VNIVERSITAT DE VALÈNCIA

MASTER DE INGENIERÍA BIOMÉDICA.

Métodos de ayuda al diagnóstico clínico.

Tema 2: Probabilidad y estadística

Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE
Sistemas de Ayuda a la Decisión Clínica, Curso 2009-2010

Objetivos del tema

Dar a conocer los conocimientos mínimos de probabilidad/estadística necesarios para aplicar procedimientos estadísticos a un conjunto de datos, sin incurrir en los errores más comunes. NO ES UN RESUMEN DE BIOESTADÍSTICA NI SE VA A DEMOSTRAR NADA

Conocer las ventajas y limitaciones que tienen estos métodos frente a otros más avanzados (redes neuronales, árboles de decisión, etc).

Conocer las implicaciones del Teorema de Bayes en clínica (¡¡ se suele utilizar a menudo !!)

Conocer las condiciones que se tienen que cumplir para realizar un determinado contraste de hipótesis.

Aprender lo que es un análisis de supervivencia.

Conocer el software que puedo usar para realizar un análisis estadístico

Probabilidad.

La Teoría de la Probabilidad analiza lo que se conoce como **experimentos aleatorios**; experimentos cuyo resultado no se conoce a priori pero que está limitado a un determinado conjunto de resultados conocido como **espacio muestral. Este espacio puede ser discreto o continuo.**

Otra definición importante es la de **evento o suceso** que es un subconjunto del espacio muestral.

Destacar que la **frecuencia relativa** de los resultados de los experimentos aleatorios, cuando se realizan un gran número de éstos en las mismas condiciones, sí es predecible; éste es el punto de apoyo para los ingenieros.

Esta frecuencia relativa de un evento A es lo que se conoce intuitivamente y viene definida por la siguiente expresión.

$$\frac{N_n(A)}{n}$$

Donde N (A) es el número de veces que ocurre el evento a sobre n experimentos. Evidentemente este cociente tiene como límites 0 (no ocurre nunca el suceso A) y 1 (ocurre siempre)

Cuando se considera que el número de experimentos tiende a ∞ y la frecuencia relativa, en ese caso converge a un valor; dicho valor se conoce como **probabilidad del evento A.**

Probabilidad.

Un ejemplo sería el lanzamiento de un dado; el espacio muestral sería $\{1,2,3,4,5,6\}$; *a priori* no se conoce el resultado del lanzamiento pero si se lanza muchas veces el dado la probabilidad de aparición de cualquier número es 1/6.

Estudiar un evento no tiene mucho sentido práctico; se analiza su frecuencia relativa y se determina si se le puede asignar una probabilidad; algo más practico (¡¡¡ y divertido!!!!) es considerar la probabilidad de la combinación de varios eventos diferentes.

Unión de dos eventos. Es el evento que consiste en todos los resultados contenidos en uno de esos dos eventos. Se representa por $\cup$.

Intersección de dos eventos. Es el evento que consiste en todos los resultados contenidos en los dos eventos. Se representa por $\cap$.

Complemento de un evento. Es el evento que consiste en todos los resultados no contenidos en dicho evento. Si E es es el evento representaremos su complementario por E(c)

A modo de ejemplo consideremos el lanzamiento de un dado. Definimos $E_1=\{4,5,6\}$; $E_2=\{2,4,6\}$.

Tendríamos entonces $E_1 \cup E_2 =\{2,4,5,6\}$; $E_1 \cap E_2 =\{4, 6\}$; $E_1(c)=\{1,2,3\}$; $E_2(c)=\{1,3,5\}$

Probabilidad.

Otra definición importante es el de **eventos mutuamente exclusivos**. Dos eventos son mutuamente exclusivos si no se pueden dar a la vez. En el lanzamiento de un dado los eventos A= {1,3,5} y B={2,4} son mutuamente exclusivos.

Lo comentado hasta ahora nos acerca a conceptos de la ***teoría de conjuntos***. De hecho es inmediato comprobar que dos eventos son mutuamente exclusivos si su intersección es cero. Existe una aproximación matemática a la probabilidad que no haría uso de las frecuencias relativas.

Axiomas de Probabilidad.

Una medida de probabilidad P[.] es una función que mapea eventos en un espacio muestral (S) a números reales cumpliéndose los siguientes axiomas.

Axioma 1. Para cualquier evento A, $0 \leq P[A] \leq 1$.

Axioma 2. $P[S]=1$.

Axioma 3. Si se tienen dos eventos, A y B, mutuamente exclusivos entonces $P(A \cup B)=P(A)+P(B)$

Probabilidad.

Los axiomas anteriormente mencionados son muy simples pero, a la vez muy potentes; a partir de ellos se puede demostrar lo siguiente

$P[\varnothing]=0$; $P[A(c)]=1-P(A)$

$P[A \cup B]=P[A]+P[B]-P[A \cap B]$

Si el evento A está incluido en B; denotado por $A \subset B$, entonces $P(A) \leq P(B)$

Si un evento B está formado por los eventos **elementales** s_i $1 \leq i \leq k$ $B=\{s_1, s_2, \ldots s_k\}$ entonces se tiene $P[B]=\Sigma P[s_i]$

Si se tiene una colección de eventos B_i $1 \leq i \leq k$ mutuamente exclusivos entre sí entonces el evento unión de todos ellos $B= B_1 \cup B_2 \cup \ldots B_k$ tiene como probabilidad la suma de las probabilidades de cada uno de ellos; $P[B]=\Sigma P[B_i]$

Independencia.

Se dice que dos eventos son **independientes** cuando la probabilidad conjunta es igual al producto de las probabilidades de cada uno de ellos. De manera intuitiva dos procesos son independientes cuando la ocurrencia, o no, de uno de ellos no influye en el otro. Esto es,

$$P(F \cap G) = P(F)P(G)$$

Generalizando esta definición se dice que una colección de eventos es mutuamente independiente si, para cualquier subconjunto de esa colección de eventos, se cumple

$$P\left(\bigcap_{i=0}^{m-1} F_{l_i}\right) = \prod_{i=0}^{m-1} P(F_{l_i})$$

Hay que tener especial cuidado con el concepto de independencia, en primer lugar el hecho que se cumpla la igualdad anterior para todo la colección de eventos no significa que se cumpla para un subconjunto.

NO es lo mismo el concepto de eventos mutuamente exclusivos o independientes; ¡no es lo mismo!.

Probabilidad condicionada. Teorema de Bayes.

En el mundo real existen muchas interacciones entre variables que forman un determinado modelo, sea este mecánico, eléctrico, electrónico, etc. Esto supone que la observación de un determinado fenómeno puede ayudar a predecir mas fácilmente el resultado de otro. Esta "prediccion más sencilla" refleja que la observación del primer fenómeno modifica o condiciona la probabilidad del segundo.

Sean dos eventos A y B definimos la probabilidad de A dado que el evento B ocurrió como P[A|B]; esta probabilidad se denomina **probabilidad condicional de A dado que B ocurrió**. Otra denominación **es probabilidad de A condicionada a B**.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

De la última expresión se puede obtener el Teorema de Bayes, fundamental a la hora de inferir probabilidades; su expresión viene dada por:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

La generalización de este teorema viene dado por la siguiente expresión; aquí los E_k son eventos mutuamente exclusivos y exhaustivos ($\cup E_k = S$).

$$P(E_s \mid A) = \frac{P(A \mid E_s) \cdot P(E_s)}{\sum_k P(E_k) \cdot P(A \mid E_k)}$$

EL DENOMINADOR DE LA ÚLTIMA EXPRESIÓN HAY QUE ANALIZARLO CON DETENIMIENTO.....MUY IMPORTANTE!!!!!

Ejemplos de lo comentado

	Menopausia		TOTAL
	SI	NO	
Normales	1750	1350	3200
Trastorno A	165	35	200
Trastorno B	45	55	100
TOTAL	1960	1440	3500

Aquí nos encontramos con algo típico; hemos recogidos datos en una población lo suficientemente grande y representativa sobre la aparición de determinados trastornos en mujeres; algunas preguntas.....

¿Probabilidad de padecer el trastorno A?=200/3500=0.057

¿Probabilidad de no padecer ningún trastorno?=3200/3500=0.914

EL ENFOQUE PRESENTADO AQUÍ ES UN ENFOQUE DE LA PROBABILIDAD FRECUENCIAL (EL OTRO PARADIGMA IMPORTANTE ES EL BAYESIANO).

Ejemplos de lo comentado

	Menopausia		TOTAL
	SI	NO	
Normales	1750	1350	3200
Trastorno A	165	35	200
Trastorno B	45	55	100
TOTAL	1960	1440	3500

¿Probabilidad de padecer el trastorno A ó el trastorno B (recordemos que si se da uno no se da el otro)?
=(200/3500)+(100/3500)=0.085

¿Probabilidad de padecer el trastorno A 0 ser menopaúsica (CUIDADO)?
=(200/3500)+(1960/3500)-(165/3500)=0.57

¿Probabilidad de, siendo menopaúsica padezca el trastorno A?= 165/1960=0.084

¿Probabilidad de menopaúsica y de padecer el trastorno A?=165/3500=0.047

¿Son independientes los sucesos de menopausia y de padecer el trastorno B?. Esto es así si se cumple

$P(M \cap B) = P(B) \cdot P(M) =$

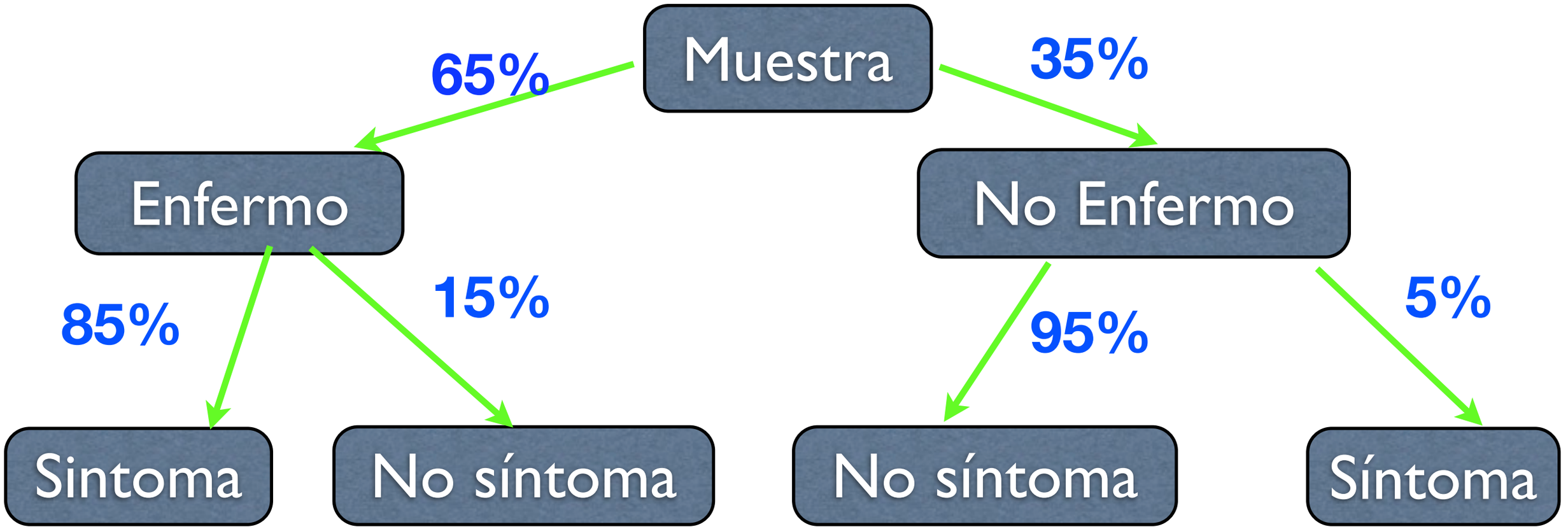
$P(M \cap B) = 45/3500 = 0.012$

$P(B) \cdot P(M) = (100/3500) \cdot (1960/3500) =$
0.016 NO LO SON (CLARO!!).

Otra manera..$P(M \cap A) = P(A|M) \cdot P(M) = (165/1960) \cdot (1960/3500) = (165/3500)$

Ejemplo del Teorema de Bayes.

Se escoge una muestra de 1000 personas de las que el 65% son enfermos. De los enfermos hay un 85% de casos que tienen un cierto síntoma y de los no enfermos el porcentaje de casos de personas que presentan síntomas es del 5%.



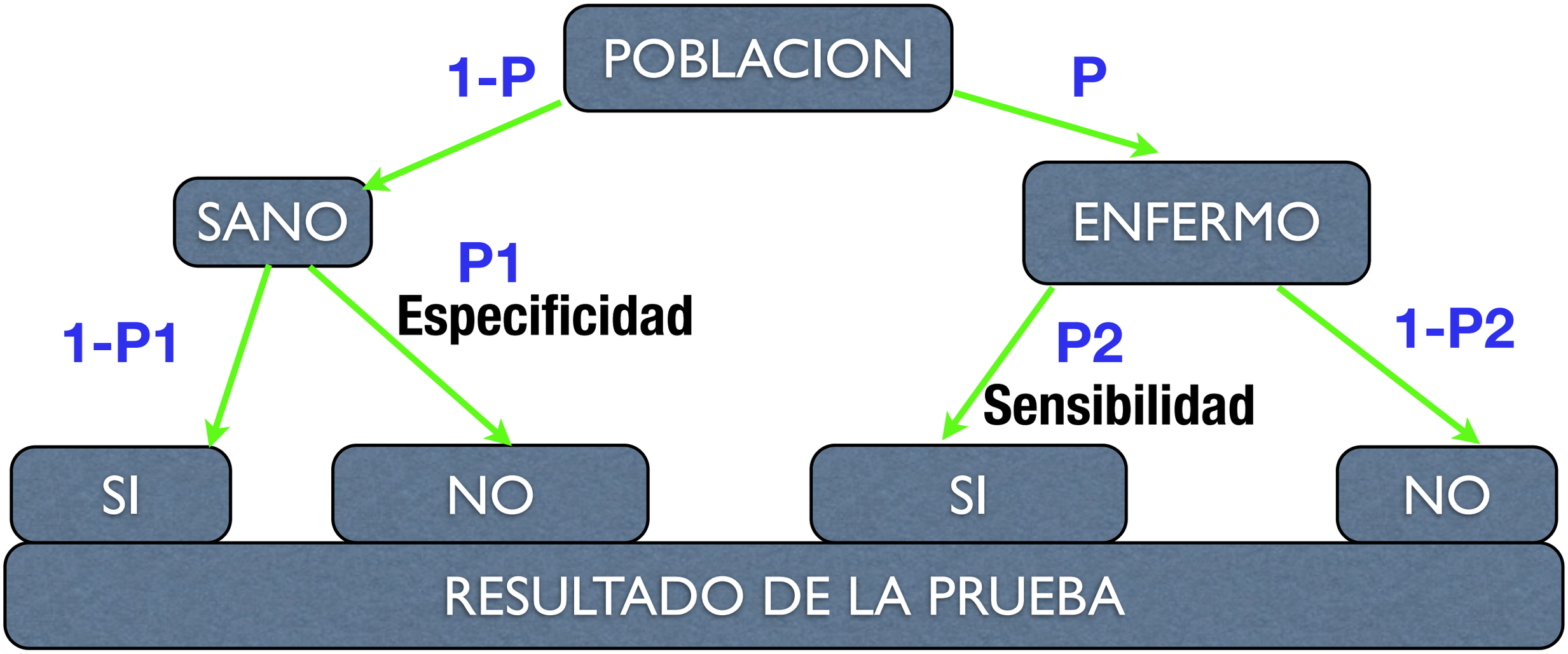
Lo primero es usar una expresión muy útil en teoría de probabilidad; a modo de ejemplo; si escogemos una persona al azar; ¿cuál sería la probabilidad de que tuviera ese síntoma?

$$P(S)=P(E)\cdot P(S|E)+P(NE)\cdot P(S|NE)=0.65\cdot 0.85+0.35\cdot 0.05=0.57$$

¿Cuál es la probabilidad que, seleccionando una persona que tiene ese síntoma la persona esté enferma?

$$P(E|S)=[P(S|E)\cdot P(E)]/P(S)=[0.85\cdot 0.65]/0.57=\mathbf{0.96}$$

Ejemplo del Teorema de Bayes. Pruebas clínicas



En este esquema tenemos que P es la probabilidad de tener una cierta enfermedad; P1 es la probabilidad que, estando sano, la prueba acierte, este parámetro se conoce como **especificidad**, por el contrario P2 es la probabilidad que estando enfermo la prueba diagnóstica acierte; este parámetro se conoce como **sensibilidad.**

Ejemplo: tomemos p=0.02; p1=0.05 y p2=0.97; con estos valores me realizo la prueba y sale positivo, ¿cúal es la probabilidad que esté enfermo?. S=Sí; N=No.

$$P(S)= [P(S|Enf)\cdot P(Enf)]+ [P(S|Sano)\cdot P(Sano)]=0.97\cdot 0.02+(1-0.05)\cdot(1-0.02)=0.95$$

$$P(Enf|S)=[P(S|Enf)\cdot P(Enf)]/P(S)=[0.97\cdot 0.02]/0.95=0.02$$ *(¡¡¡¡¡CUIDADO PORQUE EL P2=0.97 NOS PUEDE CONDUCIR A ERRORES!!!!!!)*

Ejemplo del Teorema de Bayes. Modelos(I)

Supongamos que nos plantean un problema en el que tenemos que establecer un modelo que prediga si un paciente tiene, o no, una determinada enfermedad. Dicho modelo se aplica sobre la muestra que se tiene obteniéndose lo siguiente:

	Enfermo	Sano	TOTAL
Si	25	10	35
No	5	75	80
TOTAL	30	85	115

Con esta tabla podríamos calcular los parámetros de la anterior transparencia así como algunos otros que se utilizan habitualmente en los modelos predictivos clínicos.

