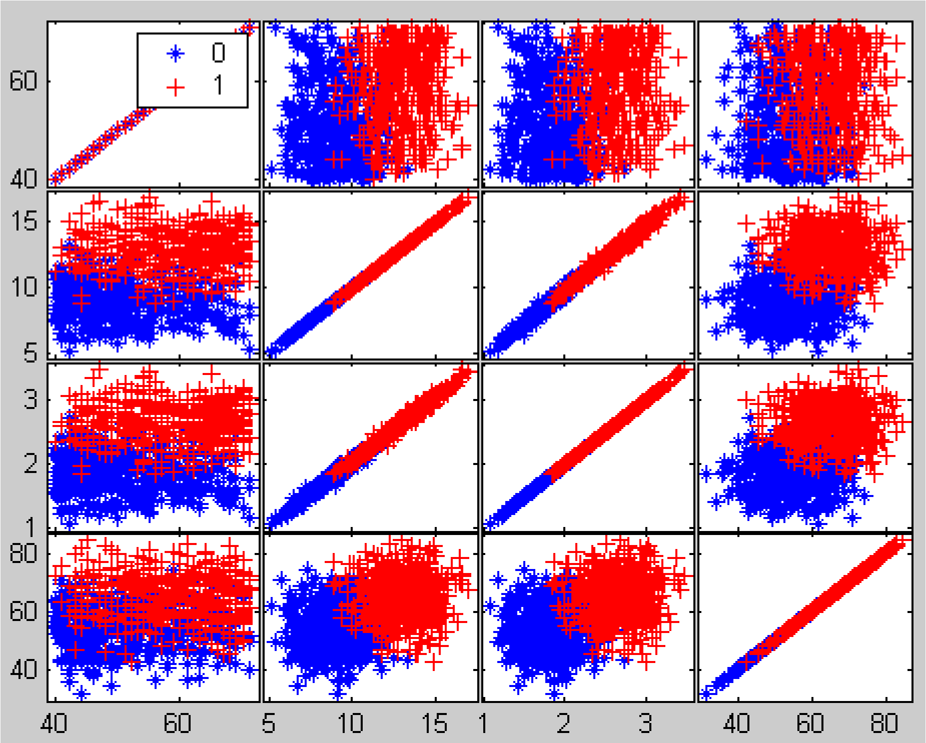
El **scatter plot** muestra todos los elementos de la base de datos por pares de características en una única figura.

Edad

Tensión arterial

Colesterol

Peso/Altura



Edad TA Colesterol P/A

Como diagnosticar si una persona posee la enfermedad X o tiene riesgo de adquirirla (hipótesis H_1)?

$$p(H_1 | x(1), x(2), \ldots, x(d)) \underset{H_1}{\overset{H_2}{\lessgtr}} p(H_2 | x(1), x(2), \ldots, x(d))$$