Sensibilidad=P(S|Enf)=25/30=0.833

Especificidad=P(N|Sano)=75/85=0.88

Si nos preocupamos de la capacidad de predicción del modelo aparecen dos cantidades importantes; que son los **valores predictivos (positivo y negativo).**

VPP=P(Enf|S)=[P(S|Enf)·P(Enf)]/P(S)

VPN=P(Sano|N)=[P(N|Sano)·P(Sano)]/P(N)

De la tabla se puede deducir que

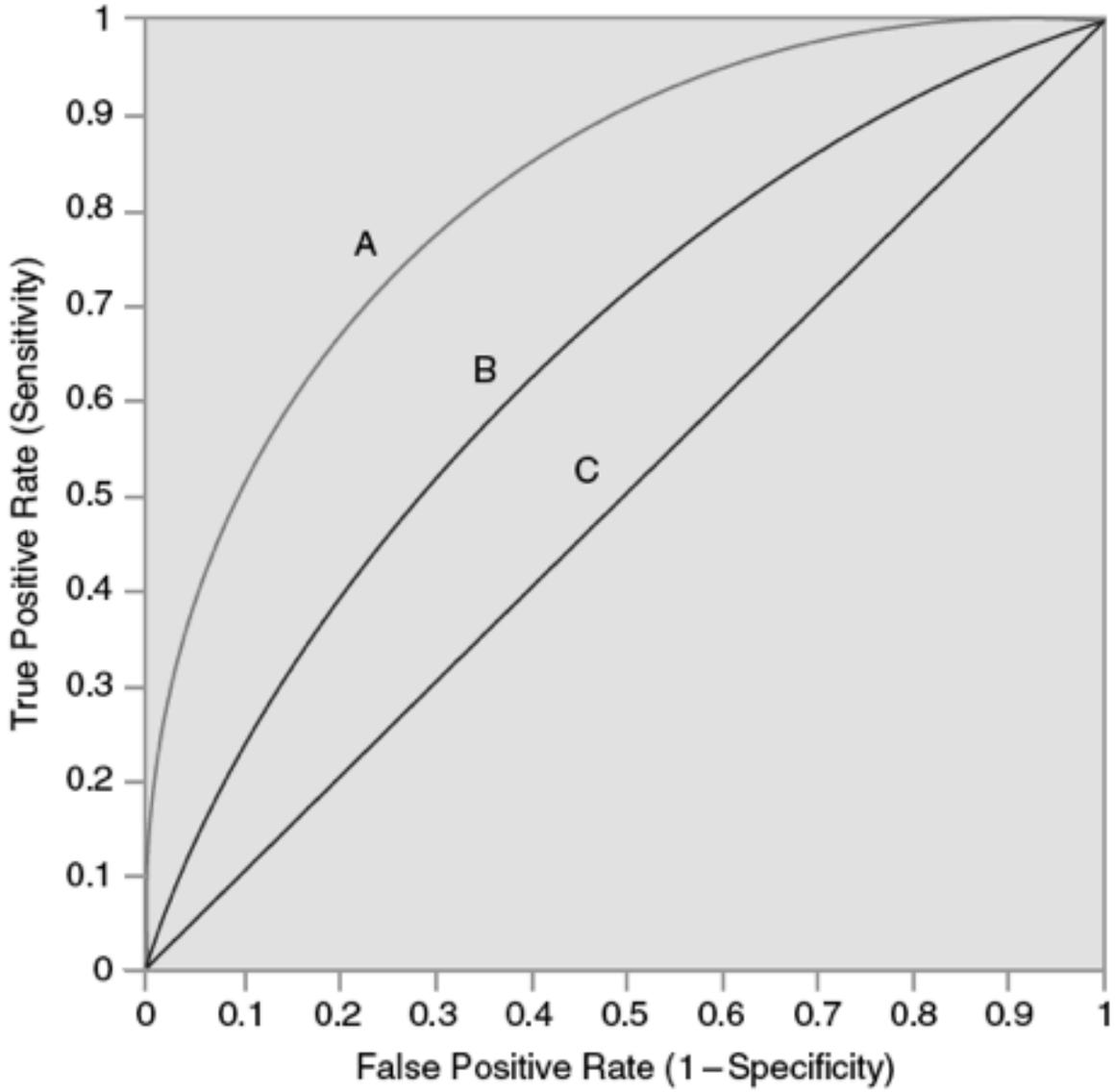
P(S)=35/115=0.304
P(N)=1-P(S)=80/115=0.695
P(Enf)=30/115=0,260
P(Sano)=85/115=0.739

por lo que usando la sensibilidad y la especificidad se tiene

VPP=[0.833·0.260]/0.304=0.71

VPN=[0.88·0.739]/0.695=0.93

Ejemplo del Teorema de Bayes. Modelos(II)



(Advanced Data Mining Techniques, Springer 2008)

Una figura muy usada es lo que se conoce como **curva ROC** *(Receiver Operating Characteristic)* donde se representan las cantidades sensibilidad y (1-especificidad) en función de un determinado parámetro de nuestro modelo o de alguna cantidad de la prueba clínica a realizar.

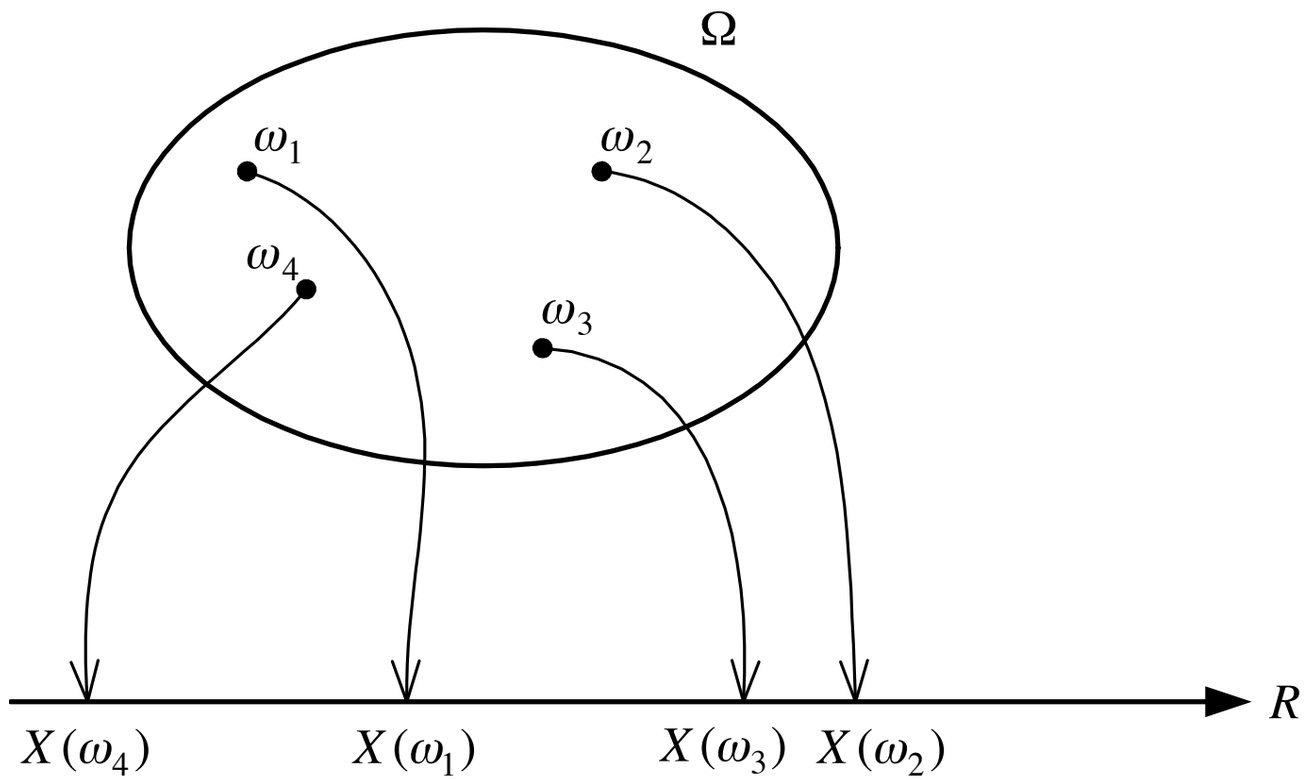
La siguiente tabla da todos los parámetros de las tablas 2x2.

		Enfermedad	
		Si	No
Test	+	A	B
	-	C	D

Sensibilidad	A/(A+C)
Especificidad	D/(B+D)
Valor predictivo positivo	A/(A+B)
Valor predictivo positivo	D/(C+D)
Aciertos	(A+D)/(A+B+C+D)

Variable aleatoria.

Se puede establecer una correspondencia entre los eventos del espacio muestral, ya sea discreto o continuo y los números reales. Se tiene entonces una variable aleatoria, bien discreta bien continua. En la siguiente figura Ω es el espacio muestral, w_k son los diferentes eventos y X es la variable aleatoria.



A modo de ejemplo tenemos las siguientes variables aleatorias:

En el lanzamiento de una moneda le asigno un 1 si sale cara y un 0 si sale cruz.

Con la misma asignación anterior puedo plantear la variable aleatoria "valor acumulado que se tendrá tras cinco lanzamientos"

No tiene por qué existir una asignación, así si considero el lanzamiento de un dado el propio valor del dado puede ser la variable aleatoria.

Variable aleatoria.

Se define la función de distribución de la variable aleatoria X aquella definida de la siguiente forma (P denota probabilidad).

$$F_X(x) = P(X \leq x)$$

Esta función puede ser continua o discreta, dependiendo de como sea el espacio muestral. Esta función tiene una serie de propiedades importantes como son:

1. $0 \leq F_X(x) \leq 1$.
2. $F_X(x)$ is nondecreasing: $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$.
3. $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.
4. $P(a < X \leq b) = F_X(b) - F_X(a)$.

Relacionada con esta función de distribución se encuentra la **función densidad de probabilidad** definida de la siguiente forma.

$$p_X(x) = \frac{dF_X(x)}{dx}$$

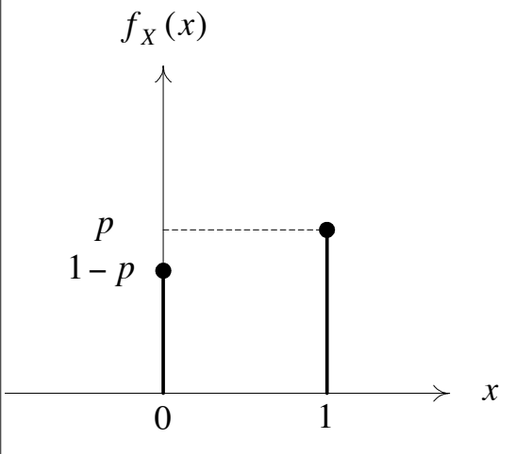
Cuando la variable aleatoria es discreta se utiliza otra función conocida como **función de probabilidad** definida como

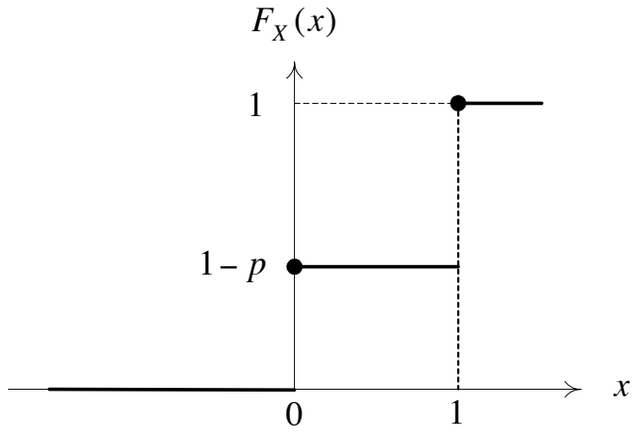
$$p_i = P(X = x_i)$$

Variable aleatoria. Ejemplos (I).

Lanzamiento de un dado. En este caso habría que determinar el valor de P(x=k) si estamos interesados en la función densidad; tenemos entonces 1/6 para todo k; o bien, si estamos interesados en la función de distribución se tendría F(1)=1/6; F(2)=1/3; F(3)=1/2; F(4)=2/3; F(5)=5/6 y F(6)=1.