Criterio MAP

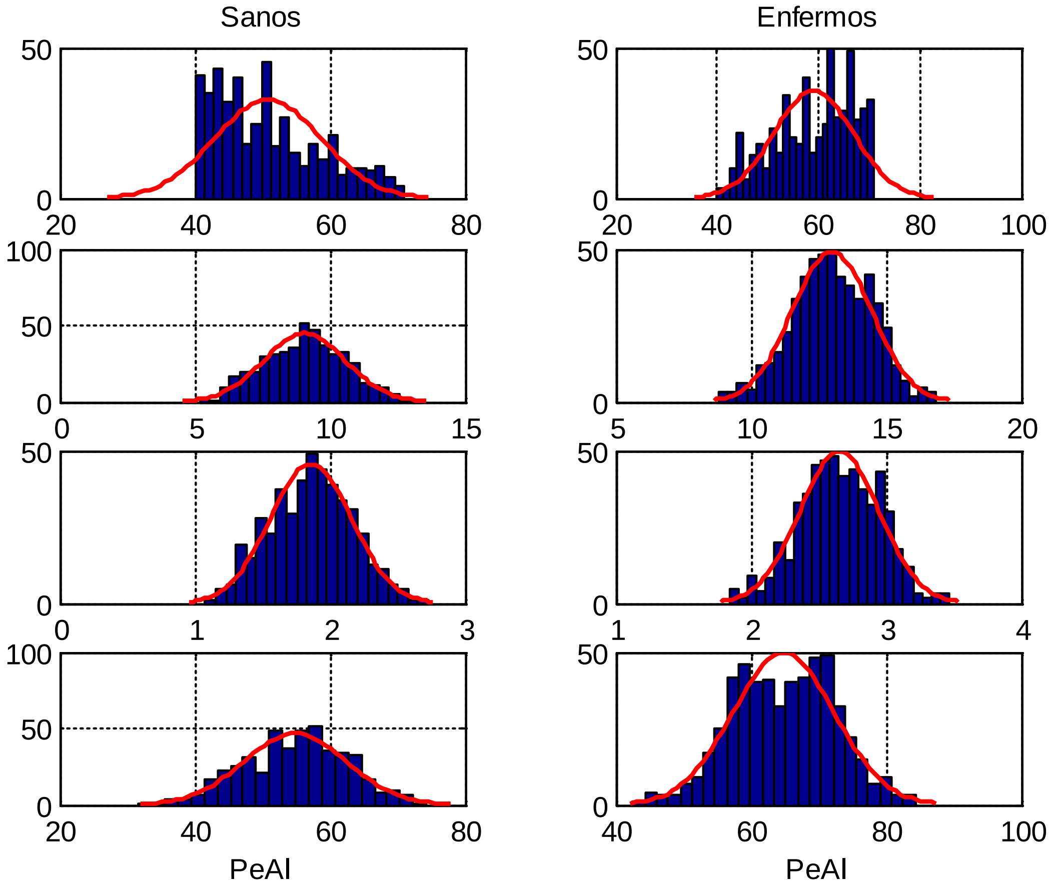
$$p(H_1 | x(1), x(2), \ldots, x(d)) =$$

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

¿Son las funciones $f(.|H_0)$ y $f(.|H_1)$ Gausianas?
Si lo son, ya podemos aplicar las técnicas del tema 2.1

3 Gaussianidad del histograma

Histograma
Cuenta el número de realizaciones para cada margen de valores. Es una forma de calcular la función de densidad de probabilidad de la variable aleatoria.



Se obtiene un desajuste grande para la primera característica (edad)

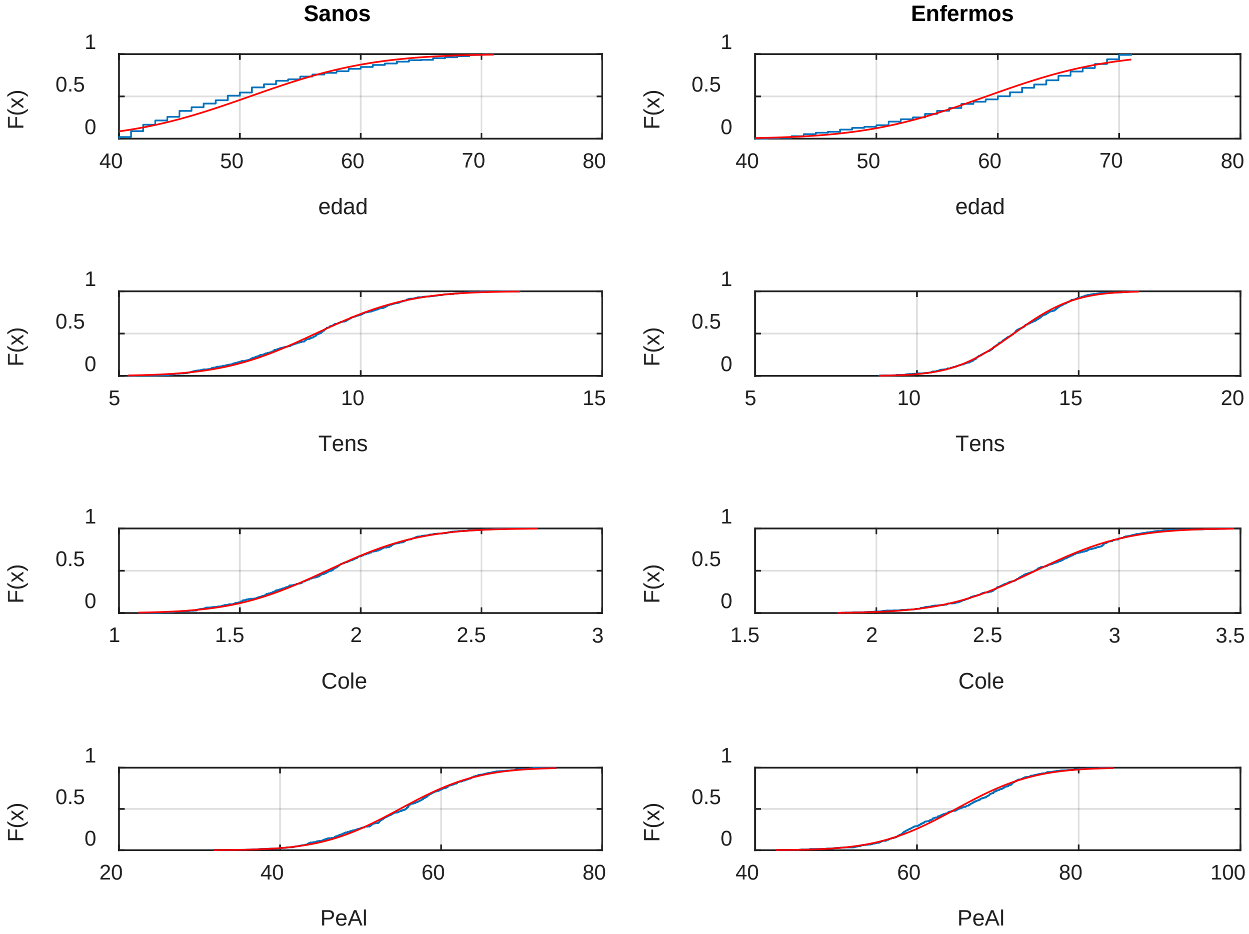
Test de Gaussianidad de cada componente

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

Cumulative Density Function
Coincide con la suma acumulada del histograma

Azul: cdf de las características.