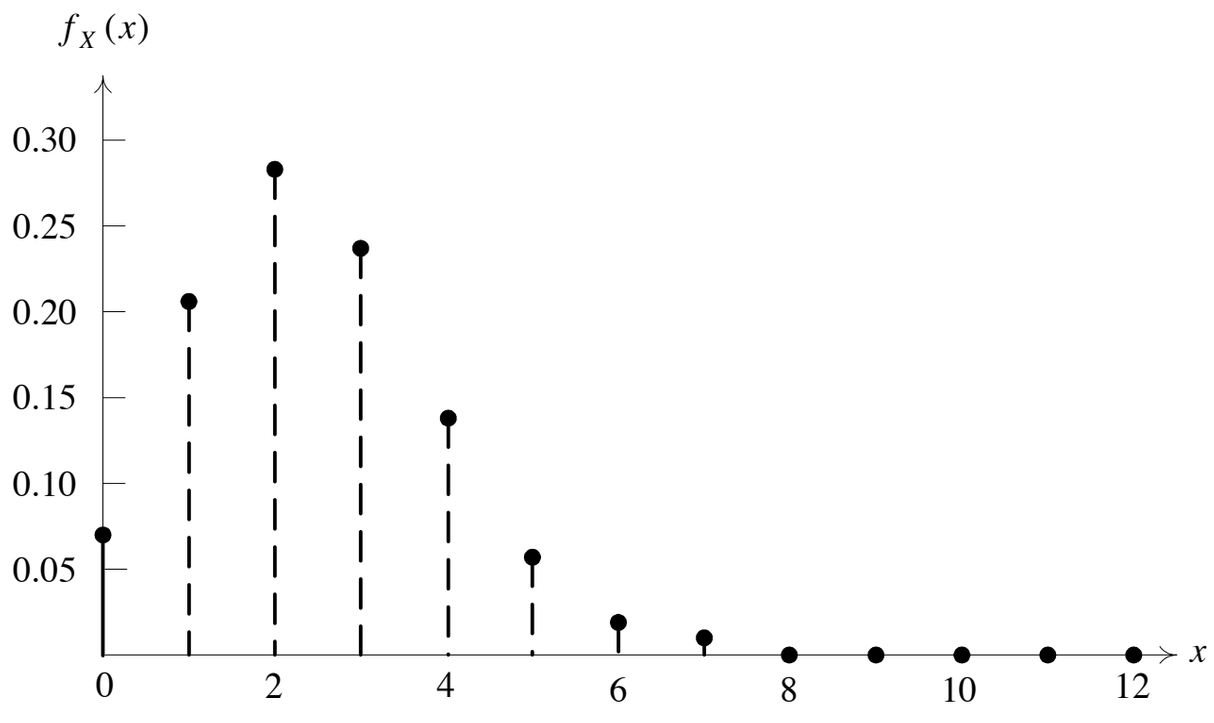
Bernouilli. Variable aleatoria que toma dos valores con probabilidades p y 1-p. Ejemplos lanzamiento de una moneda asignando 1 a cara y 0 a cruz; que un determinado tratamiento médico vaya bien.





Binomial. Variable aleatoria que da el número de eventos que suceden en una secuencia de n independientes pruebas de Bernouilli. Ejemplos número de caras tras n lanzamientos de una moneda; número de personas que padecerán una enfermedad si se tiene una cierta probabilidad de aparición.

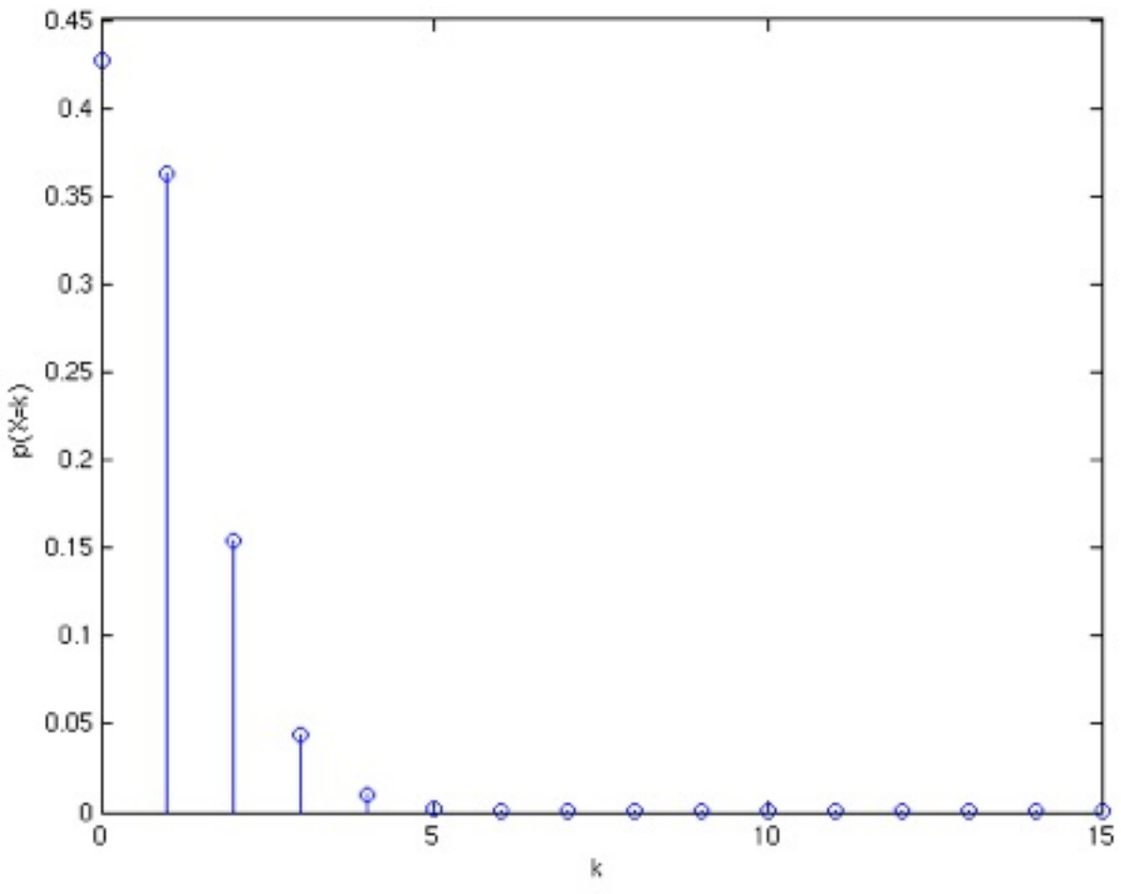
$$P(X=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & 0 \le k \le n \\ 0, & \text{otherwise} \end{cases}$$



Variable aleatoria. Ejemplos (II).

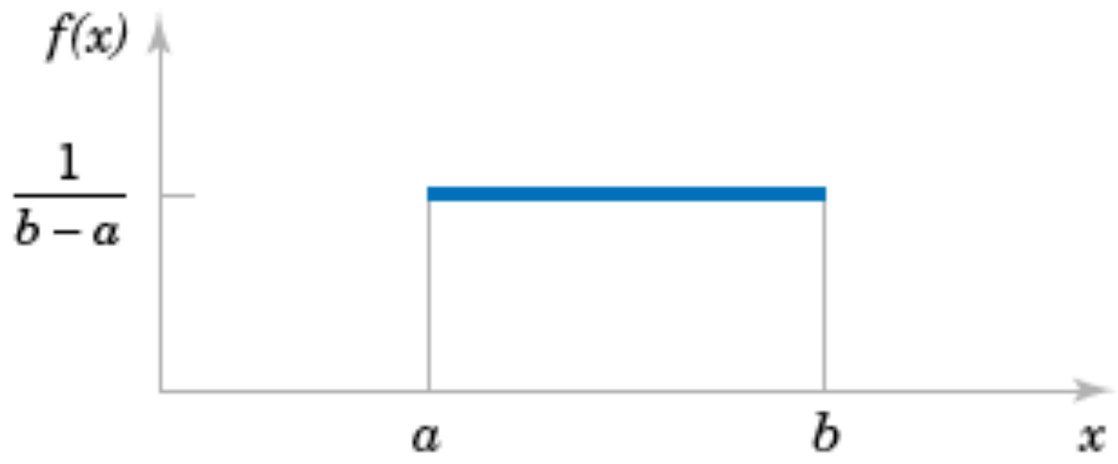
Poisson. Variable aleatoria que es una aproximacion de la binomial cuando n es grande (n>30) y el valor de p es pequeño (p<0.1). Es la que manejan compañías de seguros (sucesos raros con una población relativamente alta).

$$P[X=k] = e^{-\mu} \cdot \frac{\mu^k}{k!} \quad k = 0,1,2..$$



Uniforme. Variable aleatoria con densidad de probabilidad constante en un intervalo. Típica en problemas donde no se tiene un conocimiento "a priori" del resultado del experimento; como veremos siempre la utilizamos de forma "encubierta"

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

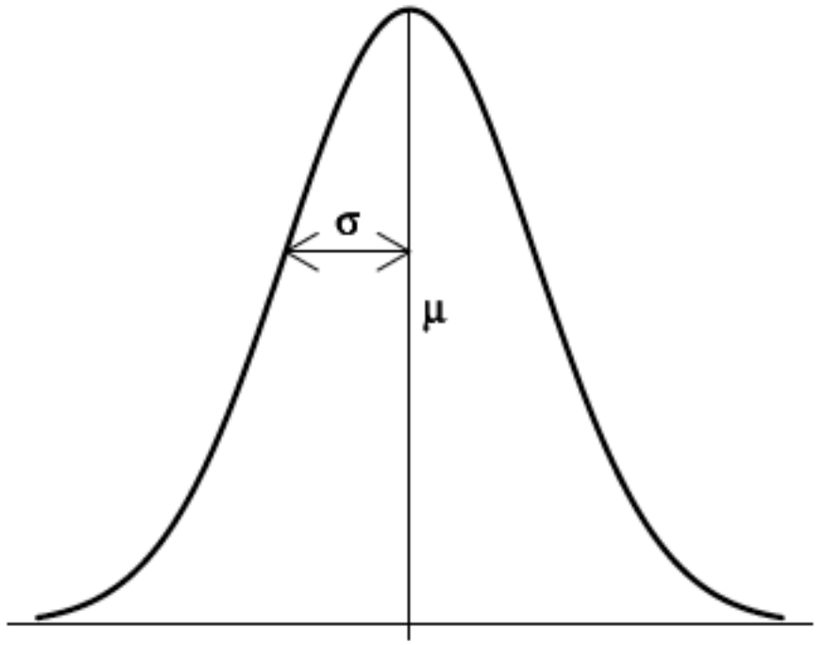


Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003

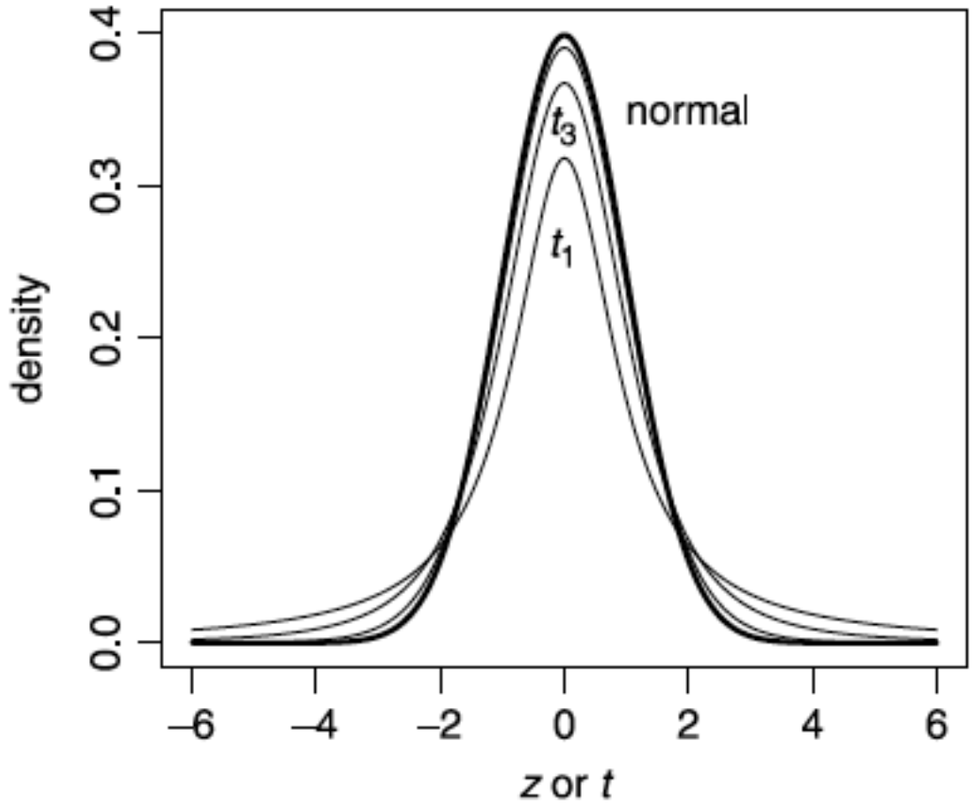
Variable aleatoria. Ejemplos (III).

Normal. Es la "reina" de las variables aleatorias a causa del Teorema del Límite Central. Este teorema viene a decir que, la suma de un conjunto de sucesos aleatorios sigue una distribución normal. Su densidad de probabilidad es

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



t de Student. Variable aleatoria parecida en forma a la normal y que se utiliza muy a menudo en los contrastes de hipótesis de tipo paramétrico. Existe un parámetro en su función de densidad que es el número de grados de libertad, ν, denotándose dicha variable por t_ν



Statistics and Data with R; An Applied Approach Through Examples, Wiley 2008

Variable aleatoria. Momentos.

Hasta ahora se tiene una serie de experimentos aleatorios que se corresponden con números reales; ¿podemos encontrar algún significado físico usando las funciones definidas anteriormente?. Aparecen entonces los momentos. (en lo que sigue f_x es la función densidad de probabilidad)

Definimos el **valor esperado de la variable aleatoria X como**

$$m_x = E[X] = \int x \cdot f_x(x) \cdot dx \quad \text{Variable continua.}$$

$$\mu_x = E[X] = \sum_k x_k \cdot P(x = x_k) \quad \text{Variable discreta.}$$

A partir de ahora se supondrá que la variable es continua, es inmediato obtener la expresión discreta.

Generalizando esta definición se tienen los **momentos de orden n,** (resaltar que la anterior definición es el momento de orden 1)

$$m_x^n = E[X^n] = \int x^n \cdot f_x(x) \cdot dx$$

Otros parámetros importantes son los **momentos centrales de orden n**

$$E[(X - m_x)^n] = \int (x - m_x)^n \cdot f_x(x) \cdot dx$$

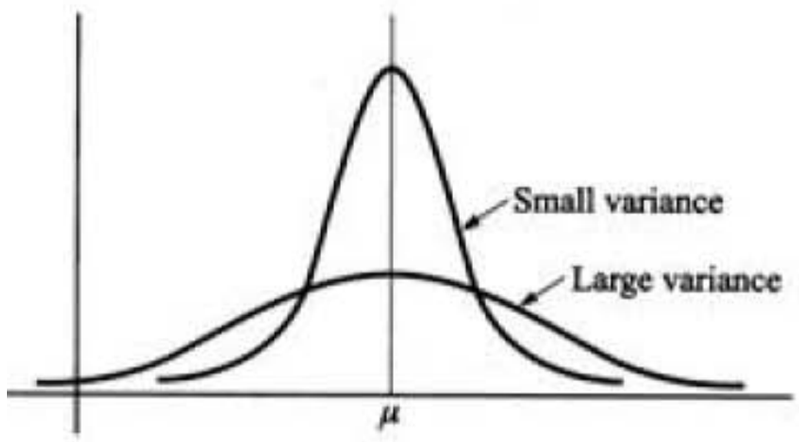
Uno de los momentos centrales más utilizados es la **varianza** definida como

$$\sigma_x^2 = E[(X - m_x)^2] = \int (x - m_x)^2 \cdot f_x(x) \cdot dx$$

Significado de algunos momentos.

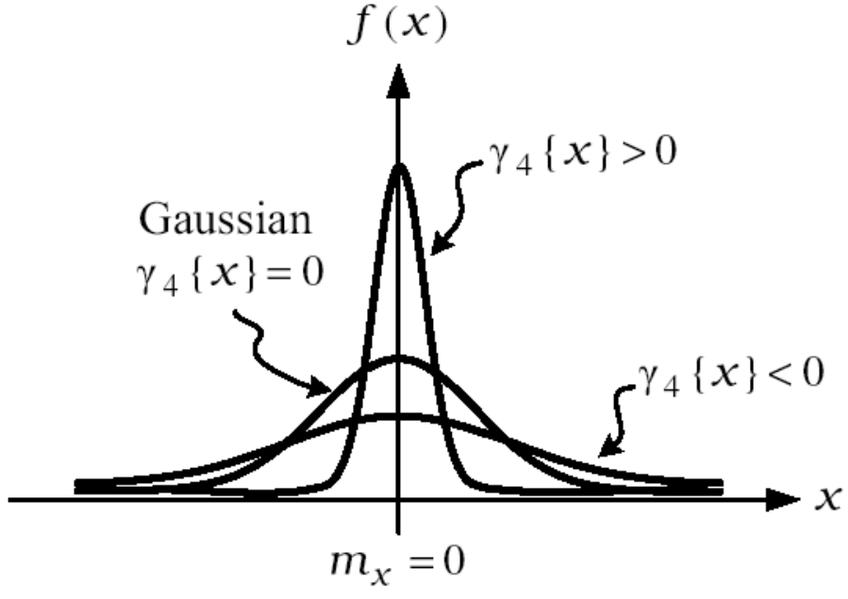
VALOR ESPERADO; da idea de la tendencia central de la variable aleatoria de acuerdo a su distribución de probabilidades.