Rojo: ajuste gaussiano



Se obtiene un desajuste grande para la primera característica (edad)

4 Gaussianidad a partir de los momentos

Mean

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

Variance

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

Skewness

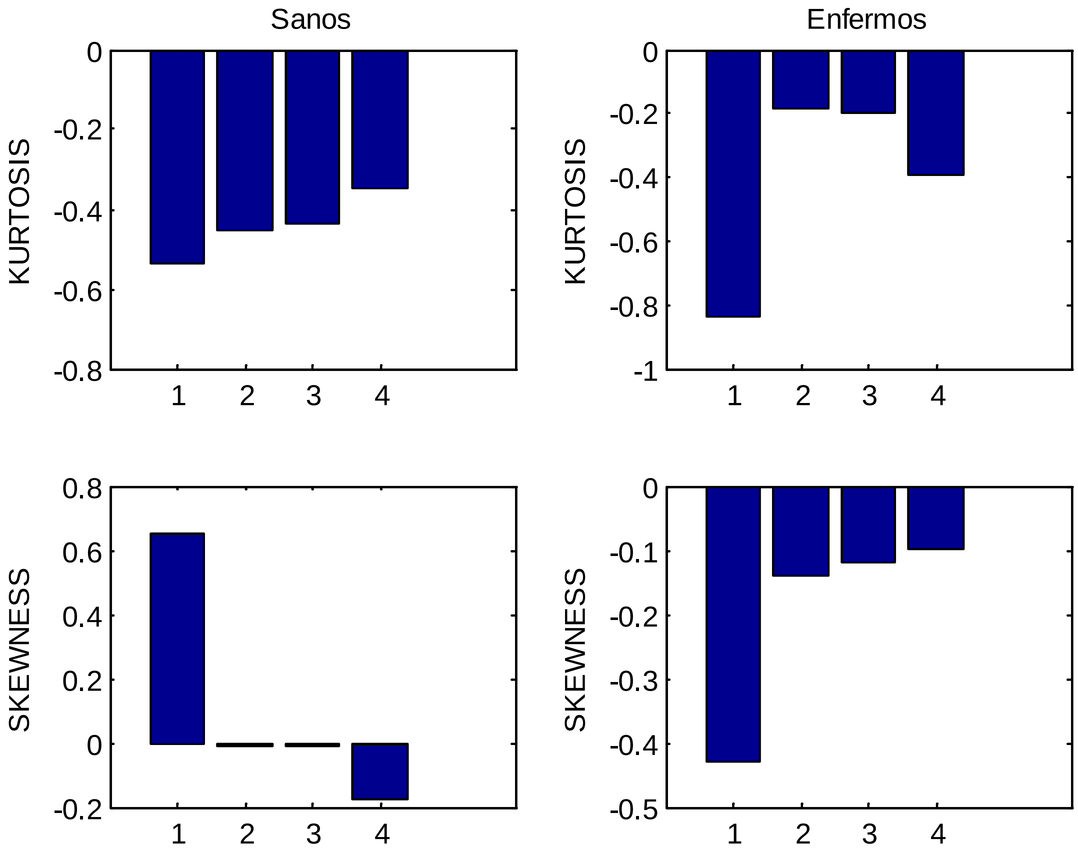
$$Sk(x) = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{E\left[\left(x-\mu\right)^3\right]}{\sigma^3}$$

Kurtosis

$$K(x) = \frac{\mu_4}{\left(\mu_2\right)^2} - 3 = \frac{E\left[\left(x-\mu\right)^4\right]}{\sigma^4} - 3$$

Si x es gaussiana entonces *Sk*(*x*)=0 y *K*(*x*)=0 (nótese que no son condiciones suficientes para concluir gaussianidad).