VARiANZA; da idea de la dispersión de la variable aleatoria (refleja la anchura de la distribución). Un parámetro que se utiliza mucho más que la varianza es la **DESVIACIÓN ESTÁNDAR, σ,** que es la raíz cuadrada de la varianza.

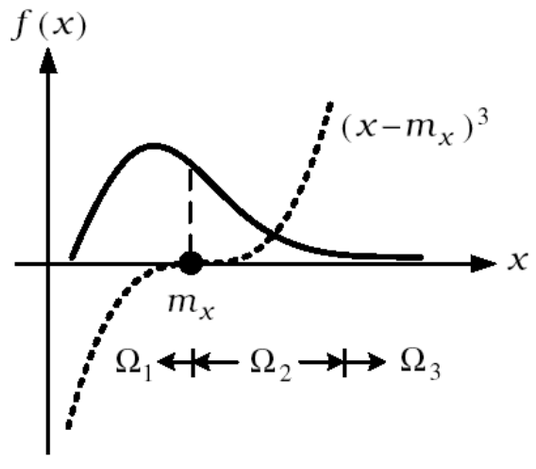


Kurtosis; da idea de lo " picuda" que es la función densidad de probabilidad de una determinada variable aleatoria tomando como referencia una distribución normal.

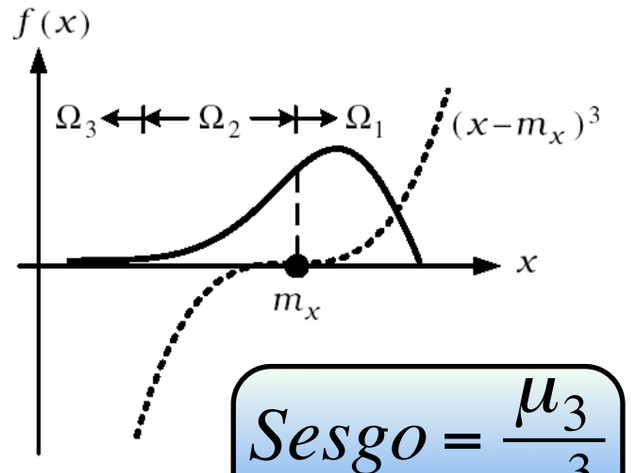
$$Kurtosis = \frac{\mu_4}{\sigma^4} - 3$$



SESGO; define el grado de asimetría de una determinada función densidad de probabilidad; el parámetro más utilizado es el cociente entre el momento central de orden 3 y la desviación estándar al cubo



(b)



$$Sesgo = \frac{\mu_3}{\sigma^3}$$

(c)

Estadística.

Hasta ahora hemos analizado las características y los parámetros que definen una magnitud que procede de un experimento aleatorio; esa aleatoriedad puede ayudarnos a explicar diferentes comportamientos en situaciones reales. Podríamos definir **la estadística como aquella parte de conocimiento que analiza procesos donde existe un determinado grado de aleatoriedad**

- Planteamos hipótesis
- Recopilamos datos (muestreo)
- Análisis de datos
- Obtención de conclusiones

Estas son las etapas clásicas de cualquier análisis estadístico.El problema que nos encontraremos en muchas ocasiones es que la toma de datos o muestreo, que es una etapa esencial en el desarrollo de modelos, se realiza sin ningún control de tal forma que se tienen los datos que el clínico ha recopilado a lo largo del tiempo pero no se ha diseñado una toma de muestras. Debemos empezar a tener en cuenta que una cosa es lo ideal y otra cosa lo que nos vamos a encontrar.....

Estadística. Definiciones.

Población; *conjunto sobre el que estamos interesados en obtener conclusiones, en la mayoría de las ocasiones es demasiado grande para poder analizarlo.*

Muestra; *subconjunto de la población y del cual tenemos datos y observaciones. Evidentemente debería ser representativo de la población.*

Estadístico; *es una función de los valores de la muestra; uno de los más sencillos es la media muestral.*

Contraste de hipótesis; *también se le conoce como prueba de significación o prueba estadística y consiste en decidir si una determinada hipótesis sobre la población debe ser aceptada, o no, analizando estadísticamente la muestra.*

Paramétrico y no paramétrico, *son los dos tipos de contrastes que puedo realizar dependiendo si la característica sobre la que se realizó la hipótesis se ajusta a una determinada distribución de probabilidad o no.*

Contraste de hipótesis.

El punto de partida de este análisis son dos hipótesis; la que se conoce como hipótesis nula y se designa por H_0 y la que se denomina alternativa y que se designa por H_1. Hay que escoger como hipótesis nula la más simple y la que conlleve (si se da el caso) un signo de igualdad.

Ejemplos de planteamiento de hipótesis podría ser:

H_0: existe igualdad de salarios entre hombres y mujeres

H_1: **no** existe igualdad de salarios entre hombres y mujeres

H_0: la edad media de jubilación anticipada es de 62 años

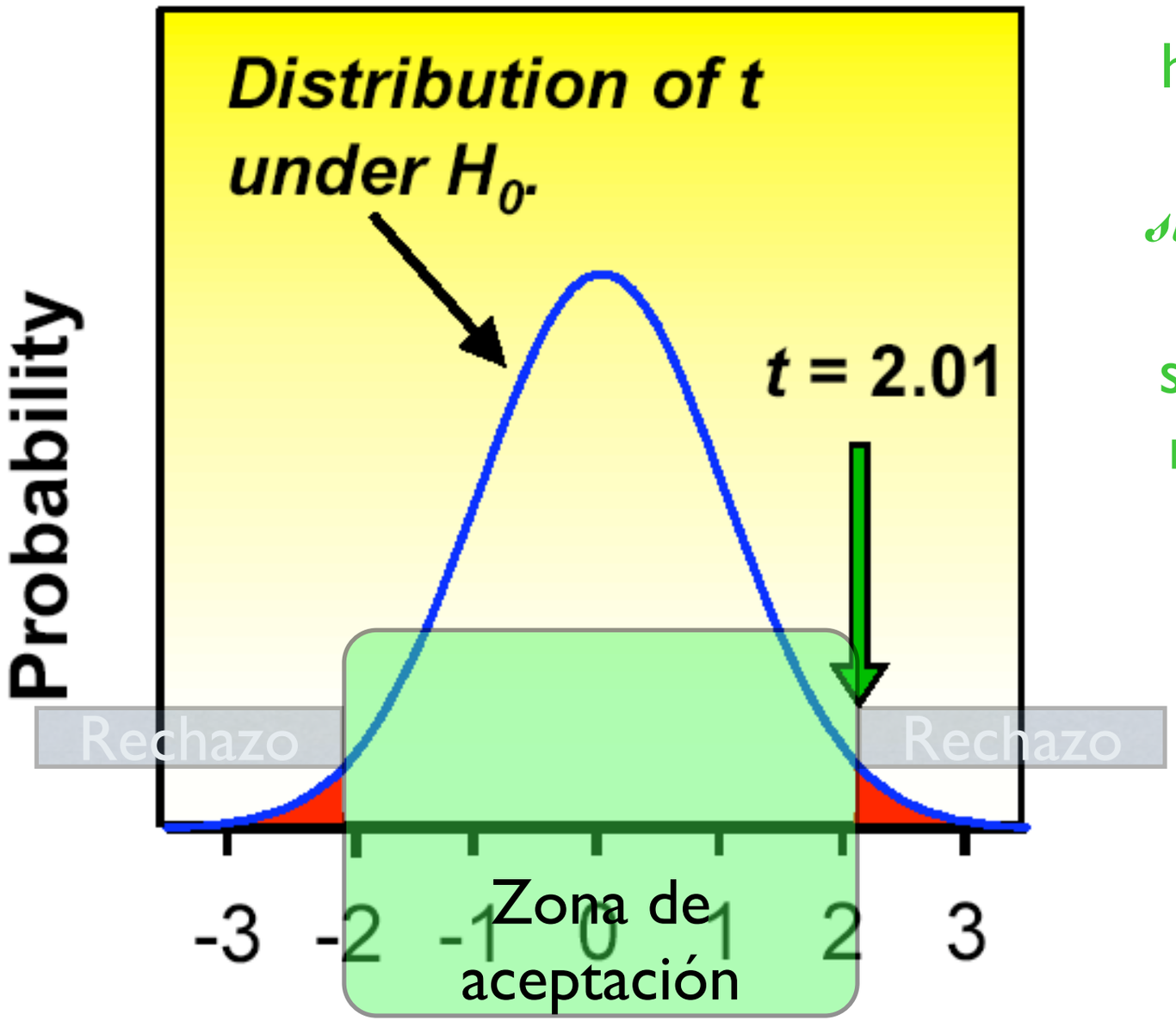
H_1: **no es de 62 años**

CLASE DE ERROR

	H_0 cierta	H_0 falsa
Acepto H_0	No hay error	Error de tipo II
Rechazo H_0	Error de tipo I	No hay error

La idea es establecer un modelo probabilístico para tomar una decisión de una determinada magnitud que se conoce como ***estadístico de contraste.*** . En dicho modelo se establecerán dos zonas disjuntas y complementarias denominadas **zona de rechazo y zona de aceptación.**

Contraste de hipótesis.



El problema aquí radica en conocer todas las posibles distribuciones que se pueden tener según el tipo de hipótesis a comprobar y según las condiciones que se cumplan en nuestros datos (si el test tiene que ser paramétrico o no paramétrico).

La pregunta evidente es; ¿qué umbral ponemos para aceptar/rechazar la hipótesis nula?. Ese umbral denotado por α, se conoce como ***umbral de significación*** y, normalmente, se toma igual a 0.05. Si se quiere mayor seguridad de cumplimiento se puedo reducir ese umbral, otras elecciones son tomarlo igual a 0.01 o a 0.001

Los paquetes estadísticos devuelven el valor de la probabilidad, p, que se conoce como ***significación muestral de la hipótesis nula***, de tal forma que se procede de la siguiente forma:

$p < \alpha$: Rechazamos H_0

$p > \alpha$: Aceptamos H_0

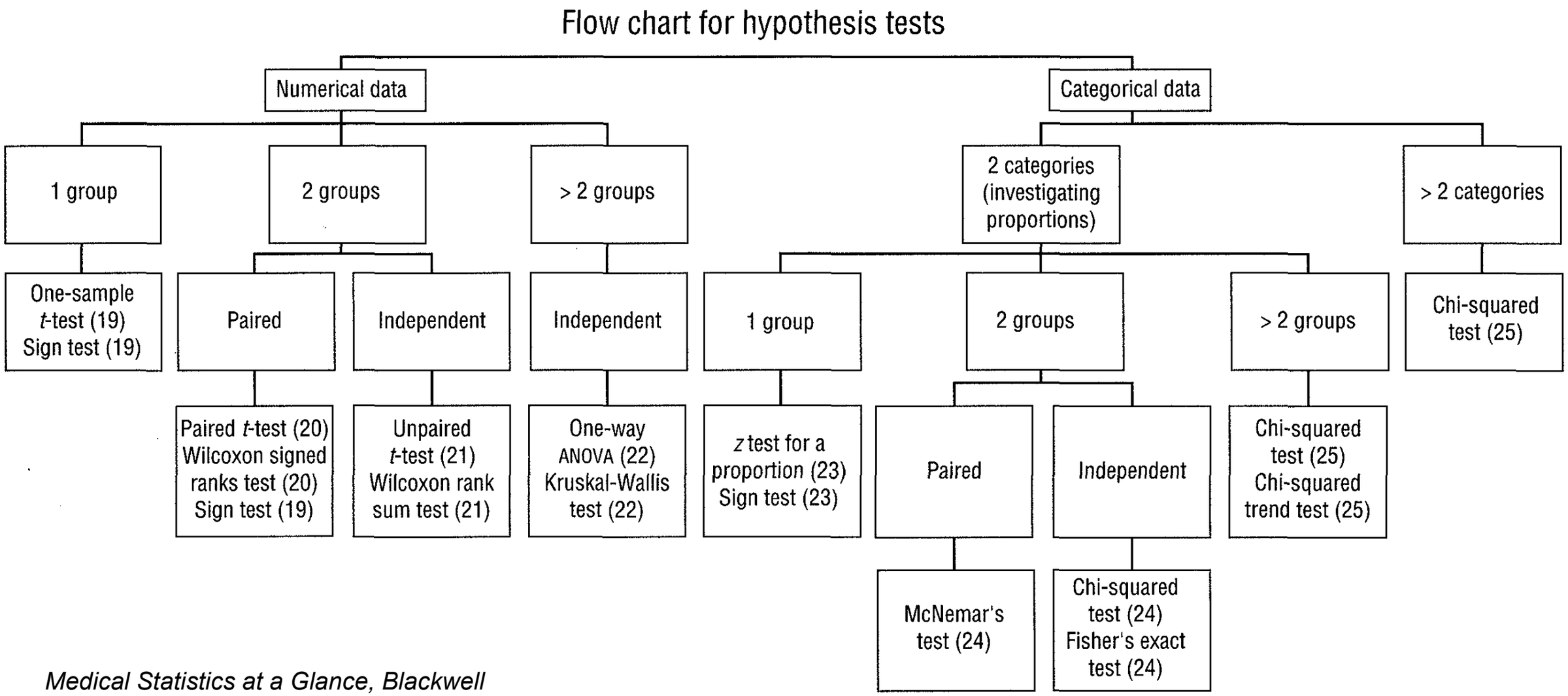
Contraste de hipótesis.

Establecemos hipótesis de trabajo
Recogemos los datos necesarios
Fijamos H_0
Determinamos el análisis a realizar de acuerdo a las condiciones que se cumplan
Aceptamos o rechazamos H_0 de acuerdo al valor de p obtenido y al de α fijado con anterioridad.

Esta parte se conoce como muestreo y existen muchas formas de hacer dicha recogida. Es la etapa crítica ya que los datos tienen que ser suficientemente representativos de lo que se quiere determinar. Lo que uno se encuentra, en muchas ocasiones es la base de datos que se tiene de la práctica diaria (la teoría está bien pero.....)

Recordemos siempre lo que estamos haciendo; el resultado de p indica la evidencia en contra de la hipótesis nula; cuanto menor es este valor mayor es la evidencia contra dicha hipótesis.

Contraste de hipótesis. Resumen.



Medical Statistics at a Glance, Blackwell

Análisis de supervivencia.

En un análisis de supervivencia estamos interesados en dos variables; por una parte la ocurrencia, o no, de un determinado suceso y, por otra parte, el tiempo que transcurre hasta que se produce dicho suceso.

Tiene aplicaciones en un gran número de campos, por ejemplo, en la industria se utiliza para evaluar el tiempo de funcionamiento de los componentes. En clínica lo podemos usar para determinar el tiempo de recuperación usando un determinado fármaco, el tiempo que el paciente sobrevive tras un determinado trasplante, etc.

Variable 1	Variable 2	Días antes del suceso
1,2	-0,3	4
0,6	2,4	2
2,2	1,1	1
0,2	-0,7	5

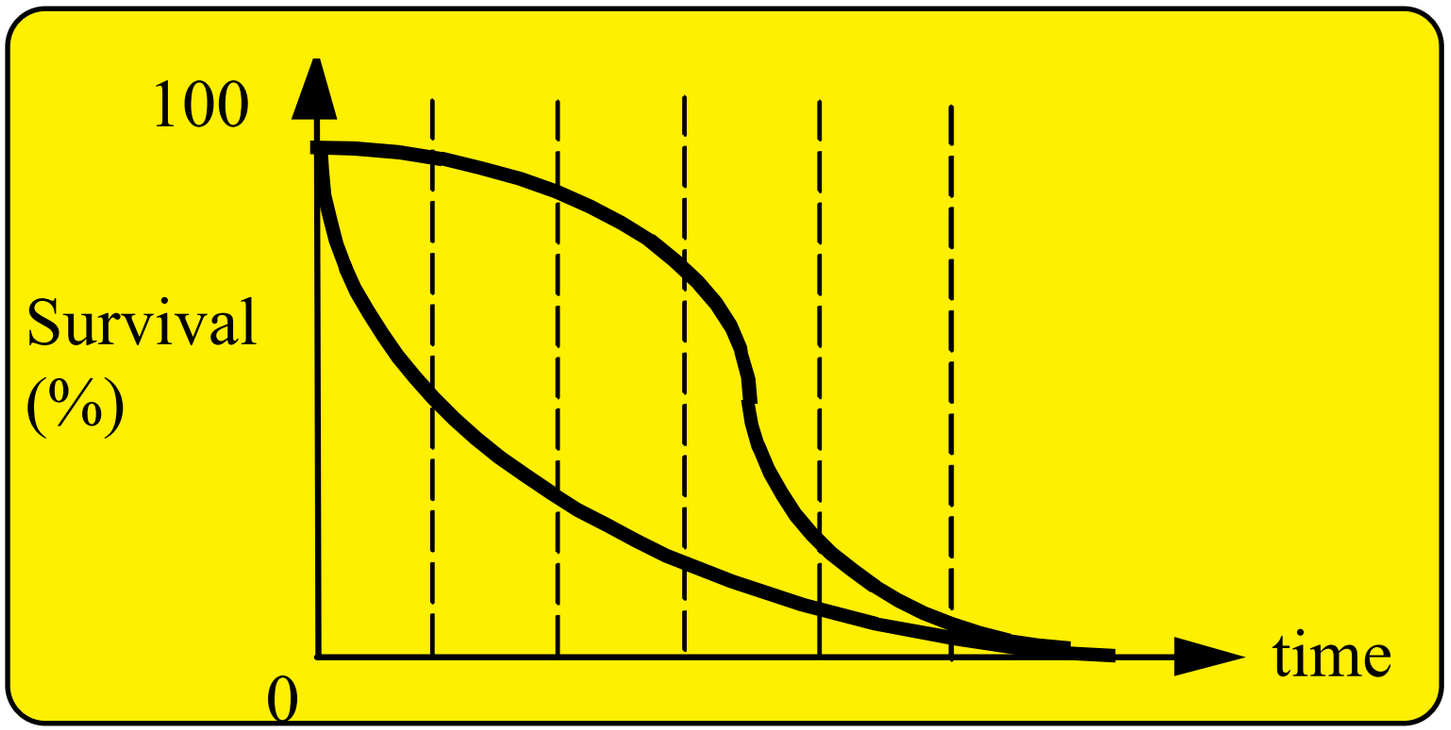
Usando estas variables

¡ PREDECIMOS EL TIEMPO!

Análisis de supervivencia.

Definimos una **función de supervivencia** que da la probabilidad que un determinado individuo no tenga la ocurrencia del suceso antes del tiempo t

$$S(t) = P(T > t)$$



S(0)=0 y S(∞)=1. Una manera de obtenerla a partir de los datos es mediante la estimación

$$S(t)=N(t)/N_{total}$$

Donde N(t) es el número de sujetos en los que no se ha producido el suceso en el tiempo t (ni anteriormente)

Aquí nos encontramos con dos problemas importantes de cara a un análisis estadístico; por una parte el tiempo no sigue una distribución normal y, por otra parte, existirán sujetos que saldrán fuera del tiempo de estudio y, por tanto, no se conocerá el tiempo en el que se produce el suceso. Estos sujetos se conocen como sujetos censurados. En este grupo están también los que son apartados del estudio por alguna razón o, simplemente abandonan el estudio. Normalmente se asume que estos sujetos se comportan igual que los sujetos al final del estudio.

Análisis de supervivencia.

Otra función relacionada con la función de supervivencia, que denotamos por F(t), es la que da la probabilidad de ocurrencia del suceso transcurrido un tiempo t; es inmediato obtener

$$S(t) = 1 - F(t)$$

Está claro que F(t) y S(t) tienen un significado de funciones de distribución de probabilidad. Obtendremos ahora una función densidad de probabilidad de ocurrencia del suceso, f(t), a partir de F(t) como

$$f(t)=\lim_{\Delta t\to 0} (N(t+\Delta t)/\Delta t)=F'(t)$$

Donde N(t+Δt) es el número de sucesos que ocurren en el intervalo de tiempo t y t+Δt y F'(t) es la derivada de la función F(t)

Se define la función de riesgo, o tasa de fallo, h(t), como la tasa instantánea de fallo en el instante t. Se calcula como

$$h(t) = f(t)/S(t)$$

$$h(t) = -S'(t)/S(t) = -d \log S(t)/dt$$

Podemos intentar determinar alguna de las funciones anteriormente mencionadas mediante dos aproximaciones; paramétrica (asumimos una determinada forma para la función y determinamos sus parámetros) o bien podemos plantear una aproximación no paramétrica (no se asume ningún modelo y son los propios datos quienes definen dichas funciones).

Análisis de supervivencia.

En relación a las aproximaciones paramétricas se tienen los modelos que consideran la exponencial y su generalización, la distribución de Weibull

$$f(t) = \frac{k}{\lambda} \cdot \left(\frac{t}{\lambda}\right)^{k-1} \cdot e^{-\left(\frac{t}{\lambda}\right)^{k}}$$

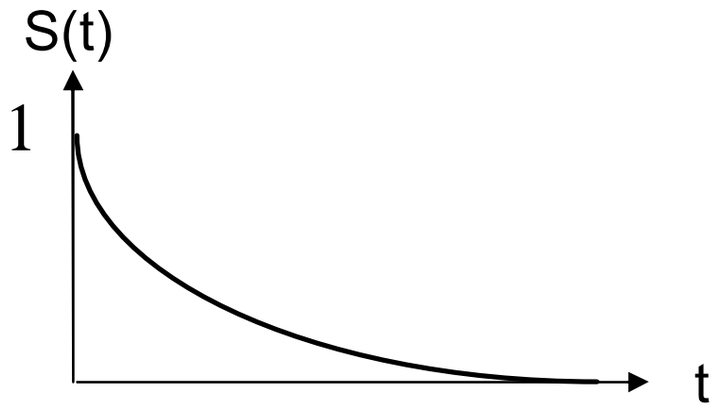
$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^{k-1}} \quad \Leftrightarrow h(t) = \frac{k}{\lambda} \cdot \left(\frac{t}{\lambda}\right)^{k-1}$$

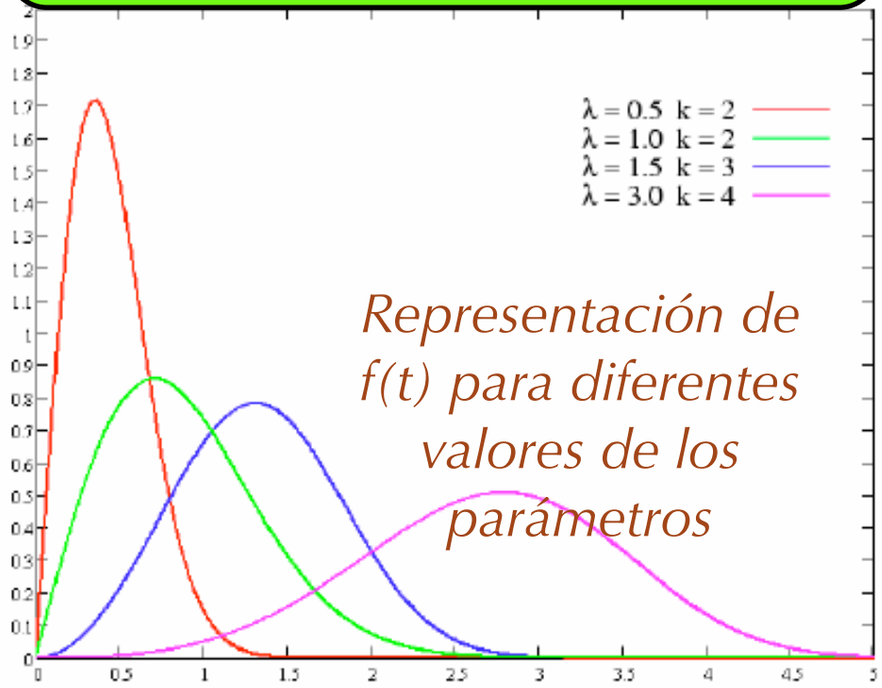
Exponencial

$f(t) = \lambda e^{-\lambda t}$

$S(t) = e^{-\lambda t}$

$h(t) = \lambda$

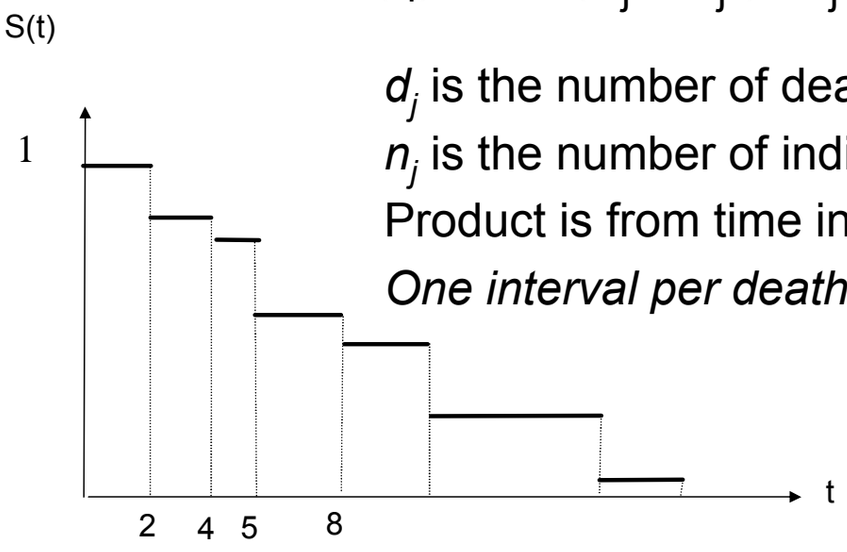




Representación de f(t) para diferentes valores de los parámetros

Modelo de Kaplan-Meier es un modelo no paramétrico en el que se tienen un producto de *probabilidades*.

$$S(t_i) = \Pi\ (n_j - d_j)/\ n_j$$



d_j is the number of deaths in interval j

n_j is the number of individuals at risk

Product is from time interval 1 to j

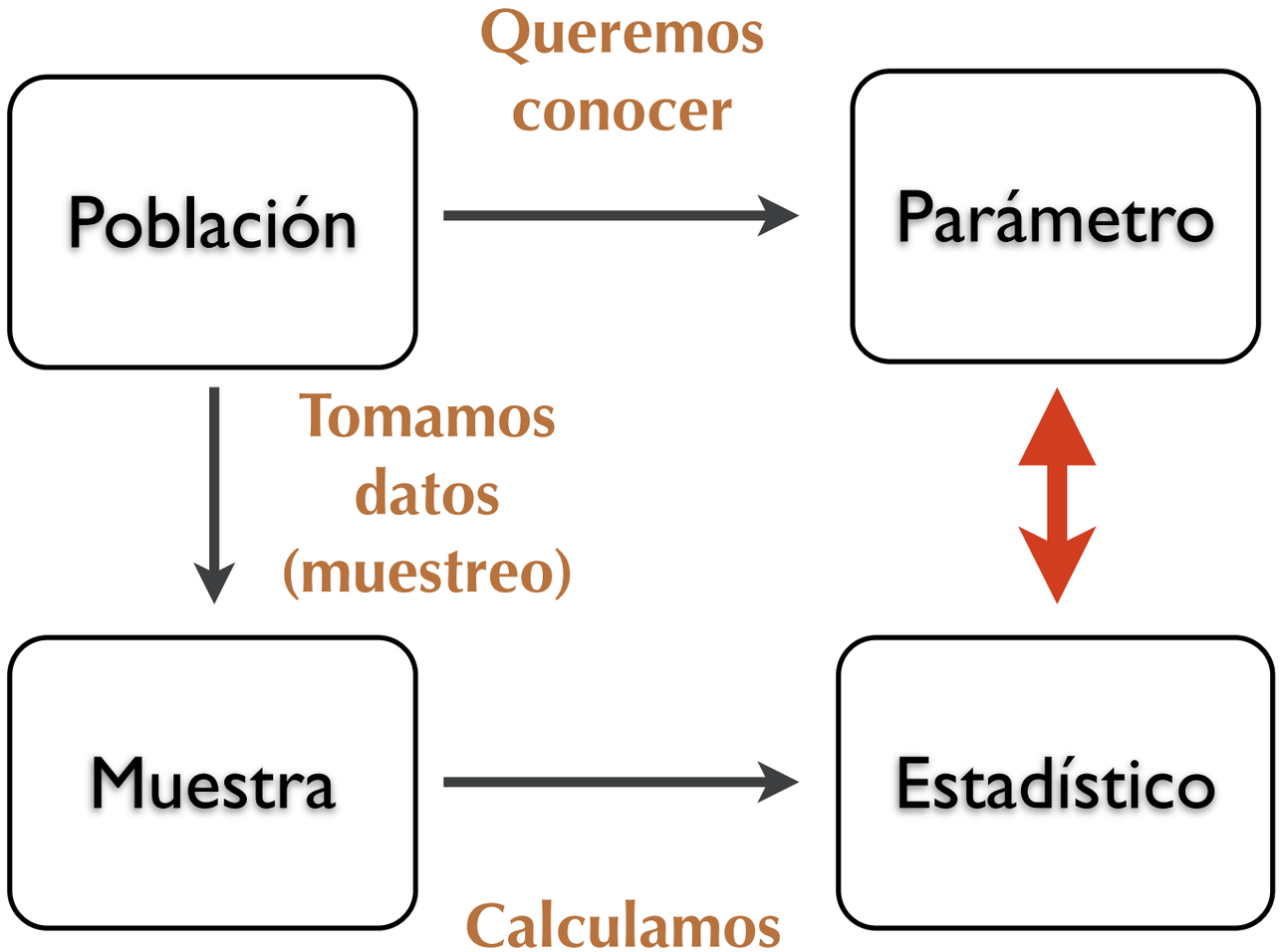
One interval per death time

Otra aproximación ampliamente extendida es el modelo de Cox donde la función de riesgo es

$$h_i(t) = h_0(t) \cdot e^{[\beta_1 \cdot x_1 + \ldots\ldots + \beta_N \cdot x_N]}$$

Donde $h_i(t)$ es el riesgo para el individuo i en el instante t, x_k son variables descriptivas de dicho individuo y, finalmente, las β_i son los parámetros que se han de determinar de acuerdo a los datos que se tienen.

Estadísticos (I)



Una definición muy general de estadístico es el de cualquier cantidad determinada a partir de los datos obtenidos de un muestreo. Esa cantidad tendrá un carácter aleatorio en cuanto que su origen es un proceso de muestreo; podemos aquí aplicar nuestros conocimientos de probabilidad y estadística.

Hasta ahora se han analizado las diferentes distribuciones/densidades de probabilidad mediante una serie de magnitudes (valor medio, desviación estándar, sesgo, etc). ¿Como se procede cuando no se conoce EXACTAMENTE la función que genera los datos obtenidos?.