Ejemplo

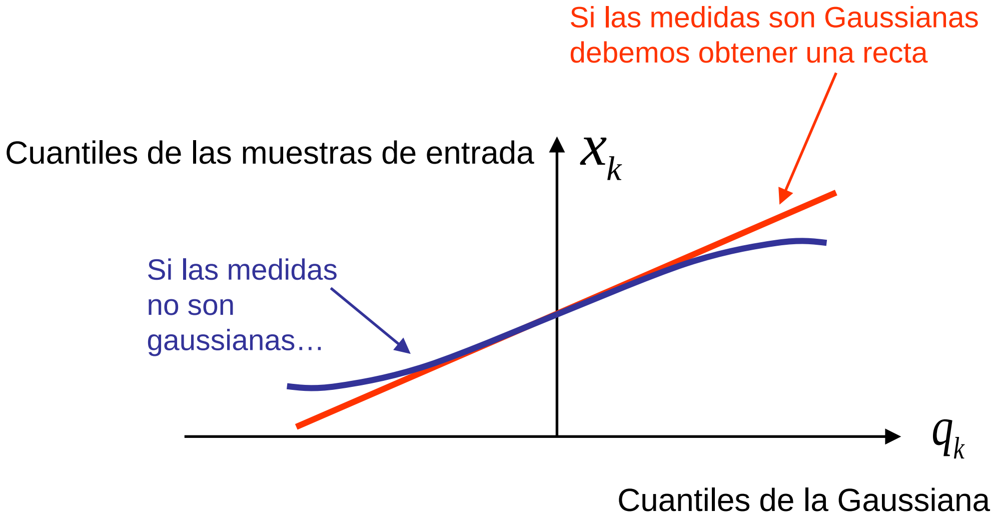

Sanos
Enfermos
KURTOSIS
0
-0.2
-0.4
-0.6
-0.8
1
2
3
4
KURTOSIS
0
-0.2
-0.4
-0.6
-0.8
-1
1
2
3
4
SKEWNESS
0.8
0.6
0.4
0.2
0
-0.2
1
2
3
4
SKEWNESS
0
-0.1
-0.2
-0.3
-0.4
-0.5
1
2
3
4

5 Gaussianidad a partir de los cuantiles

Norm Plot: Representación de los cuantiles

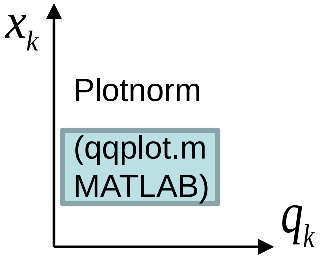
$$p_k = \frac{k-0.5}{n} \cong F(q_k) = \int_{-\infty}^{q_k} \frac{1}{\sqrt{2\pi}} e^{-\frac{\lambda^2}{2}} d\lambda = \begin{cases} Q(-q_k) & q_k < 0 \\ 1-Q(q_k) & q_k > 0 \end{cases}$$