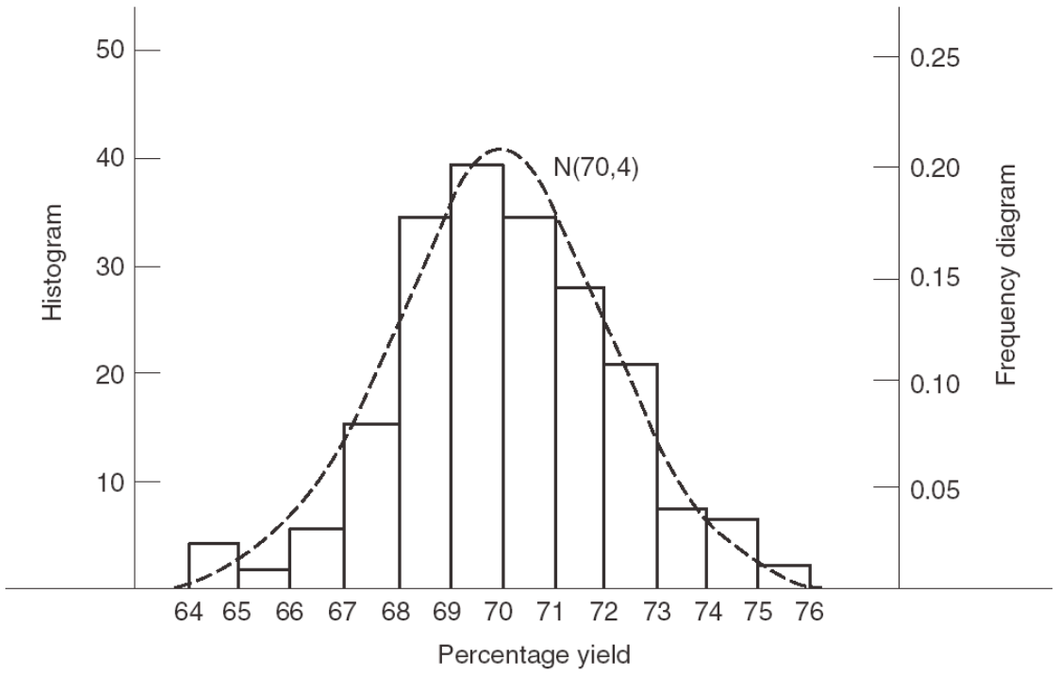
La manera de proceder será estimar las magnitudes anteriormente comentadas a partir de los datos que se tienen. En las expresiones que siguen se supone una distribución uniforme para el cálculo de eses magnitudes. En todas las expresiones se puede utilizar la frecuencia de aparición del dato para los diferentes cálculos.

Estadísticos (II)

Estadístico	Se calcula como
Valor medio	$\overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$
Mediana	Es el valor central que queda tras ordenar los valores; (semisuma si tengo un número par de valores)
Moda	Es el valor que más aparece
Varianza	$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{x})^2$
Desviación estándar	$\sigma(x_1 \ldots x_N) = \sqrt{\mathrm{Var}(x_1 \ldots x_N)}$
Sesgo	$\mathrm{Skew}(x_1 \ldots x_N) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma}\right]^3$
Kurtosis	$\mathrm{Kurt}(x_1 \ldots x_N) = \left\{\frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma}\right]^4\right\} - 3$

Con estos índices se puede tener una idea de la tendencia central (los tres primeros), de la dispersión (varianza y desviación estándar) y de la forma (sesgo y curtosis)en cuanto a la distribución de la variable.

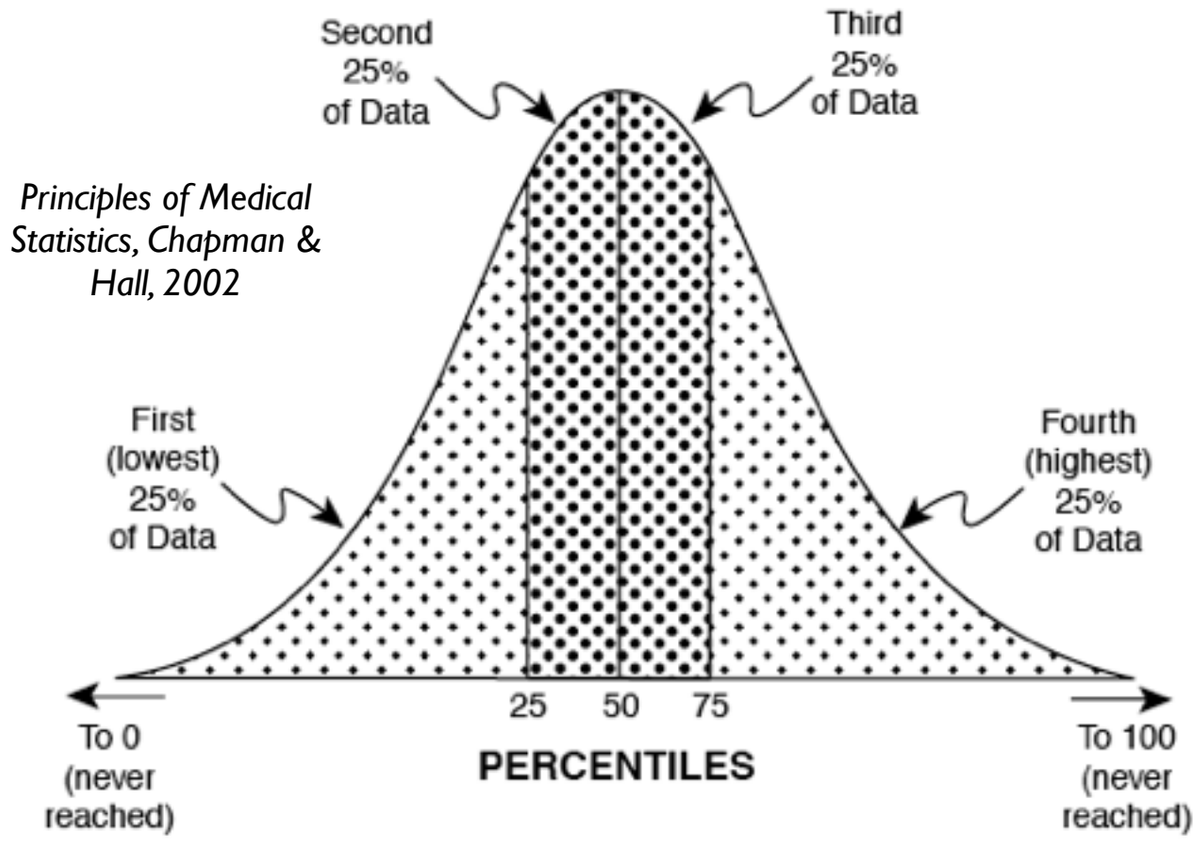
Todos estos índices, se pueden entender a través del histograma. La variable se divide en intervalos regulares y se representa el número de casos en cada intervalo.



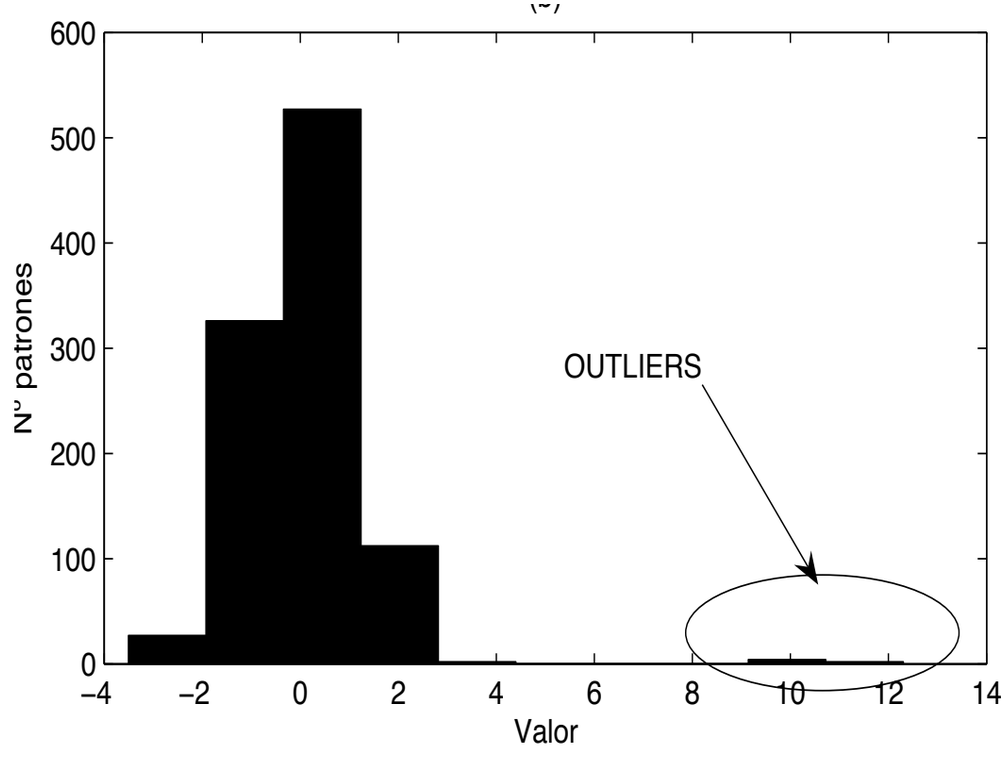
Fundamentals of Probability and Statistics for Engineers, Wiley, 2004

Estadísticos (III)

Estadístico	Se calcula como
Percentil	Es el valor por debajo del cual hay un p% de los valores
Cuartil	Es el valor por debajo del cual hay un p% de los valores (p=25,50,75)
Rango	Diferencia entre el valor máximo y mínimo
Rango Intercuartil (IQR)	Diferencia entre el tercer cuartil y el primero



Principles of Medical Statistics, Chapman & Hall, 2002



Con estos parámetros, junto con el histograma, se pueden determinar los outliers, valores atípicos y que, en la mayoría de aplicaciones, se eliminan.

Valor medio y proporción

De todos los estadísticos destacamos, por su importancia para nosotros, el valor medio y la proporción. Esta proporción se entiende como el número de veces que se da una determinada posibilidad frente al total.

$$p = \frac{N_A}{N_{Total}}$$

Un concepto importante es el de N% intervalo de confianza para algún estadístico p; es un intervalo en el que se se tiene un P% de probabilidad de contener a p.

El valor medio sigue una distribución normal si se conoce la desviación estándar de la población y una t-Student si hay que estimar dicha desviación. El intervalo de confianza viene definido por lo que se conoce como **error estandar de la media (SEM)**. Si consideramos un intervalo de confianza del 95% se tiene:

$$\left[m - 1.96 \cdot SEM, m + 1.96 \cdot SEM\right]$$

$$SEM = \frac{\sigma}{\sqrt{N}}$$

$$\left[m - t_{0.05} \cdot SEM, m + t_{0.05} \cdot SEM\right]$$

$$SEM = \frac{S}{\sqrt{N}} \Leftrightarrow S = \sqrt{\frac{1}{N-1} \cdot \sum_{k=1}^{N} (x_k - m)^2}$$

En el caso de una proporción podemos, en principio, asimilar esta variable a una función de distribución binomial. Se sabe que, si en una distribución binomial el número de elementos de la muestra es alto se puede considerar una distribución normal . conn esta distribución podemos establecer un intervalo de confianza de la siguiente forma (aquí error(n) es la proporción).

$$error_C \pm z_N \cdot \sqrt{\frac{error_C \cdot (1 - error_C)}{N}}$$

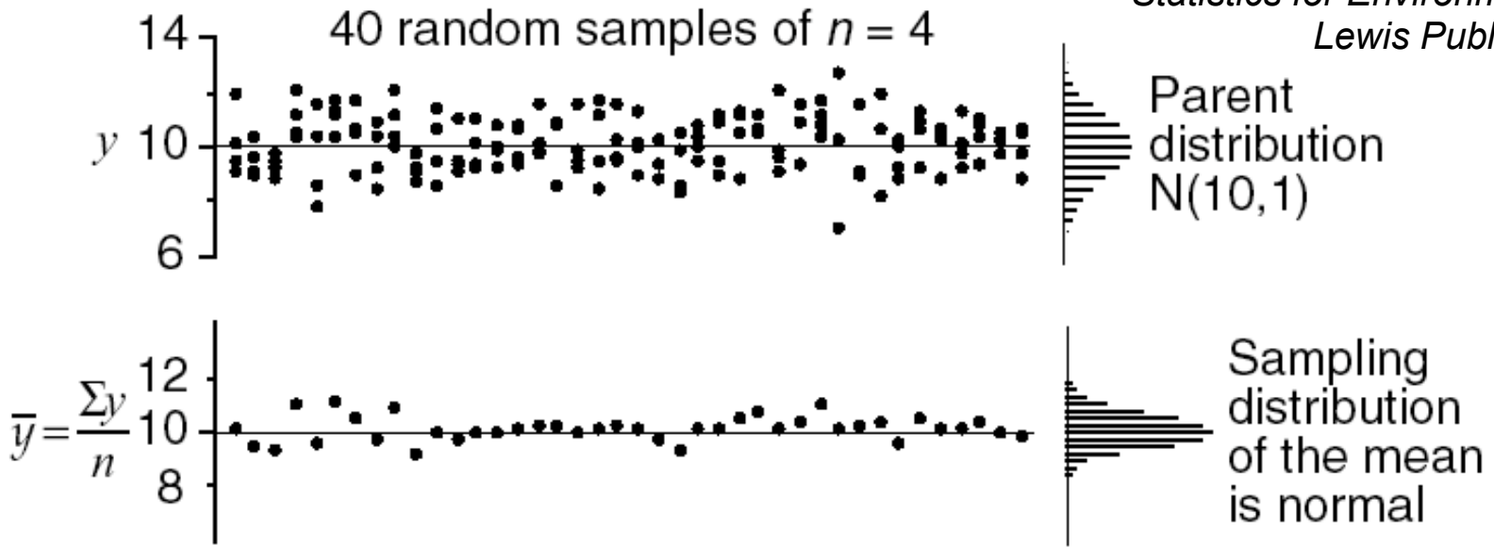
Confianza %	80	90	95	99
z_N	1,28	1,64	1,96	2,58

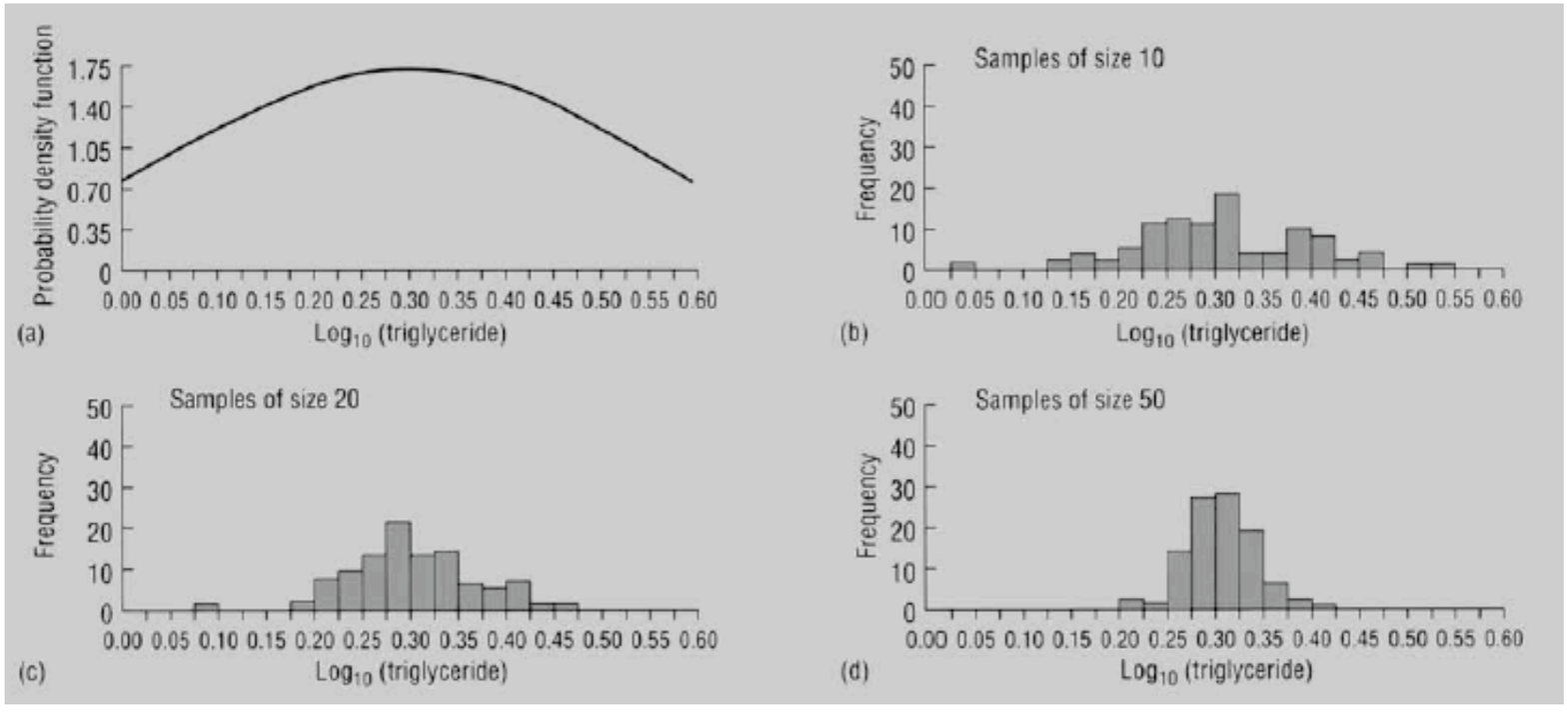
¿Qué relación hay entre este intervalo y el definido para el valor medio?

Valor medio (II)

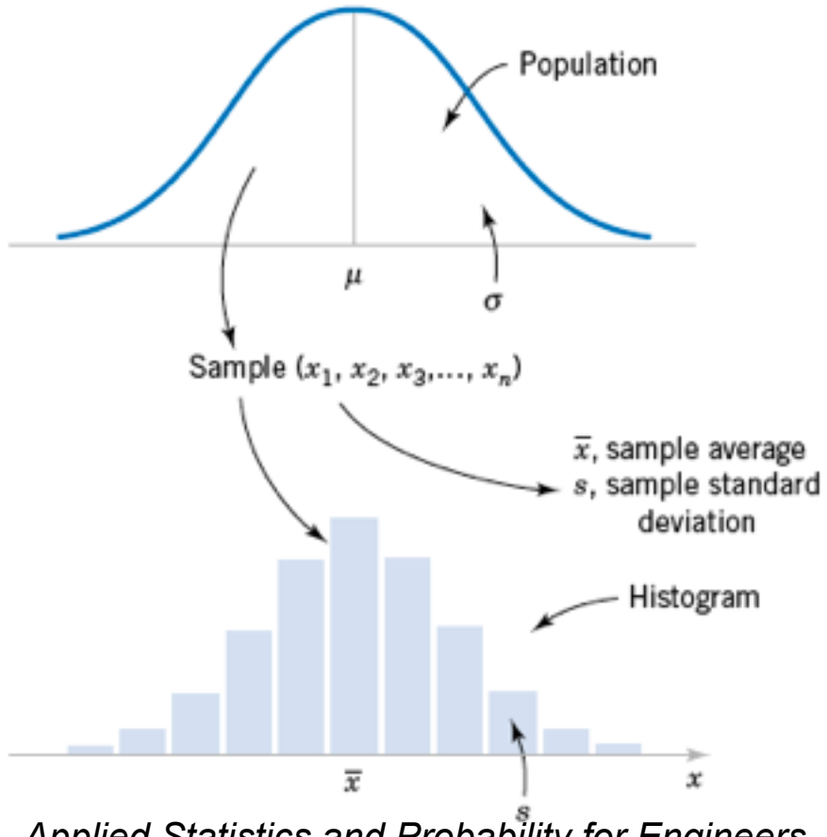
Con estos tres ejemplos gráficos se intenta mostrar lo que supone tomar una muestra y calcular su correspondiente valor medio; una cosa es la población y su valor medio como parámetro y otra son la muestra y su estadístico

Statistics for Environmental Engineers, Lewis Publishers





Medical Statistics at a Glance, Blackwell



Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003

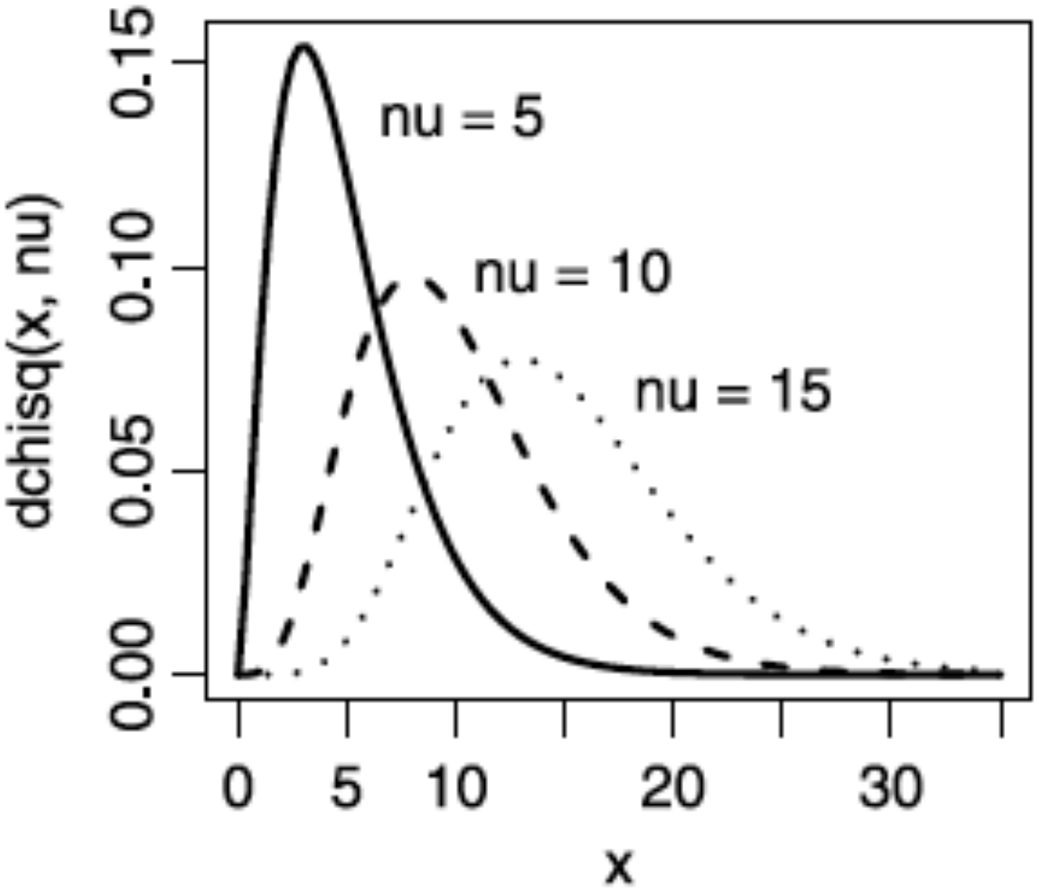
Variables cualitativas (I)

Ahora cabe preguntarse, ¿qué podemos hacer con las variables cualitativas?. Aquí podemos plantear dos tipos de pruebas; la de **homogeneidad e independencia**. En la de homogeniedad se busca determinar si los datos que se tienen son los mismos respecto de la categorización establecida. En la segunda buscamos conocer si las categorías de las filas son indepenedientes de las categorías de las columnas cuando los datos se dispoenen en tablas.

Este tipo de análisis se basa en obtener un estadístico definido por la siguiente cantidad

$$X^2 = \Sigma\left[\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right]$$

Estas cantidades de observado/ esperado hacen referencia a los que uno observa (los datos que se tienen) y a lo que se tendría si se cumplieran las condiciones de homogeneidad o de independencia que se intentan comprobar con este tipo de análisis.



Statistics and Data with R; An Applied Approach Through Examples, Wiley 2008

Este estadístico sigue una distribución de tipo chi-cuadrado. Este tipo de distribución queda caracterizada por un parámetro, ν, conocido como grados de libertad que es igual a su valor medio. La varianza de esta distribución es el doble de dicho parámetro.

Variables cualitativas (II)

¿Existe diferencia en cuanto a hombres para los dos síntomas que se tienen (homogeneidad)?. Es decir en muestra muestra tenemos 32/50 para el síntoma 1 y 28/50 para el síntoma 2; ¿esta diferencia la tengo para la población?

	Sintoma 1	Sintoma 2
Hombres	32	28
Mujeres	18	22
TOTAL	50	50

El siguiente paso sería calcular el estadístico comentado en la anterior transparencia

$$X^2 = \Sigma\left[\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right]$$

	(o-e)	$(o-e)^2$	$(o-e)^2/e$
H-1	2	4	0,133
M-1	-2	4	0,200
H-2	-2	4	0,133
M-2	2	4	0,200

El valor esperado de hombres con el síntoma 1 sería de de 30; tenemos 60/100=proporción de hombres, este valor se multiplica por 50 (número de personas con problemas). El resto de términos se calcula igual. Tendríamos entonces la siguiente tabla de valores esperados.

	Sintoma 1	Sintoma 2
Hombres	30	30
Mujeres	20	20
TOTAL	50	50

El término X^2 queda 0.666; ahora queda por determinar el número de grados de libertad que viene definido por el factor $(c-1)\cdot(f-1)$ donde f y c son el número de filas y de columnas de nuestra tabla. En nuestro caso es una tabla 2x2 por lo que tenemos que el número de grados de libertad es igual a 1. Para 1 grado de libertad y un intervalo de confianza al 95% este valor debería ser mayor que 3.841; no podemos rechazar entonces que las dos relaciones son iguales.

Variables cualitativas (III)

En el siguiente ejemplo estamos interesados en conocer si el lugar donde se juega interviene en el resultado de un partido (problema de independencia).

	Casa	Fuera	TOTAL
Ganador	97	69	166
Perdedor	42	83	125
TOTAL	139	152	291

El siguiente paso sería calcular el estadístico

$$X^2 = \Sigma\left[\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right]$$

	(o-e)	$(o-e)^2$	$(o-e)^2/e$
G-C	17,7	316,3	3,99
P-C	-17,7	316,3	5,30
G-F	-17,7	316,3	3,65
P-F	17,7	316,3	4,84

Hacemos otra tabla igual suponiendo que el lugar de partido no influye en el resultado, por ejemplo en el caso de Ganador-Casa tendríamos, por una parte que, la proporción de los que se ganan es de 166/291; si este factor se multiplica por los que se juegan en casa (139) se tendría 79.3. Si nos fijamos en ganador-fuera tendríamos 166/291 multiplicado por 152 se tendría 86.7. Si nos fijamos en lo perdido se tendría que tenemos una proporción de 125/291; esta proporción hay que multiplicar por los partidos jugados en casa y fuera para acabar la tabla.

	Casa	Fuera
Ganador	79,3	86,7
Perdedor	59,7	65,3

El término X^2 queda 17.78; el número de grados de libertad es igual a 1. Para 1 grado de libertad y un intervalo de confianza al 95% este valor debería ser mayor que 3.841; EXISTE evidencia que la posibilidad de victoria depende del lugar del partido.

Variables cualitativas (IV)

Hemos visto dos ejemplos para tablas 2x2; se puede generalizar dicho resultado para tablas mayores. En el siguiente ejemplo se intenta determinar si los miembros de 3 partidos están de acuerdo con la importancia de la tasas (problema de homogeneidad).

	Very Important 1	2	3	Not Important 4	Total
Democrats	42	26	19	13	100
Republicans	55	21	14	10	100
Independents	38	30	22	10	100
Total	135	77	55	33	300

Test statistic:

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} = number of occurrences in the ijth cell

$$e_{ij} = \frac{(o_{i.})(o_{j})}{o_{..}}$$

$$o_{i.} = \sum_j o_{ij}$$

$$o_{.j} = \sum_i o_{ij}$$

$$o_{..} = \sum_i \sum_j o_{ij}$$

Class	o_{ij}	e_{ij}	$(o_{ij} - e_{ij})^2/e_{ij}$
Democrats			
1	42	45.0	0.200
2	26	25.7	0.004
3	19	18.3	0.027
4	13	11.0	0.364
Republicans			
1	55	45.0	2.222
2	21	25.7	0.860
3	14	18.3	1.010
4	10	11.0	0.091
Independents			
1	38	45.0	1.089
2	30	25.7	0.719
3	22	18.3	0.748
4	10	11.0	0.091

$\chi^2 = 7.425$

Region of rejection:

$$\chi^2 \geq \chi^2_{\alpha,v} \qquad v = (r-1)(c-1)$$

r = number of rows

c = number of columns

$$v = (r-1)(c-1) = (3-1)(4-1) = 6$$

$\chi^2_{0.05,6} = 12.592$

No rechazamos H_0

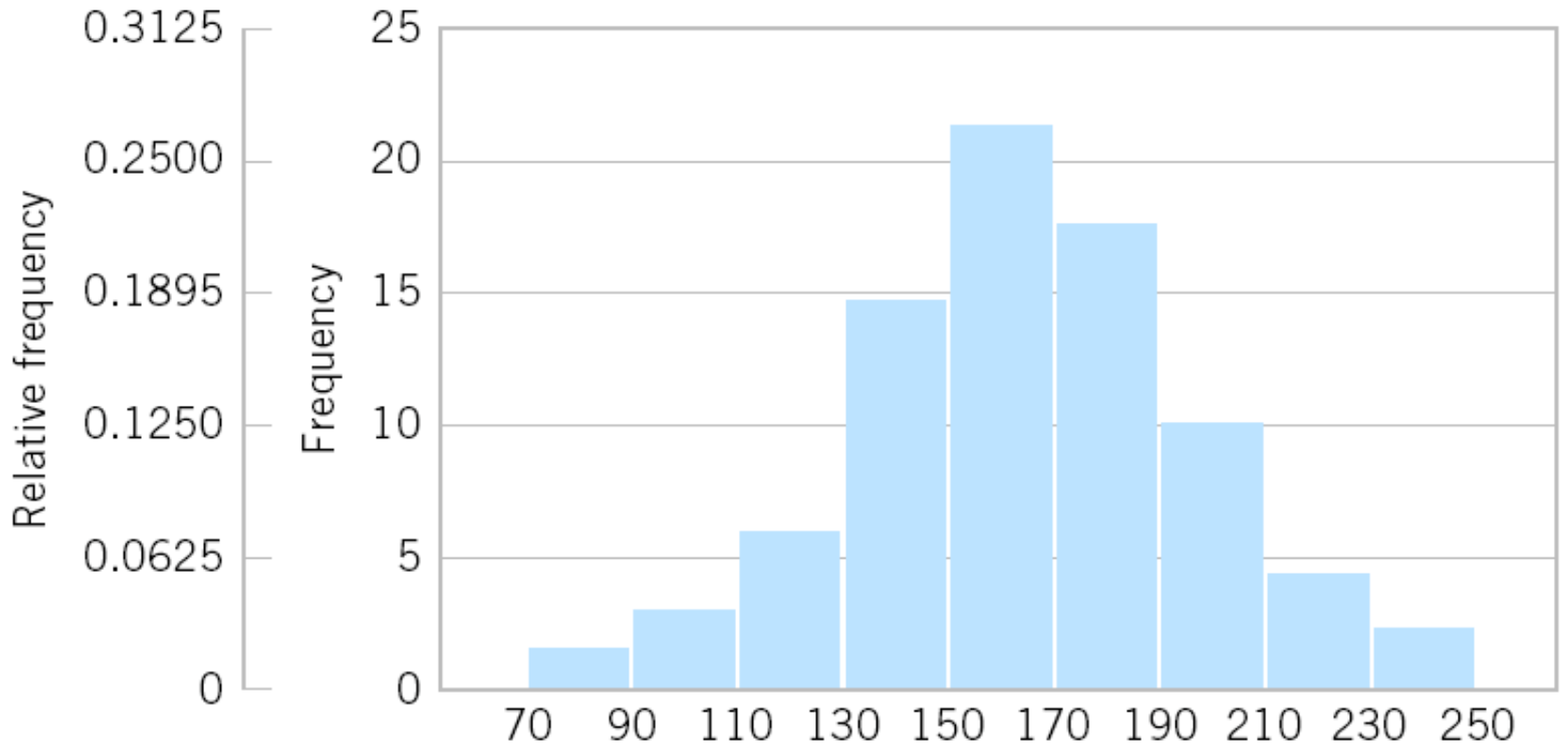
Ejemplo extraído de Statistics for Research. Wilev

Representaciones gráficas (I).