$$p_k = \int_{-\infty}^{x_k} f(\lambda) d\lambda$$


Cuantiles de las muestras de entrada
x_k
q_k
Cuantiles de la Gaussiana

Si las medidas son Gaussianas debemos obtener una recta

Si las medidas no son gaussianas…

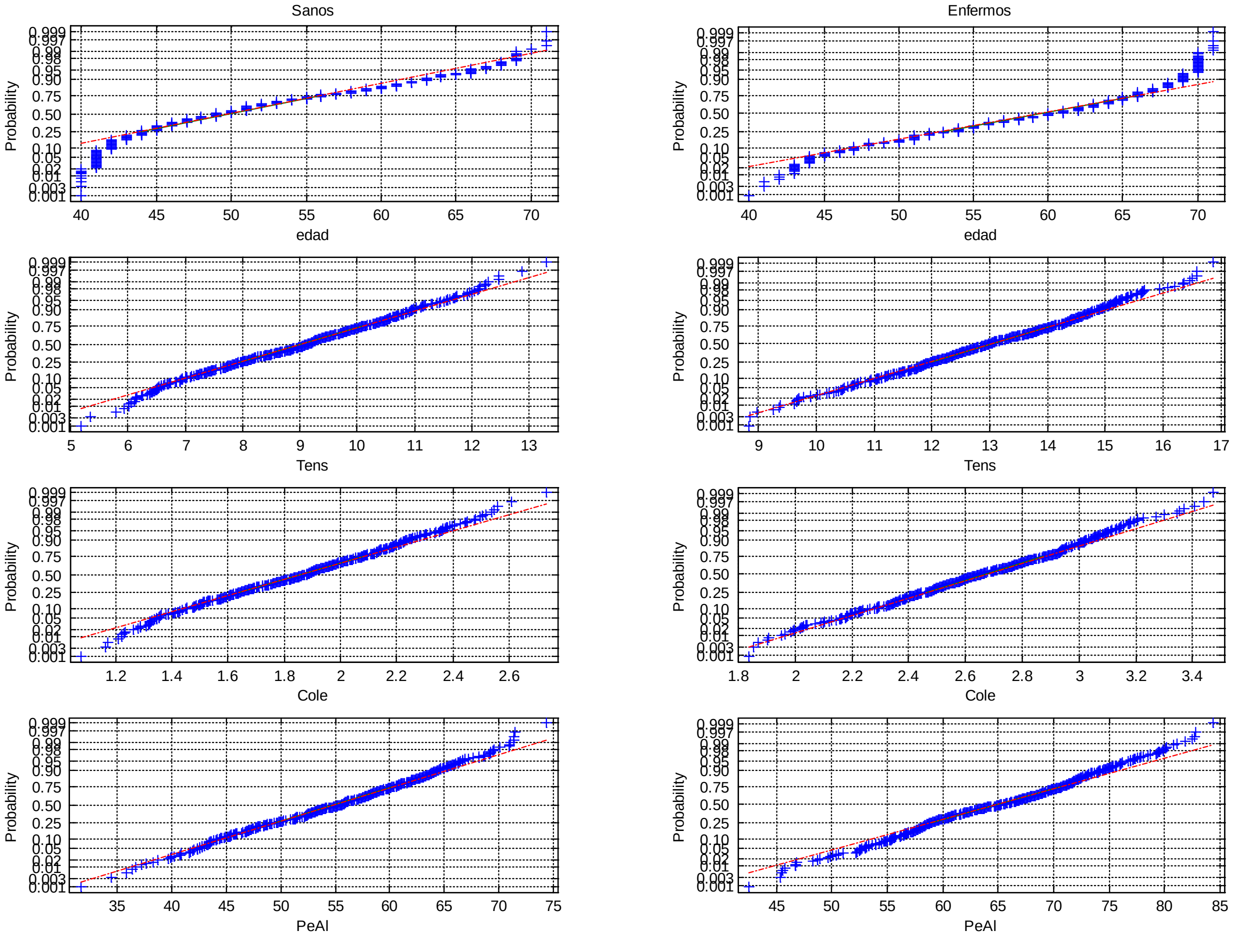


Plotnorm

(qqplot.m MATLAB)

Ejemplo:

Norm Plot: Representación de los quantiles



6 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

  H_1: the alternate hypothesis

- So, there are two possible outcomes:

  - Reject H_0 (and accept H_1) because of insufficient evidence in the sample 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.

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

	H_0 is true	H_1 is true
Do not reject H_0	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	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!

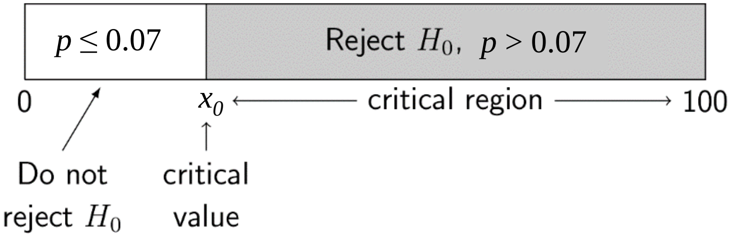
- The lower the significance level α is, the less likely we are to commit a type I error. **Generally, we would like small values of α**, typically 0.05 or less.

Case study 1

A company manufacturing RAM chips claims the defective rate of the population is 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.



- Let *X* denote the number of defective 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?**

Evaluate one of the two equivalent expressions:

$$1-\alpha = \Pr\left\{X \le x_0 \left| H_0 \right.\right\} = \sum_{X=0}^{x_0} \binom{100}{X} p^X (1-p)^{100-X}$$

$$\alpha = \Pr\left\{X > x_0 \left| H_0 \right.\right\} = \sum_{X=x_0+1}^{100} \binom{100}{X} p^X (1-p)^{100-X}$$

In this example, the density function is binomial: $\Pr\left\{X = k \left| H_0 \right.\right\} = \binom{100}{k} 0.07^k (1-0.07)^{100-k}$

If the level of significance is $\alpha = 0.05$, for $p = 0.07$,

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

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

And hence $X > 11$ implies rejection of H_0 with 95,31% of certainty (or 4,69% of error).
Equivalently, $X \le 11$ implies acceptance of H_0 with 95,31% of certainty (or 4,69% of error).

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

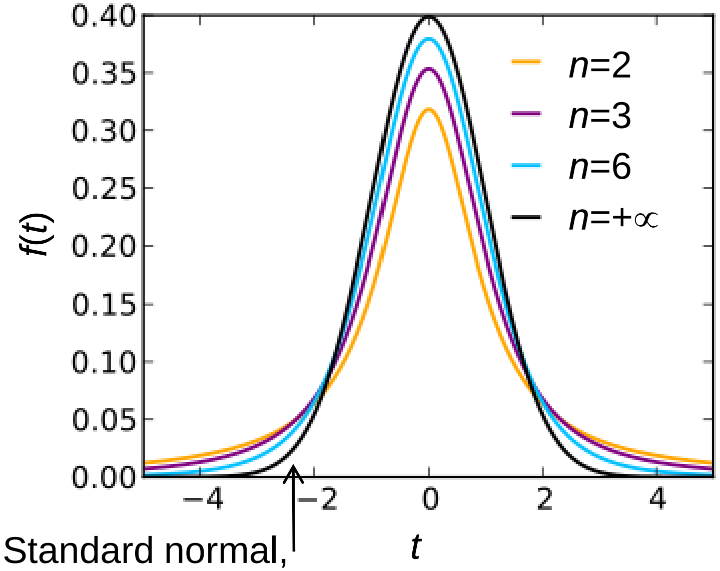
For N=1000, $X \ge 84$ implies rejection of H_0

Case study 2: significance level for the sample average

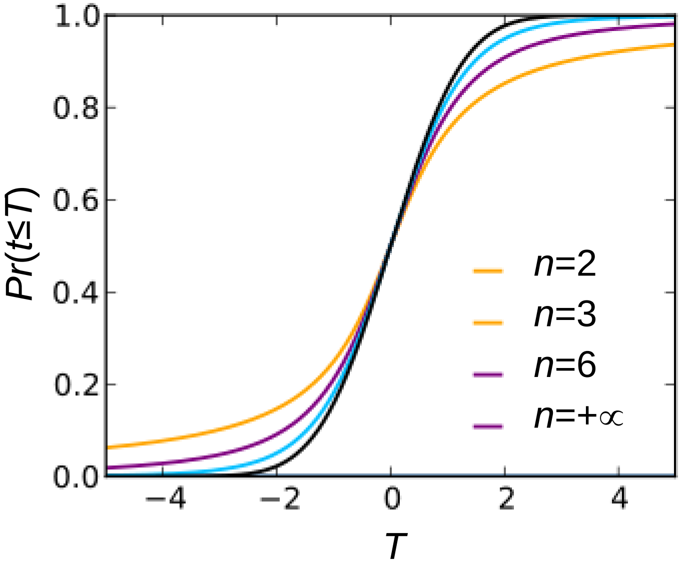
Define the hypothesis testing as $\mathrm{H}_0 : \mu = \mu_0 \quad \mathrm{H}_1 : \mu \neq \mu_0$

Let us assume the samples have been randomly selected from a normal random process with unknown parameters (mean and variance). Under hypothesis H_0, the standardized variable *t* follows a *t*-student distribution with *n*-1 degrees of freedom…

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



Standard normal,
infinite degrees of freedom



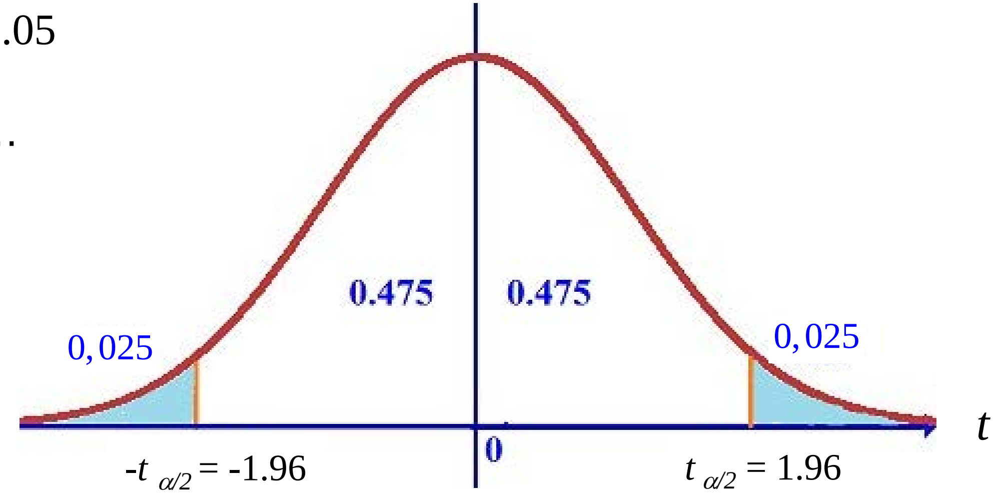
By operating with the expression of *t:* $\bar{x} \in \left(\mu_0 - t_{\alpha/2} \frac{s}{\sqrt{n}}, \mu_0 + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$

where the significance level is defined as $\frac{\alpha}{2} = \Pr\{t \leq -t_{\alpha/2}\} = \Pr\{+t_{\alpha/2} \leq t\}$

Then, we cannot reject with (1-α)% confidence H_0 if $\bar{x}$ lies within the interval.

A typical value is $\alpha = 0.05$

For a given value of *n*…



By reformulating the equation above, we can also state that the true mean value μ_0 is within the following interval

$$\mu_0 \in \left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

with (1-α)% confidence.

7 Hypothesis testing through the p-value

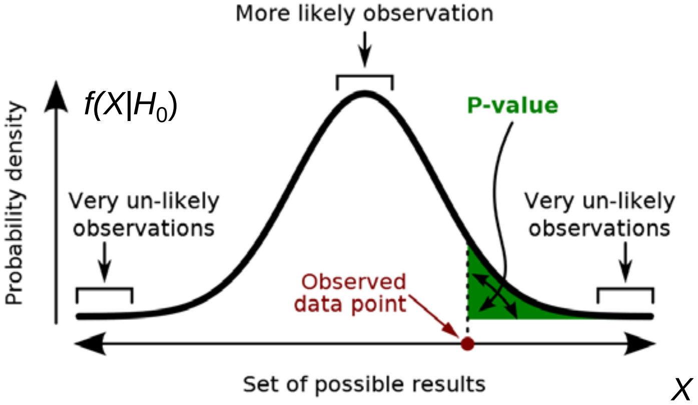
- The **p-value** (or observed significance value) is the probability (calculated assuming H_0 is true) of obtaining a test statistic value at least as contradictory to H_0 as the value obtained for the sample
- That is: the probability, assuming H_0, of obtaining a result equal to or more extreme than 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, $\text{p-value} = \Pr(X > d \mid H_0)$

$(\text{or } \Pr(X < d \mid H_0))$

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

else, do not reject H_0



Example

- Suppose that, for a given hypothesis test, the p-value is 0.09
- Can H_0 be rejected?
- 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

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

Case study 3: fitness for Gaussian distribution

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

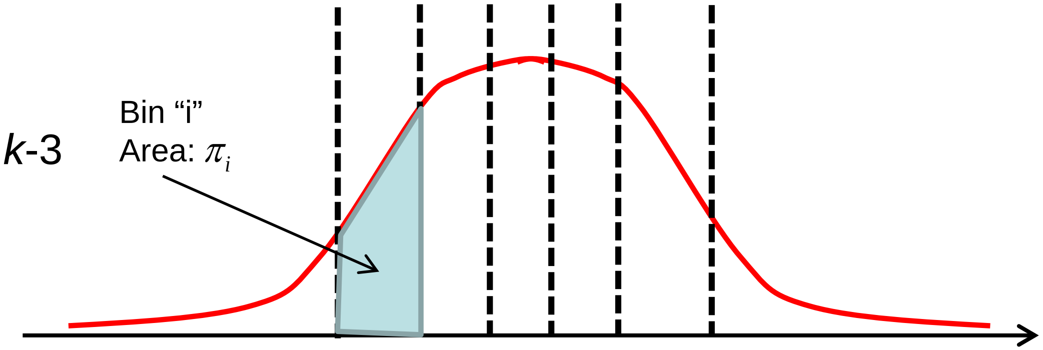
H_0: the distribution is Gaussian H_1: not Gaussian

1. Full range of *n* sample values divided into *k*-bins
2. Assume H_0: π_i probability of samples falling in bin *i*, π_{i0} prob. for a Gaussian distrib.
3. Test statistic definition…

$$H_0 : \pi_1 = \pi_{1,0}; \pi_2 = \pi_{2,0}; \ldots; \pi_k = \pi_{k,0}$$

$$X^2 = \sum_{i=1}^{k} \frac{\left(n_i - n\pi_{i,0}\right)^2}{n\pi_{i,0}}$$

it has a chi-squared distribution *df* = *k*-3



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

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

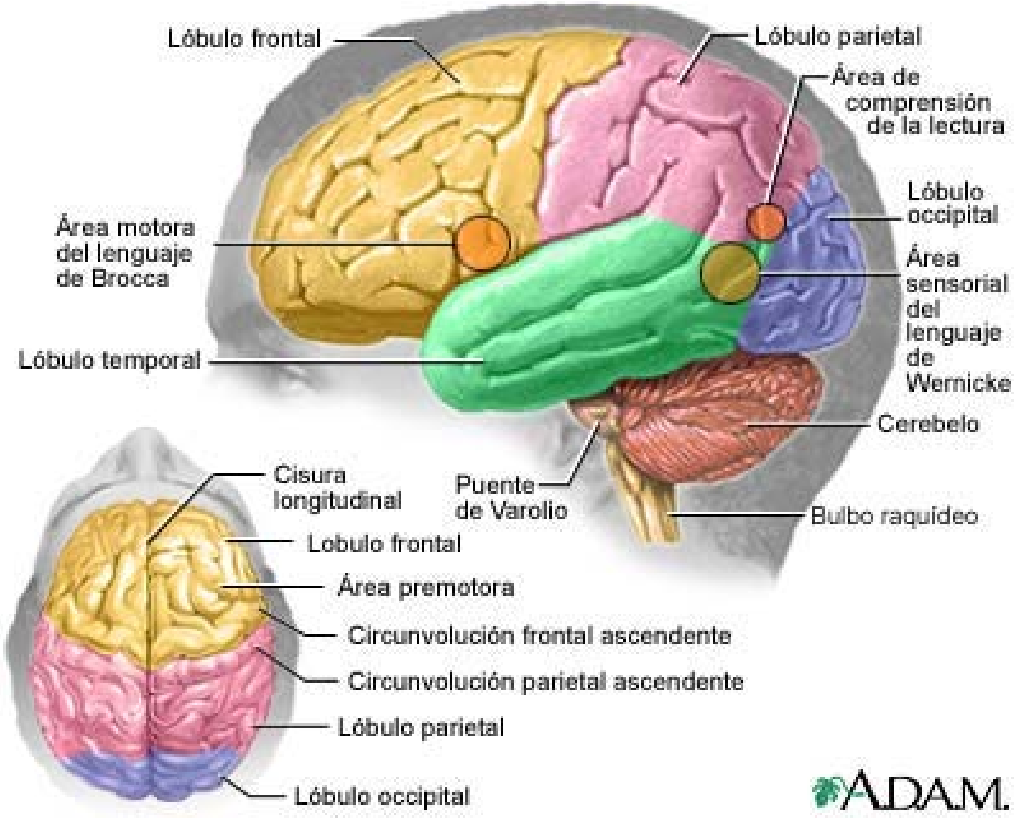
5. Reject H_0 if $p - value \le \alpha$ for a significance value α

`Matlab: chi2gof`

8 The database PRBB_Brain: The Human Brain

The Brain:

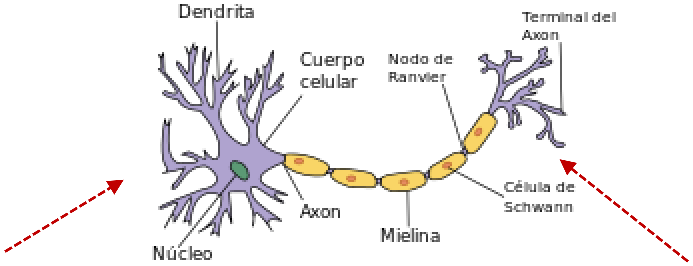
El cerebro es la parte antesuperior del encéfalo y el centro supervisor del sistema nervioso. Consta de la materia gris (parte superficial llamada corteza y el núcleo) y la materia blanca (partes profundas a excepción del núcleo). Las áreas principales del cerebro tienen una o más funciones específicas .



8 The database PRBB_Brain: The Human Brain

Brain Images:

El tejido llamado "**materia gris**" presente en el cerebro y en la médula espinal es también conocido como sustancia gris y está compuesto por cuerpos celulares. La "**materia blanca**" o sustancia alba está compuesta por fibras nerviosas.

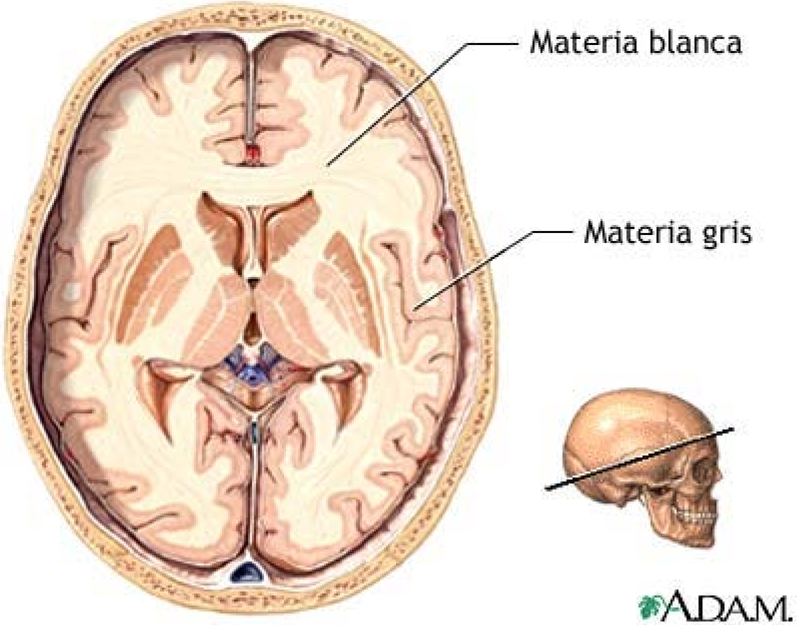


Materia Gris

(procesamiento de información o razonamiento)

Materia Blanca

(transmisión de información a otra célula nerviosa)



8 The database PRBB_Brain

The database is formed by 8 Images.
All of them correspond to a single human brain cut

- Images 1, 2, 3, 4: Obtained by Magnetic Resonance (MRI)
- Images 5: Positron emission tomography (PET)
- Images 6, 7, 8: Probabilities for each pixel to be in one of the three classes (Not used in Prac0):
  - White (Materia Blanca) Class 2
  - Grey (Materia Gris) Class 1
  - LCR (Líquido Cefaloraquídeo) Class 3
  - BackGround (Class 4) (Not used in Prac0):
- Each image has 256x256 pixels = 65536 = N patterns

  just N=10682 in Prac0
- Feature vectors are formed taking d=5 pixels. Each feature “j” in a vector is obtained from the same horizontal-vertical pixel at the “j” image.

$$\mathbf{X}_{\text{Base de Datos}} = \underset{\xleftrightarrow{\qquad d=5 \qquad}}{\begin{pmatrix} v_1(1) & v_1(2) & : & v_1(d) \\ v_2(1) & v_2(2) & : & v_2(d) \\ : & : & : & : \\ v_N(1) & v_N(2) & : & v_N(d) \end{pmatrix}} \updownarrow \quad N = 10682$$

3 The DataBase:
PRBB_Brain

Images 1:5