El uso de estadísticos y de representaciones gráficas para la obtención de conclusiones sobre los datos se conoce como Análisis Exploratorio de los Datos (EDA). A continuación se verán gráficas poco conocidas pero que proporcionan mucha información sobre los datos. En prácticas repasaremos todas las representaciones gráficas (diagramas de barras, sectores, líneas, etc).

Histograma, proporciona información gráfica sobre la distribución de los datos, los outliers quedan rápidamente identificados. Al dividir por el número total de datos cada uno de los diferentes intervalos tenemos un "estimador gráfico" de la función densidad de probabilidad.

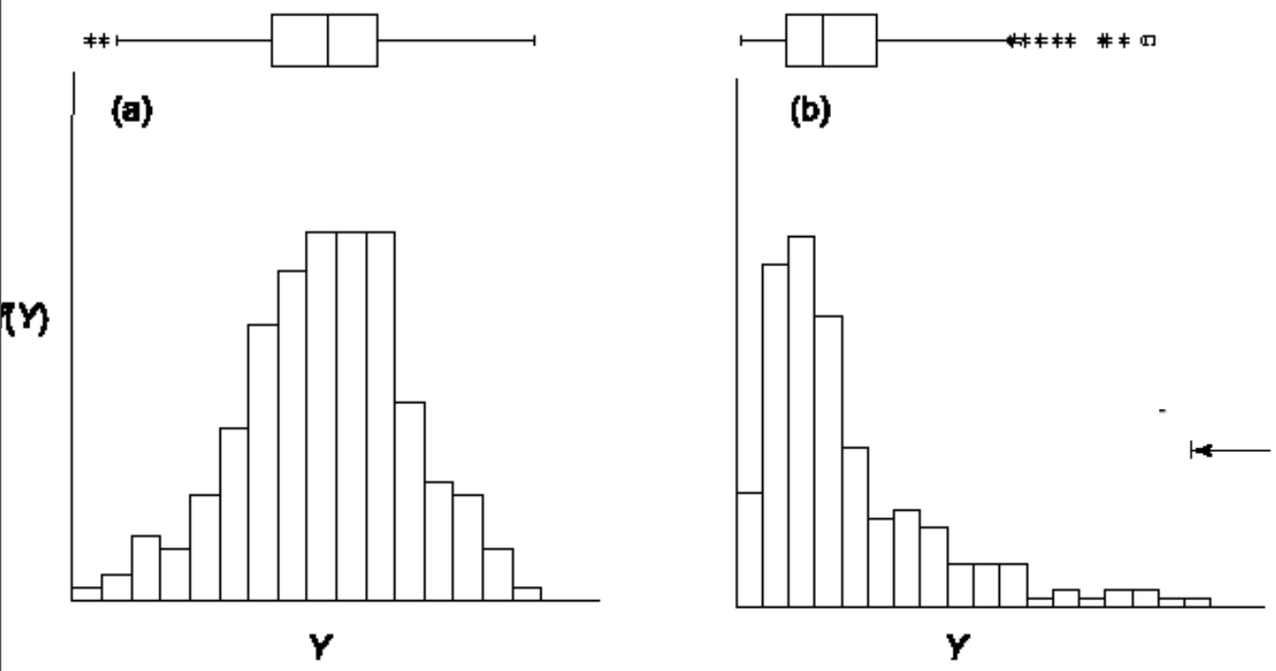
Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000



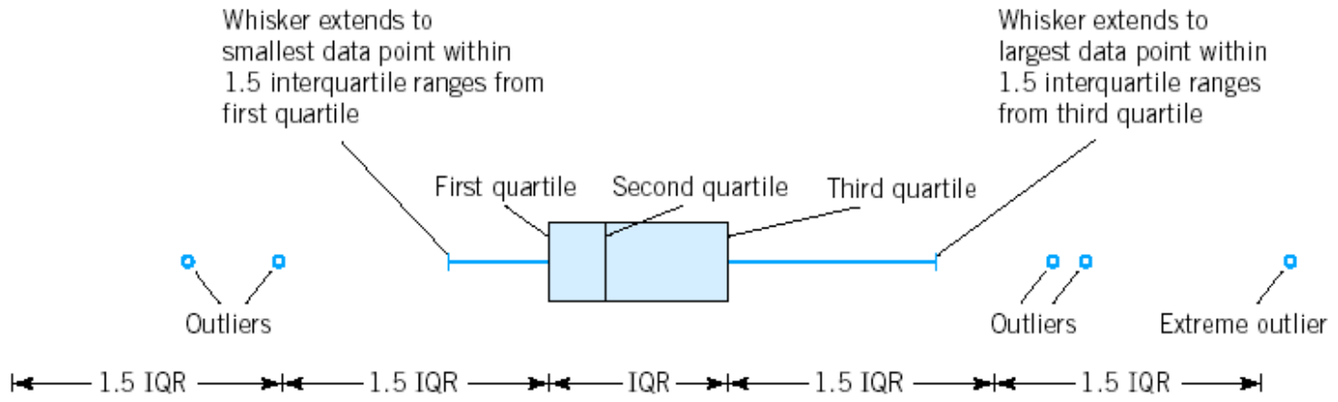
Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003

Representaciones gráficas (II).

Boxplot; proporciona información visual sobre los 3 cuartiles y los valores máximo y mínimo. Estos son los 5 números que describen completamente un conjunto de datos.



Experimental Design and Data Analysis for Biologist, Cambridge University Press



Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003

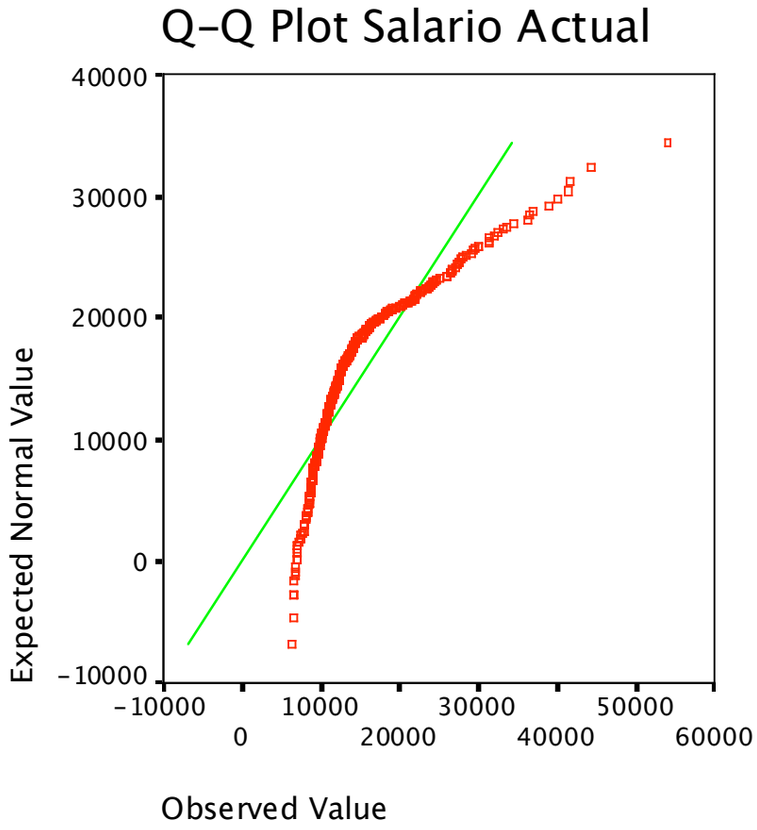
En muchos paquetes informáticos no se representan los valores máximo y mínimo sino que se representan valores por encima/por debajo 1.5 veces el rango intercuartil a partir del tercer y primer cuartil respectivamente. Esto se hace así para evitar problemas cuando existen outliers ya que estos valores podrían provocar que los intervalos fueran muy grandes y poco representativos.

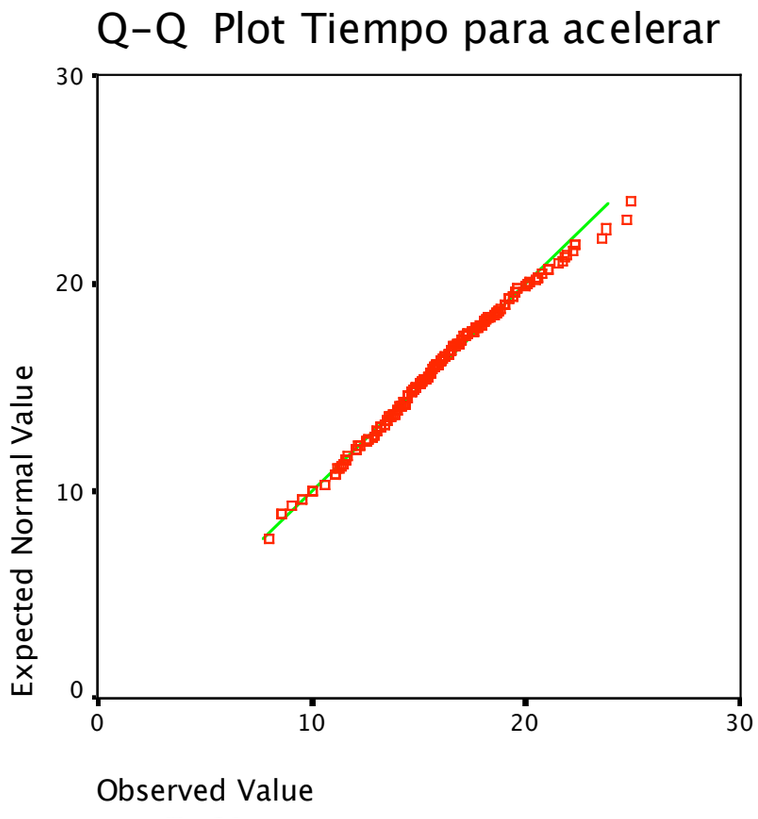
Este tipo de representación es muy útil en problemas de clasificación donde queremos ver si existen diferencias entre dos grupos (especialmente indicado cuando queremos relacionar variables cualitativas con cuantitativas).

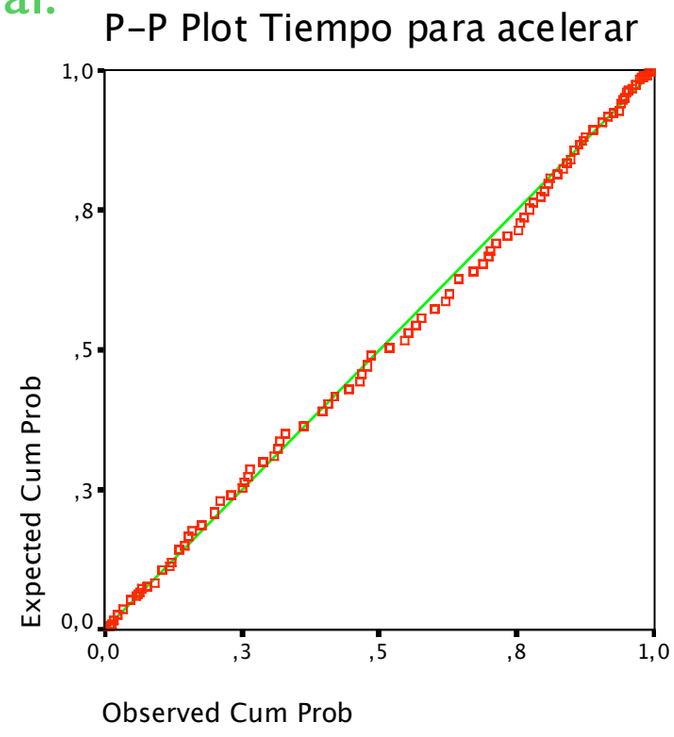
Representaciones gráficas (III).

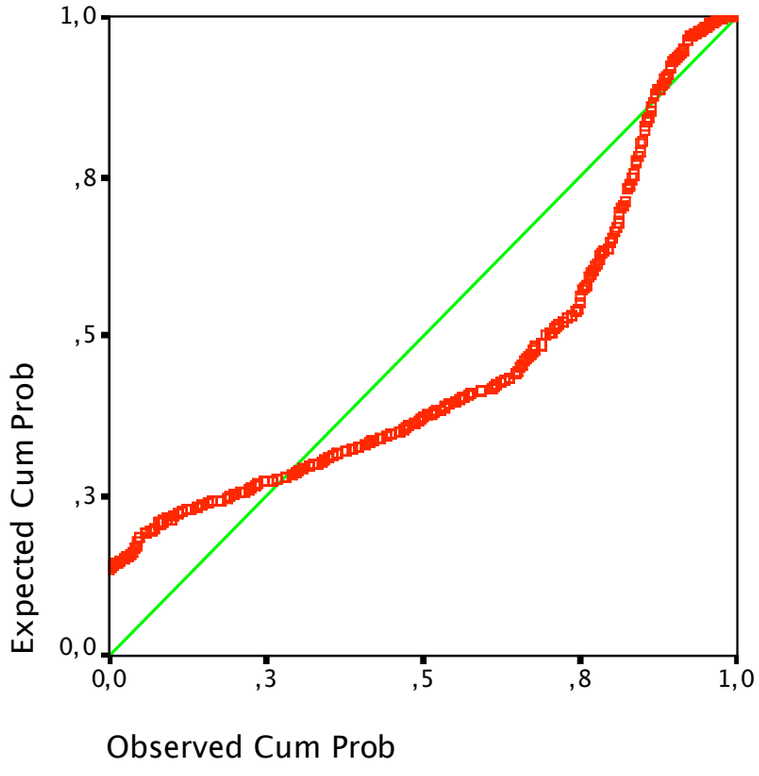
Los gráficos de probabilidad se usan para visualizar si unos datos siguen, o no, una distribución de probabilidad. Los más extendidos son los de normalidad. La idea básica consiste en representar, en un mismo gráfico, los datos que han sido observados frente a los datos teóricos que se obtendrían de una distribución normal. Si la distribución de los datos es una normal los puntos se concentrarán en una línea recta.

Existen 2 tipos de gráficos de probabilidad; en los gráficos P-P se representan las proporciones acumuladas de una variable con las de una distribución normal. Los gráficos Q-Q se obtienen representando los cuantiles de los datos que se tienen respecto a los cuantiles de la distribución normal.









VNIVERSITAT DE VALÈNCIA

MASTER DE INGENIERÍA BIOMÉDICA.

Métodos de ayuda al diagnóstico clínico.

Tema 2: Probabilidad y estadística

Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE