Bioestadística para no estadísticos
Tome decisiones con criterio

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Capítulo 6:

Modelos para variables continuas

Erik Cobo, Jordi Cortés y Roser Rius
Jose Antonio González, Rosario Peláez, Marta Vilaró y Nerea Bielsa

Septiembre 2014

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

equator network

UPC

ELSEVIER DOYMA

MEDICINA CLINICA

TRIALS

Modelos para variable continuas

Presentación **2**

1. Distribuciones continuas **3**

1.1. Probabilidades en variables continuas 3

1.2. Distribución uniforme 4

1.3. Distribución normal 5

1.3.1. Función de distribución F_X 6

1.3.2. Distribución normal tipificada 10

1.4. Exponencial 11

1.5. Ajuste 13

2. Curva ROC **17**

Soluciones a los ejercicios 23

Presentación

Si la variable es continua, la probabilidad de un valor concreto no tiene interés, pero sí las probabilidades acumuladas o las de un intervalo.

La distribución del Gauss-Laplace, llamada en campana o "Normal" es muy útil para representar una gran cantidad de variables. Menos frecuentes, pero más simples, son la uniforme y la exponencial. En este capítulo, mediante ejercicios de dificultad progresiva, el lector se habituará al uso de la distribución Normal.

En Ciencias de la Vida, la variabilidad es la norma, y ciertas diferencias con el valor central son, por definición, "normales", en el sentido de no-patológicas. Por tanto, hay que aprender a valorar qué distancias, por su magnitud, pueden ser sospechosas de patológicas.

Contribuciones: EC y JC escribieron la versión de septiembre de 2013 a partir de los apuntes de EC, RR y JAG de la asignatura de Probabilidad y Estadística de la Facultad de Informática de la UPC, que fue editada por RR. MV, RP y JAG revisaron la versión de enero de 2104 y NB y EC la de septiembre de 2014.

1. Distribuciones continuas

1.1. Probabilidades en variables continuas

Si la variable es continua, la probabilidad de observar un valor concreto es insignificante.

> **Ejemplo 1.1:** La altura es continua. Por tanto, entre 2 señores, uno de 180 y otro de 181 cm siempre podremos encontrar otro. Cada vez intervalos más pequeños. Y así indefinidamente. Por ello, la probabilidad de observar un valor concreto es "infinitamente pequeña".

La probabilidad de un valor concreto es 0, nula. En cambio, la "Función de Distribución" sigue siendo útil.

Definición

La **función de distribución** F_X de una variable continua para un cierto valor x proporciona la probabilidad acumulada hasta ese valor x

Ejemplo 1.1 (cont.): Algún valor de la función de distribución F_X de la altura podría ser:

$F_X(180) = P(X \leq 180) = 0.82$

$F_X(190) = P(X \leq 190) = 0.96$

Ahora, bien, como la probabilidad de un valor concreto es cero:

$F_X(190) = P(X \leq 190) = 0.96 = P(X < 190)$

Es decir, no distinguimos entre "≤" y "<".

Recuerde

En las continuas, **Función de distribución** $F_X = P(X \leq x) = P(X < x)$

Nota: Esta función de distribución como acumulación de probabilidad que es, no puede disminuir, no puede ser menor para valores mayores de X. Quizás no crezca, pero no puede disminuir. Crecerá más en aquellas zonas o intervalos con mayor probabilidad. La derivada o primitiva de una función valora este incremento en un punto concreto. La operación contraria a derivar es integrar. Y una integral es como una suma pero aplicada a funciones continuas.

Tranquilo, no debe recordar los detalles técnicos. Sólo que la derivada de F_X es la función de densidad f_X y valora cuánto crece la probabilidad acumulada F_X.

Definición

La **función de densidad f_X** de una variable continua informa de la intensidad del crecimiento de F_X en un punto concreto de X.

1.2. Distribución uniforme

Ejemplo 1.2: Un paciente ingresado sabe que su médico pasa visita entre las 8 y las 9 am y que decide el orden al azar por dónde empezar, de forma que se espera la misma probabilidad para cada momento del tiempo entre las 8 y las 9 am. La persona acompañante del paciente debe irse a trabajar a las 8h40': ¿Cuál es la probabilidad de que el médico haya pasado antes? Como dispone de 40' sobre un total de 60', P(X<8h40')= 2/3.

Definición

En una variable con distribución **uniforme** entre dos puntos a y b:

$$F_X(x) = P\ X < x\ = \frac{x-a}{b-a}$$

Ejemplo 1.2 (cont.): La Figura 1.1 muestra las formas de sus funciones de distribución y de densidad.

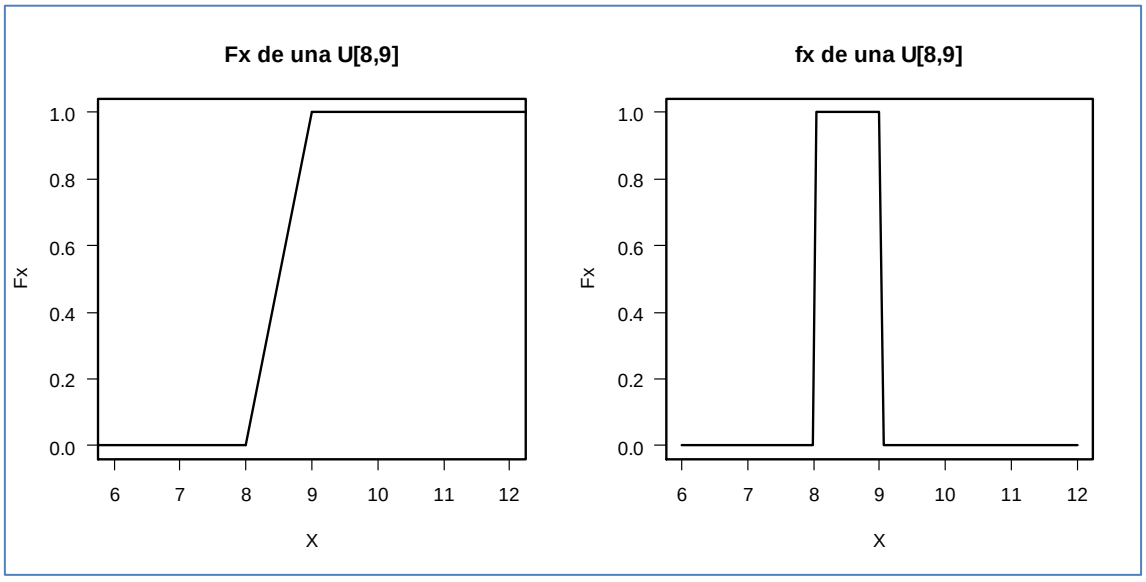


Figura 1.1 Representación de F_X y f_X de la uniforme con a=8 y b=9

Ejercicio 1.1

La llegada de pacientes con la enfermedad E sigue una distribución Uniforme a lo largo del día, entre las 0 y las 24h. Calcule la proporción de pacientes que serán visitados antes de las 8 am y durante el turno de mañana (8 a 15h).

Ejemplo de R

```
# Cálculo de Fx: P(X<7) si X~U(5,10)

> punif(7,5,10)
[1] 0.4
```

1.3. Distribución normal

La distribución Normal tiene la conocida forma de campana o montaña, simétrica alrededor de la media (μ) y con la desviación típica (σ) marcando la distancia entre la media y el punto de máxima pendiente —que marca la *inflexión* o cambio de giro de la curva.

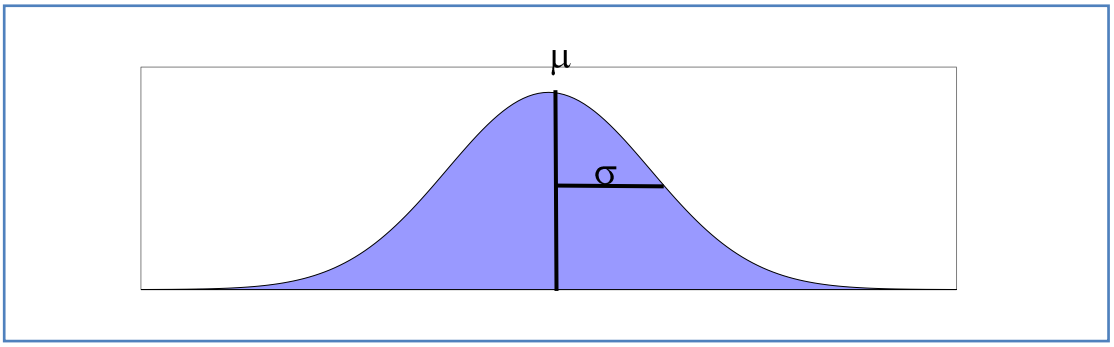


Figura 1.2 Representación de la distribución Normal con esperanza μ y desviación estándar σ

Recuerde

En la Normal, esperanza y desviación típica tienen interpretación visual: μ es el centro; y σ, su distancia al punto de máxima pendiente.

Este modelo matemático reproduce la distribución real de un buen número de variables.

Nota: Recuerde que decir "una variable biológica sigue la distribución Normal" o "la variable es Normal" implican un abuso de lenguaje. Lo correcto sería decir "el modelo Normal reproduce el comportamiento de dicha variable". Disculpen si, por brevedad, usamos expresiones como "variable Normal".

La distribución Normal aparece cuando la variable en estudio es el resultado de la actuación de muchos fenómenos independientes y con igual influencia.

Ejemplo 1.3: La distribución Normal, en sus inicios, fue utilizada para representar la distribución de los errores de medida. Pero no los errores groseros, pocos y evidentes; sino los muchos, pequeños e inapreciables que acompañan ciertos procesos de medida, como la balanza de fiel.

Nota: Las leyes de la combinatoria muestran que la probabilidad de que todos estos pequeños fenómenos actúen en el mismo sentido, generando valores extremos, es muy pequeña. En general, estos efectos se compensan unos con otros y los valores se acercan a una cierta media.

La **máquina de Galton** muestra físicamente la aparición de la distribución Normal cuando confluyen muchos factores aleatorios. Puede ver **varios** vídeos en la red.

Ejemplo 1.4: la altura de los varones adultos y sanos de una determinada población puede aproximarse, razonablemente bien, por la distribución Normal. Para decir que es Normal, ha sido preciso especificar primero la edad, el género y la población, ya que éstas características podrían originar diferencias notables, remarcables. Si, por ejemplo, se mezclan ambos géneros, la distribución resultante tendría dos montañitas o jorobas que marcarían los intervalos modales de hombres y mujeres.

Nota: La dispersión de los valores de la distribución Normal es, por tanto, el resultado de establecer un modelo sobre el elevado número de fenómenos con muy pequeña influencia. Éstos son tantos y tan pequeños que no aportan información y representan el "ruido".

Recuerde

La media μ de la Distribución Normal representa la señal "relevante"; y la desviación típica, las oscilaciones "irreproducibles".

Notación

Representamos el modelo Normal de parámetros μ y σ por $N(\mu, \sigma)$

Ejemplo 1.5: La altura de los varones adultos sanos es N(170 cm, 8 cm)

1.3.1. Función de distribución F_X

Historieta: Hubo épocas en las que aquí se explicaban tablas como éstas. Ahora, gracias a R, Vd. se las ahorra.

α	0'001	0'01	0'05	0'10	0'20	0'32
α/2	0'0005	0'005	0'025	0'05	0'10	0'16
Z	3'29	2'58	1'96	1'64	1'28	1

z	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9
0,0_	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1_	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2_	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3_	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4_	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5_	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6_	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7_	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8_	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9_	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0_	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1_	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2_	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3_	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4_	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5_	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6_	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7_	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8_	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9_	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0_	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1_	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2_	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3_	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4_	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5_	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6_	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7_	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8_	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9_	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0_	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1_	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2_	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3_	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4_	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5_	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6_	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,7_	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,8_	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Ejemplo de R

```
# Cálculo de Fx: P(X≤180) si X ~ N(170 cm, 8 cm)
> pnorm(180,170,8)
[1] 0.8943502
# Casi un 90% miden menos de 180 en una población ~ N(170 cm, 8 cm)
```

Ejercicio 1.2

a) Calcule con R las probabilidades de encontrar alguien que mida menos de 170; menos de 162, y menos de 154 cms.

b) Haga un dibujo y represente las probabilidades anteriores. Sin necesidad de acudir a R, deduzca las probabilidades de medir más de 170, más de 178 y más de 186 cm.

c) Sin necesidad de recurrir a R, calcule las probabilidades de medir entre 162 y 178. Y entre 154 y 186 cms.

Ejemplo de R

```
# Cuantil: k tal que 0.9 = P(Y≤k) si ~N(170cm, 8cm)
> qnorm(0.90, 170, 8)
[1] 180.2524 cms
# 180.25 es aquél valor que deja debajo el 90% de los casos
```

Ejercicio 1.3

a) Suponga que N(170,8) es la distribución de la altura de las pacientes. Si quiere garantizar que el 99% cabrán sin tener que doblar las piernas, las camas deben medir…

b) Suponga ahora que en ciertas condiciones hormonales, la altura se hace mayor. Si quiere establecer un umbral (cut-point) que tenga una especificidad (% de sanos que dan negativo) del 95%, ¿cuál sería este valor?

c) Suponga también que otras condiciones hormonales provocan valores bajos y debe establecer 2 límites "de normalidad" con la misma especificidad. ¿Qué valores serían?

Los ejemplos y ejercicios anteriores muestran que, si se toma una vez hacia arriba y una vez hacia abajo el valor de la desviación típica (±1σ), se engloba el 68% de las observaciones. Y si en lugar de hacer una vez el valor de la desviación típica, se toma dos veces dicho valor (±2σ), se incluye al 95% de las observaciones.

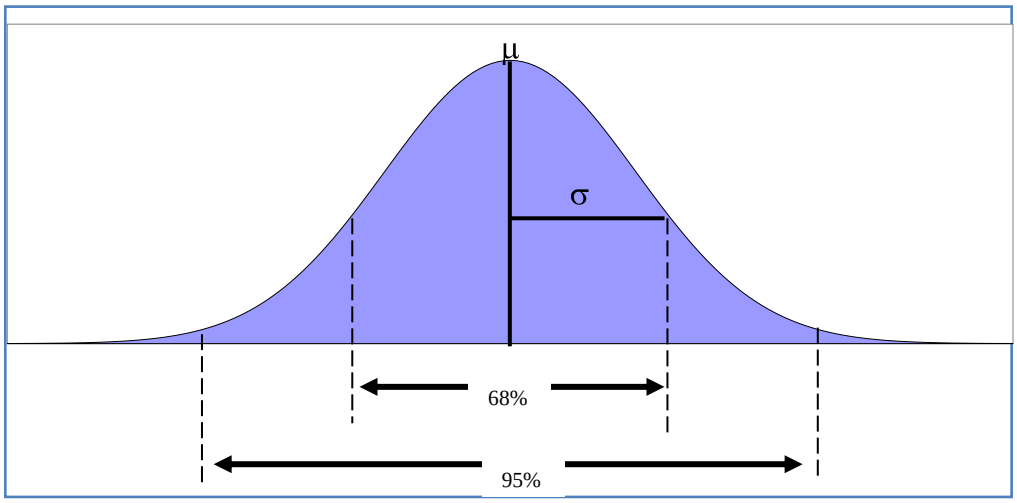


Figura 1.3 Representación de las regiones que contienen el 68% y el 95% de las observaciones en una distribución Normal con media μ y desviación estándar σ.

Recuerde

Más y menos 2 veces σ alrededor de μ contiene el 95% de los casos.

Hay pues 2 aplicaciones complementarias de las funciones de distribución: (1) encontrar la probabilidad acumulada hasta un cierto valor; y (2) encontrar el valor al que corresponde una probabilidad acumulada.

Recuerde

Hay 2 usos recíprocos: (1) dado el valor X, calcular las probabilidades que delimita; y (2) dadas ciertas probabilidades, calcular el valor X que las limita.

En la Normal, el 1º se obtiene con pnorm, y el 2º con qnorm.

Ejemplo 1.6: ¿Cuál es el límite de la glicemia que deja por encima el 5% de los sanos?

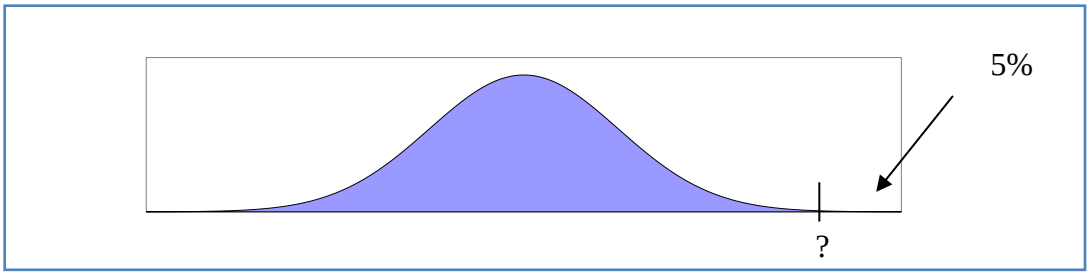


Figura 1.4 ¿Qué valor deja por encima el 5% de la distribución?

Ejemplo 1.7: Un paciente tiene, en cierta prueba, índice o escala (por ejemplo, de inteligencia) una puntuación de 112 unidades. Este valor no aporta nada a un inexperto en dicha prueba, pero sí que lo haría decirle que ocupa el percentil 70, es decir, que un 70% de las unidades de su población tiene puntuaciones inferiores.

Ejercicio 1.4

Un estimulador tiene un umbral con cierta variabilidad: unos voluntarios responden ante un estímulo de unos voltios; y otros, de tantos voltios. . La distribución del umbral en los sanos es aproximadamente normal con una media de 5 voltios y una desviación típica de 0.5.

Rellene los siguientes espacios en blanco:

a) El 95% de los voluntarios tienen un umbral que se sitúa entre __y__ voltios.
b) En el 95% de los voluntarios, el umbral se sitúa por encima de ___ voltios.
c) En el 95% de los voluntarios, el umbral se sitúa por debajo de ___ voltios.
d) El 90% de los voluntarios tienen un umbral que se sitúa entre _ y _ voltios.

e) En el 84% de los voluntarios, el umbral se sitúa por encima de ____voltios.
f) En el 84% de los voluntarios, el umbral se sitúa por debajo de ____voltios.
g) ¿Cuál es la probabilidad de que el umbral supere 6.3 voltios?
h) ¿Cuál es la probabilidad de que un voluntario tenga un umbral entre 4.5 y 5.5?

Ejercicio 1.5

En unidades del Sistema Internacional, el cloruro plasmático tiene unos límites de "normalidad" de 95 y 105 mmol/l.

a) ¿Es posible que una persona sana supere estos límites?
b) ¿Cuál cree Vd. que es el valor de la media y de la desviación típica de esta variable en los "normales"?
c) ¿Existe alguna condición (premisa) para este cálculo?
d) Para la Ferritina, estos límites son 15-200µg/l ¿Cómo se imagina su distribución?

Ejercicio 1.6

Busque variables relacionadas con su trabajo que presumiblemente sigan una distribución normal.

Ejercicio 1.7

Invente aplicaciones "útiles" para las variables del punto anterior. Invente condiciones o situaciones en las que sea razonable que las variables del ejercicio anterior dejen de seguir una distribución normal.

1.3.2. Distribución normal tipificada

Como hemos visto, las probabilidades de la distribución Normal dependen de su media y desviación típica. Para comparar diferentes variables, es interesante disponer de un resultado "estandarizado" o tipificado.

Definición

El desvío tipificado Z se obtiene:

$$Z = \frac{Valor - Media}{desviación\_típica} = \frac{X - \mu}{\sigma}$$

Z tiene media 0 y desviación típica 1: está "reducida".

Historieta: Un marciano al que ha conocido por internet le cita en la plaza de su ciudad y le dice: "ya me verás, mido 160 cms". Primero Vd. piensa "será un marciano bajo", pero luego cae en cuenta de que no conoce la media de sus alturas. Se la pregunta y le dice que es 150 cms. "Vale, es un marciano alto", razona. Pero "¿sobresale o es un alto típico?". Y ahora le pregunta σ, que resulta ser 2 cms. Y Vd. interpreta: "destaca". Así es: su mayor altura es 5 veces la distancia típica. Vamos, que si fuera terrícola, donde σ=8, ¡se distanciaría 40 cms de la media!

Ejercicio 1.8

En la distribución Normal tipificada, Z~ N(0, 1) , ¿qué proporción de casos quedan por encima de -1.96 y por debajo de +1.96?

Recuerde

En la Normal tipificada, Z, "±1.96" (o redondeado: "±2") son los límites que contienen el 95% de las observaciones.

Como Z tiene media 0, valores negativos representarán observaciones por debajo de la media; y como su desviación típica es 1, una observación prototípica se aleja de la media, por arriba o por debajo, en 1 unidad.

Ejercicio 1.9

¿Qué proporción de casos están por encima de z = 1.66? Es decir, ¿cuál es la probabilidad de que Z > 1.66?

Ejercicio 1.10

Un gabinete psicológico valora los resultados de la inteligencia abstracta A según una escala N(100, 15) y la emocional E según una escala N(1000, 10). Un paciente tiene A=120 y B=1020. Vd. observa que ambas inteligencias están por encima de la media. Pero relativo a sus conciudadanos, ¿destaca más en A o en E?

1.4. Exponencial

El modelo de Poisson permite, a partir de una tasa λ de eventos por unidad de tiempo, modelar la probabilidad de observar *x* casos en esa unidad de tiempo: P(X=x). El modelo exponencial, a partir de la misma tasa λ de eventos por unidad de tiempo, modela la probabilidad de que el tiempo T hasta el próximo evento sea menor que un cierto valor t: P(T<t).

Ejemplo 1.8: En Barcelona, el número diario de accidentados con lesiones craneoencefálicas vale $\lambda=1$ casos/día. La exponencial permite calcular que el tiempo hasta el siguiente evento será inferior a 1 día en un 63.21% [P(T<1)=0.6321206] y a 2 días en un 86.5% [P(T<2)= 0.8646647].

Notación

Representamos el modelo Exponencial por E(λ)

Recuerde

Como en la Poisson, en el Exponencial la tasa λ indica casos/tiempo.

Ejemplo de R

```
# Cálculo de F_X(2): P(T<2) si X~E(1)
> pexp (2,1)
[1] 0.8646647
```

Ejercicio 1.11

Si la tasa diaria de traumatismos craneoencefálicos vale 1, ¿qué proporción de veces estaremos 3 o más días sin observar ninguno?

Ejemplo de R

```
# Cuantil: k que cumple 0.95=P(T≤k) si X~E(1)
> qexp(0.92,1)
[1] 2.995732
# Note la concordancia con el resultado del ejercicio anterior
```

Ejercicio 1.12

Cierto equipo anota 50 canastas por hora de juego. Si Vd. desea garantizar con una seguridad del 95% que antes de un tiempo t ya habrán anotado 1 canasta, ¿cuánto vale este tiempo t?

Las expresiones de la esperanza y la varianza de una variable Exponencial son sencillas:

Fórmulas

Si $T \sim E(\lambda)$, $E(T) = 1/\lambda$ y $V(T) = 1/\lambda^2$

Ejercicio 1.13

En unidades por semana, la tasa del número de traumatismos craneoencefálicos es 7 casos/sem. ¿Cuál es el valor esperado del tiempo hasta el próximo?

Como en el modelo de Poisson, la premisa más importante del modelo Exponencial es que λ es constante: el proceso no tiene memoria.

Ejemplo 1.9: El hecho de que llevemos tanto tiempo sin que nos toque la lotería no aumenta ni disminuye la probabilidad de que nos toque en el siguiente sorteo.

Recuerde

El azar no tiene memoria.

1.5. Ajuste

Cita

Todos los modelos son falsos, pero algunos son útiles (**George Box**).

Historieta: La Ciencia ha abandonado la soberbia de emular al Hacedor y escribir las leyes del Universo. Su objetivo, sólo ligeramente más modesto, es proponer modelos que permitan reproducirlo y, el de la técnica, mejorar las condiciones en que lo disfrutamos.

La pregunta de interés es: "si actúo como si el modelo fuera correcto, ¿cuál es la magnitud de mis errores?"

Figura 1.5 Viñeta representativa de la cita de **George Box**

Definición

La **bondad del ajuste** describe la similitud entre un modelo estadístico y unos datos.

La Figura 1. superpone las funciones de densidad observadas (sombreado fuerte) con las teóricas (sombreado claro) de una Normal con media y varianza igual a las observadas. En el primer caso los datos provienen realmente de una Normal, pero en el segundo, de una variable muy asimétrica; y en el tercero, de una uniforme.

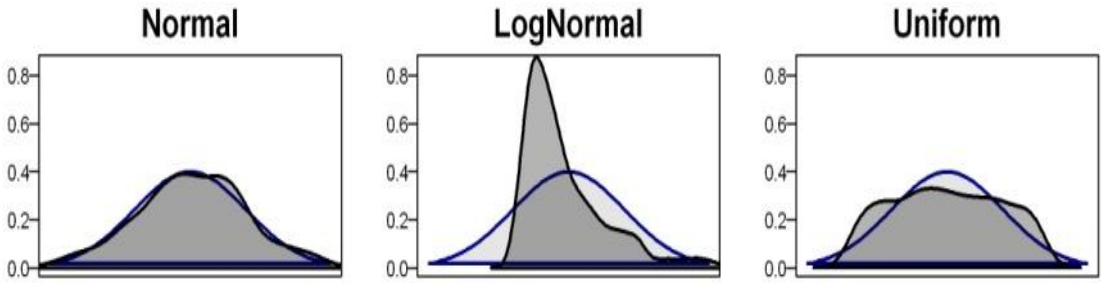


Figura 1.6 Funciones de densidad observadas y teóricas

Nota: Este gráfico es muy visual, pero poco estable y difícil de valorar.

También interesa disponer de medidas que permitan valorar la calidad del ajuste a nivel global, es decir, a lo largo de toda la distribución. Disponemos de 2 medidas populares.

Definición

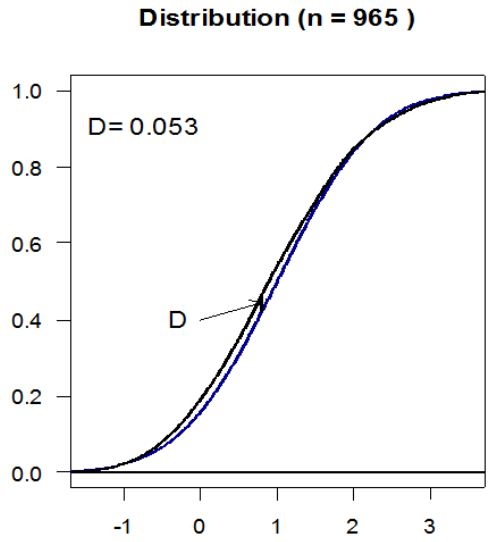
La **distancia de Kolmogoroff** es el valor máximo de la diferencia, para todos los puntos de la variable, entre la Función de Distribución teórica y la probabilidad acumulada observada.

El **estadístico Shapiro-Wilks** es la correlación entre el cuantil teórico y el observado.

Ambos toman valores entre 0 y 1, pero el primero es una medida del desajuste, con mayores valores cuanto menor es el ajuste.

Nota: Más adelante consideraremos su fluctuación en las muestras. Por ahora, veremos el significado de estas medidas.

Un análisis gráfico más fino consiste en superponer las funciones de distribución, como hicimos entre Binomial y Poisson. Ahora en lugar de 2 modelos teóricos, enfrentaremos modelo con datos.



Ejemplo 1.10: La Figura 1. muestra las probabilidades acumuladas de una N(1,1) y las proporciones acumuladas de una muestra, grande, de n=965 observaciones aleatorias extraídas de este modelo. Como el modelo es correcto, el desajuste, que puede ser explicado por el azar, es pequeño: el estadístico D de Kolmogoroff vale 0.053, indicando que la probabilidad acumulada observada difiere 5.3% de la teórica en el punto de mayor desajuste.

Figura 1.7 La diferencia máxima entre probabilidades acumuladas teóricas y las proporciones acumuladas observadas es 0.053.

Recuerde:

La diferencia máxima entre las probabilidades teóricas y las proporciones observadas (ambas acumuladas) es la D de Kolmogoroff.

Un tercer gráfico, menos intuitivo pero más visual, enfrenta cuantiles de la variable tipificada en lugar de probabilidades acumuladas (Figura 1.). Un eje muestra el cuantil observado; y otro, el teórico. Por ejemplo, un punto concreto enfrenta el valor observado del caso que deja un 50% de las observaciones por debajo (la mediana) con el valor que tendría, en una N(0,1), el caso con F_X=0.5 (el "0"). Es más fácil de interpretar porque, si el ajuste es bueno, observaremos una línea recta.

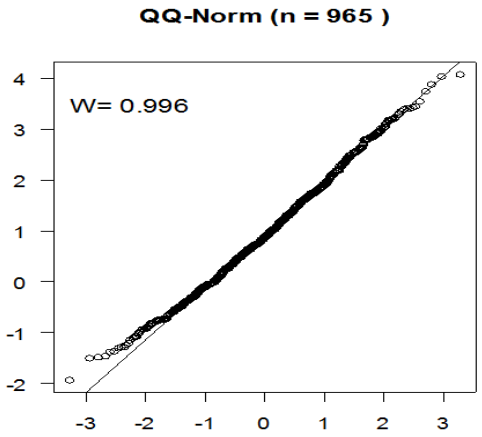


Figura 1.8 El estadístico de Shapiro-Wilks valora la correlación entre los cuantiles teóricos y los observados.

Definición

El **gráfico QQ o QQ-plot** enfrenta los cuantiles observados con los teóricos.

Nota: Conocido originalmente por **recta de Henry**, cuando estudia el ajuste a la Normal, recibe también los nombres de 'gráfico de probabilidad normal' y QQ-norm, como en la figura anterior.

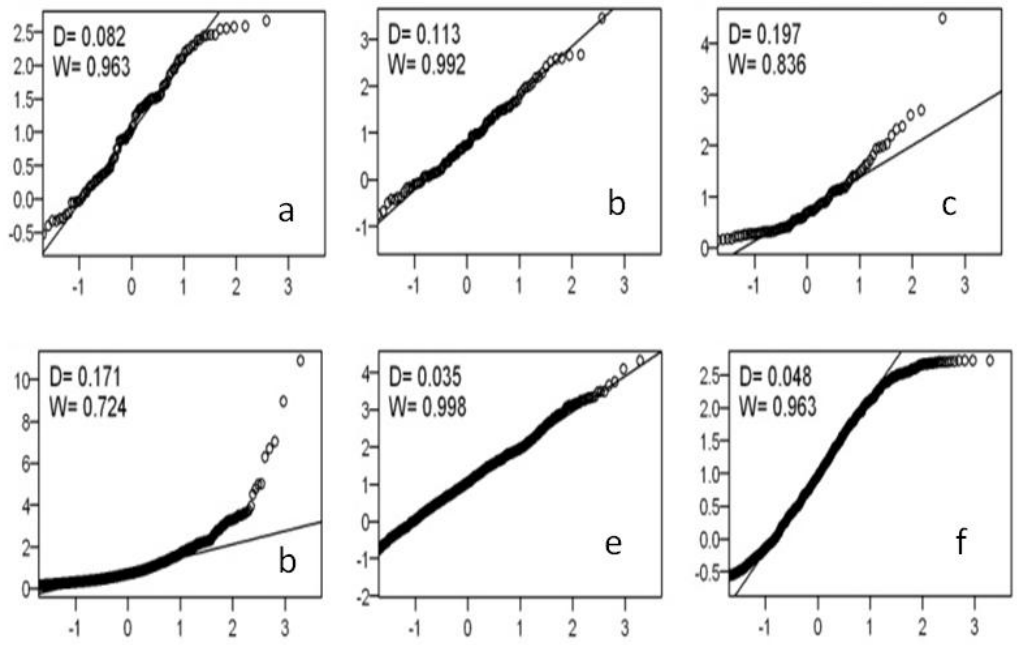
Ejemplo 1.11: La Figura 1. muestra que el QQ-norm en este ejemplo ajusta muy bien a una línea recta. Nótese la menor estabilidad en los extremos. La medida de Shapiro-Wilk cuantifica esta correlación entre cuantiles observados y teóricos en W=0.996l, muy cerca de 1, su valor máximo.

Recuerde

Si el ajuste es bueno, el **QQ-norm** mostrará una recta, **D** será próximo a 0 y **W** a 1.

Ejercicio 1.14

Las 6 figuras muestran, para 2 tamaños muestrales, n=100 en la primera fila y n=1000 en la segunda, los QQ-norm de diferentes variables y las medidas de Kolmogoroff y Shapiro Wilks. Diga cuál es el peor ajuste en cada fila según el gráfico y los valores de las medidas D y W.

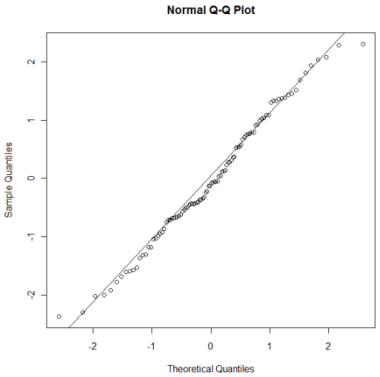


Ejemplo de R

```
# ks.test y shapiro.test proporcionan también p valores. Vd debe interpretar sólo D y W.
# Obtención de QQ-norm, D y W
> x <- rnorm(100)
> qqnorm(x)
> qqline(x)

> ks.test(x,pnorm)
One-sample Kolmogorov-Smirnov test
data:  x
D = 0.0856, p-value = 0.456
alternative hypothesis: two-sided

> shapiro.test(x)
Shapiro-Wilk normality test
data:  x
W = 0.9868, p-value = 0.4256
# La instrucción rnorm(100) proporciona 100 números aleatorios con distribución Normal estándar.
```



2. Curva ROC

Muchos indicadores pueden tomar más de 2 valores. Sean ordinales o numéricos, la definición de sensibilidad y especificidad requiere establecer un límite o umbral ('*cut-point*') que separe el conjunto de resultados en dos grupos, positivo y negativo.

> **Ejemplo 2.1:** Un posible ejemplo es el resultado de una prueba que mide la concentración de glucosa en plasma, en condiciones basales. Dicho resultado, expresado en mg/dl, puede ser muy variado: 50, 75, 110, 128, 165, 192, etc. Ninguna cifra de éstas es, por sí misma, ni positiva ni negativa.

Sin embargo puede ser útil considerar que cifras de 100 o superiores definen un resultado positivo; y las inferiores, negativo.

Recuerde

En los indicadores numéricos, es común establecer un **umbral**.

Ejemplo 2.2: el Ejercicio 1.4 dice que el límite de estimulación de los voluntarios sanos sigue una N(5, 0.5). Supongamos, además, que en cierto tipo de enfermos sigue una N(6, 0.5). Figura 2.1

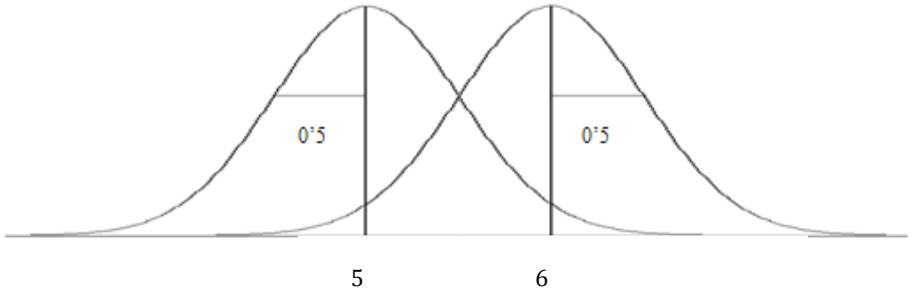


Figura 2.1 Distribución del umbral de estimulación en sanos y enfermos

Si el criterio diagnóstico se establece en 5.5, los valores de sensibilidad y especificidad serán:

Sens = P(+|E) = P(Y > 5.5 | Enfermo) = P(z > (5.5-6)/0.5) = P(z > -1) ≈ 84.13%

Esp = P(-|S) = P(Y < 5.5 | Sano) = P(z < (5.5-5)/0.5) = P(z < 1) ≈ 84.13%

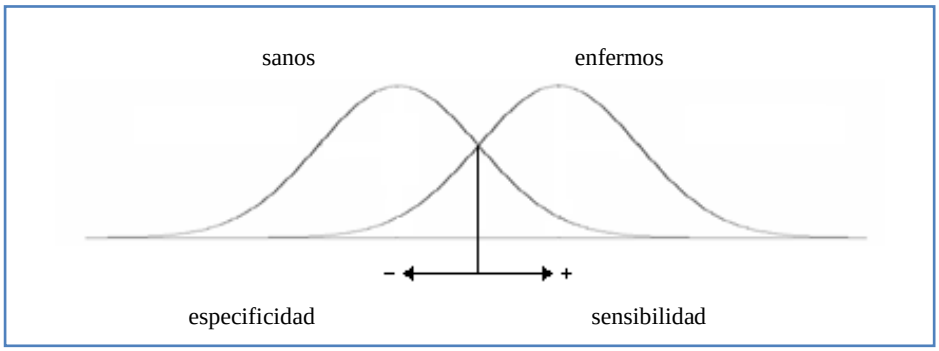


Figura 2.2 Sensibilidad es la proporción de la curva de enfermos que queda por encima del criterio diagnóstico y especificidad la de sanos que queda por debajo

En cambio, si el criterio se hubiera establecido en 5.2, serían:

$$\text{Sens} = P(+|E) = P(Y > 5.2 \mid \text{Enfermo}) = P\left(Z > \frac{5.2-6}{0.5}\right) = P(z > -1.6) \approx 94.52\%$$

$$\text{Esp} = P(-|S) = P(Y < 5.2 \mid \text{Sano}) = P\left(Z > \frac{5.2-5}{0.5}\right) = P(z < 0.4) \approx 65.54\%$$

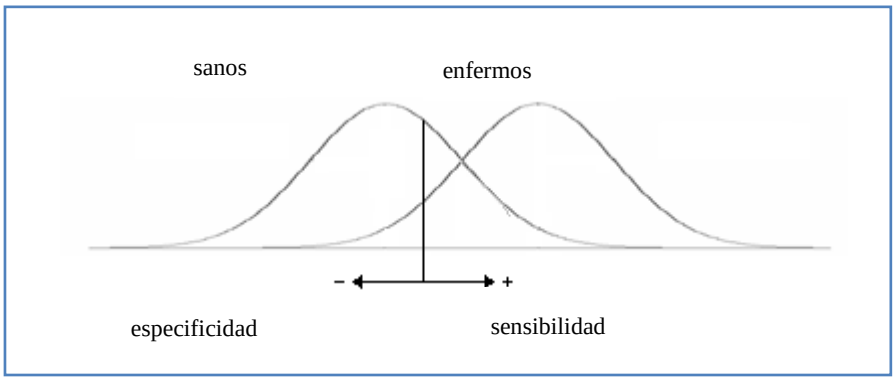


Figura 2.3 Al desplazar el umbral hacia la izquierda aumenta la sensibilidad y disminuye la especificidad

Moviendo el punto de corte se cambian los valores de especificidad y sensibilidad. Si se desea aumentar la sensibilidad, la especificidad disminuye. Y viceversa. Nótese que habrá tantos "pares" de valores de sensibilidad y especificidad como posibles puntos de corte. Cada indicador diagnóstico tiene unos pares de valores de sensibilidad y especificidad que le "caracterizan".

Lectura: Receiver Operating Characteristic (ROC) curves: a plot of the sensitivity of a diagnostic test against one minus its specificity as the cut-off criterion for indicating that a positive test is varied. Often used in choosing between competing tests, although the procedure takes no account of the prevalence the disease being tested for.

Nota: Vea este video. Mejor ahora, son sólo unos minutos.

Ejercicio 2.1 (extraído de Radiology 1982; 143: 29-36)

Se desea estudiar la calidad diagnóstica que ofrece la evaluación por radiólogos de una tomografía computarizada. Se dispone de la prueba de referencia y de 109 pacientes. La tabla de frecuencias es:

		Prueba de referencia	
		Sano=58	Enfermo=51
Clasificación de la tomografía computarizada (prueba índice)	**1** Seguramente normal	33	3
	2 Probablemente normal	6	2
	3 Dudosa	6	2
	4 Probablemente anormal	11	11
	5 Seguramente anormal	2	33

Calcule las proporciones de casos positivos en los enfermos y en los sanos si sitúa el umbral en el máximo (declarar positivo sólo si resultado = 5). Ídem si fuera al revés (negativo sólo si es 1).

Recuerde

Al bajar el umbral aumentan los positivos. Tanto en enfermos como en sanos.

Una vez más, bajar el umbral, implica aumentar la sensibilidad. Pero también, bajar la especificidad.

Definición

La curva característica (ROC: *Receiver Operating Characteristic*) dibuja los pares de las proporciones de positivos en las 2 muestras. Cada umbral marca un par.

Ejercicio 2.2

Defina la curva ROC en términos de sensibilidad y especificidad.

Ejercicio 2.3 (Cont. Ejercicio 2.1)

¿Qué umbral proporcionaría una sensibilidad del 100%? ¿Y una especificidad del 100%?

Convierta la tabla de frecuencias dada en una nueva que contenga el número de casos bien y mal clasificados para cada punto de corte en cada muestra.

Calcule la proporción de positivos para cada posible punto de corte en las 2 muestras, sanos y enfermos.

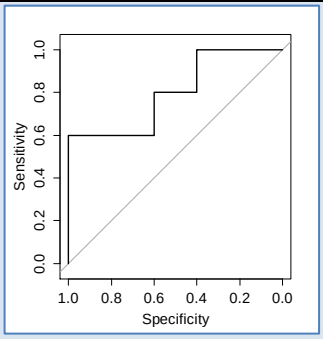
Calcule sensibilidad y especificidad para cada punto de corte.

Nota: Más adelante veremos cómo elegir el 'mejor' umbral. Ahora los estudiamos todos.

La curva ROC pone la proporción de positivos en los enfermos (sensibilidad) en el eje vertical de ordenadas y la de positivos en los sanos (1 – especificidad) en el horizontal de las abscisas.

Ejemplo de R

```
# Instale en R el paquete pROC
# y genere 2 vectores con la con-
# dición y el resultado del test
> install.packages('pROC')
> library(pROC)
> respuesta <- c(0,0,1,0,1,0,0,1,1,1) # 0=S, 1=E
> test <- 1:10 # Valores de 1 a 10
> roc(respuesta,test,plot=TRUE)
```



Ejercicio 2.4 (cont Ejercicio 2.1)

Dibuje, a ojo, la curva ROC.

Recuerde

La curva ROC informa sobre el rendimiento:

1) de cada punto de corte de una prueba determinada.

2) global de cada prueba dentro de un conjunto de pruebas.

La mayor exactitud diagnóstica de una prueba se traduce en un desplazamiento hacia la esquina superior izquierda de la curva ROC: el Área Bajo la Curva ROC (ABC) indica la exactitud global de la prueba, con máximo en 1 y mínimo en 0.5.

Nota: Un valor menor de 0.5 indica clasificación cruzada (los sanos tienen más tendencia al positivo que los enfermos), por lo que debería invertirse el criterio de positividad de la prueba.

Recuerde

El ABC de ROC se interpreta como la proporción de parejas sano-enfermo en las que el enfermo tiene un valor más alto que el sano.

En términos probabilísticos, si X_E y X_S son los valores del indicador en los enfermos y los sanos, $ABC = P(X_E > X_S)$.

Nota: ABC coincide con el valor del estadístico del promedio de la suma de rangos de Wilcoxon, W, que permite contrastar la hipótesis $P(X_E > X_S) = ½$.

Recuerde

ABC es la probabilidad de que un enfermo tenga mayor valor que un sano.

Agradecimiento: La Figura 1. es de Enrique Ventura y ha sido publicada en Casino G, coord. Bioestadística para periodistas y comunicadores. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 26. Barcelona: Fundación Dr. Antonio Esteve; 2013.

Soluciones a los ejercicios

1.1. X~U(0,24);

P(X≤15) = (15-0)/(24-0) = 15/24:

P(X≤8) = (8-0)/(24-0) = 8/24;

P(8≤X≤15) = P(X≤15)–P(X≤8) = 15/24 –8/24 = 7/24

1.2. a) `> pnorm (162, 170,8)` `[1] 0.1586553` P(X<162) = P(X≤162) ≈ 15.86%

`> pnorm (154, 170,8)` `[1] 0.02275013` P(X<154) = P(X≤154) ≈ 2.28%

P(X<170) = P(X≤170) = 50% (No necesitamos R)

b) Por simetría: P(X>170) = P(X<170) =50%;

P(X>178) = P(X<162) ≈ 15.86%

P(X>186) = P(X<154) ≈ 2.28%

c) P(162 ≤ X ≤ 178) = P(X<178) – P(X<162) = [1-P(X>178)] – P(X<162) ≈

≈ [1-0.1586] – 0.1586 = 0.6828 ≈ 68.28%

P(154 ≤ X ≤ 186) = P(X<186) – P(X<154) = [1-P(X>186)] – P(X<154) ≈

≈ [1-0.0228] – 0.0228 = 0.9544 ≈ 95.44%

1.3. a) Para acotar el 99% de las observaciones, se debe calcular el cuantil 0.99 de la distribución Normal, con parámetros μ=170 y σ=8.

```
> qnorm(0.99,170,8)        [1] 188.6108
```

Por lo tanto, para garantizar que el 99% de los pacientes cabrán, las camas deben medir por lo menos 188.61 cm.

b) Se tiene que calcular *a* tal que P(X ≤ *a*) = 95%, siendo X~N(170,8).

```
> qnorm(0.95,170,8)    [1] 183.1588
```

El valor del umbral seria de 183.16 cm: el 95% de los sanos daría negativo (especificidad).

c) Al poder tener tanto valores altos como bajos se debe repartir la α del 5% entre las dos colas. Es decir que tenemos que encontrar a_1 y a_2 tal que P(X ≤ a_1) = 2.5% y P(X ≤ a_2) = 97.5%.

```
> qnorm(0.025,170,8)       [1] 154.3203
> qnorm(0.975,170,8)       [1] 185.6797
```

Por lo tanto los límites de "normalidad" estarían entre 154.32 y 185.68 cm.

Nota: Dejar 2.5% a cada lado es la más *bonita* de las posibles soluciones, pero también cumpliría con una especificidad del 95% dejar, por ejemplo, un 4% abajo y un 1% arriba.

1.4. Partiendo de que X~N(5,0.5) obtenemos:

a) `> qnorm(0.025,5,0.5)` `[1] 4.020018`

`> qnorm(0.975,5,0.5)` `[1] 5.979982`

El umbral está entre 4.02 y 5.98 Voltios en el 95% de los casos.

b) Debemos encontrar *a* tal que P(X>*a*) = 0.95, que por simetría de la distribución es lo mismo que encontrar *a* que cumpla P(X ≤ *a*) = 0.05.

```
> qnorm(0.05,5,0.5)          [1] 4.177573
```

En el 95% de los voluntarios, el umbral se sitúa por encima de 4.18 Voltios.

c) Debemos encontrar *a* tal que $P(X \leq a) = 0.95$.

```
> qnorm(0.95,5,0.5)          [1] 5.822427
```

En el 95% de los voluntarios, el umbral se sitúa por debajo de 5.82 Voltios.

d) Se trata de encontrar los cuantiles de $\alpha/2=0.05$ para la variable X. Teniendo en cuenta los resultados de los apartados anteriores el 90% de los voluntarios tienen un umbral entre 4.18 y 5.82.

e) Utilizando el mismo razonamiento que en el apartado b) obtenemos:

```
> qnorm(1-0.84,5,0.5)        [1] 4.502771
```

En el 84% de los voluntarios, el umbral se sitúa por encima de 4.50 Voltios.

f) Utilizando el mismo razonamiento que en el apartado c) obtenemos:

```
> qnorm(0.84,5,0.5)          [1] 5.497229
```

En el 84% de los voluntarios, el umbral se sitúa por debajo de 5.50 Voltios.

g) `> 1-pnorm(6.3,5,0.5)  [1] 0.004661188` $P(X>6.3) = 1-P(X\leq 6.3)$

h) $P(4.5 \leq X \leq 5.5) = P(X\leq 5.5) - P(X\leq 4.5)$

```
> pnorm(5.5,5,0.5) - pnorm(4.5,5,0.5)        [1] 0.6826895
```

La probabilidad de que un voluntario tenga un umbral entre 4.5 y 5.5 es de aproximadamente 68%.

1.5. a) Convendría estudiar cómo se han definido estos límites. Dado que (con pequeña probabilidad) puede haber personas sanas que tengan valores muy alejados, suelen definirse estos límites de forma que incluyan el 95% de los sanos. Por tanto, en principio es posible que una persona sana supere estos límites, si bien con una probabilidad pequeña, conocida y decidida previamente.

b) A partir de estas cifras, si se asume la forma de montañita simétrica de la normal, la media sería el punto central, 100, y la desviación típica, la mitad de la distancia de los extremos, 2.5.

c) Que la variable siga la distribución normal.

d) Parece difícil imaginar una distribución simétrica para la Ferritina. El cálculo anterior no sería correcto. A veces, trasformar logarítmicamente estas variables positivas permite descubrir detrás una forma de ¡montañita simétrica!

1.6. Por la experiencia previa, parece que las cifras de colesterol son relativamente simétricas, con más casos por el centro.

1.7. Por favor, consulte sus propuestas o en el foro o con su tutor o con los directores del curso.

1.8. Dada la simetría de la distribución Normal, la proporción de casos por encima de -1.96 y la proporción de casos por debajo de 1.96 es la misma. Como el valor 1.96 deja por encima el 2.5% de los casos, por debajo de 1.96 se encuentran el 97.5% de los casos —así como por encima de -1.96.

1.9. `> 1-pnorm(1.66,0,1)  [1] 0.04845723` $P(Z>1.66) = 1- P(Z\leq 1.66)$

1.10. En E porqué se aleja 2σ, mientras que en A solo se aleja 1.75σ.

1.11. `> 1-pexp(3,1)  [1] 0.04978707` $P(T\geq 3) = 1-P(T<3)$, con $T\sim E(1)$

1.12. Debemos encontrar *a* tal que P(T≤*a*) = 0.95.

```
> qexp(0.95,50)        [1] 0.05991465
```

Aproximadamente 0.06h es decir en el minuto 3.6 de partido.

1.13. 1/7 de semana, es decir 1 día.

1.14. En la primera fila, de 100 casos, los órdenes serían: visualmente quizás b > a > c ; con D, a > b > c; y con W, b > a > c. Luego el peor ajuste es c. En realidad, la muestra c ha sido obtenida de una lognormal, muy asimétrica, b de una normal y a de una uniforme. En la segunda fila, de 1000 casos, coinciden los 3 criterios: e (normal) > f (uniforme) > d (lognormal).

2.1. Declarando positivo solo si el resultado es igual a 5:

P(+|E)=33/51=0.65 P(+|S)=2/58=0.03

Declarando positivo solo si el resultado es igual a 1:

P(+|E)=48/51=0.94 P(+|S)=25/58=0.43

2.2. Ahora, en lugar de decir "La curva ROC dibuja los pares de las proporciones de positivos en las 2 muestras para cada umbral", diremos "La curva ROC representa los pares de sensibilidad y el complementario de la especificidad para cada punto de corte".

2.3. Hay que fijarse que para cinco categorías de resultados en la prueba índice tendremos cuatro puntos de corte posibles. Vamos ahora a interpretar esta tabla.

		n=58		n=51	
		Clasificación Sanos		Clasificación Enfermos	
		Correcta	Incorrecta	Correcta	Incorrecta
Punto de corte	Probablemente normal	33	25	48	3
	Dudosa	39	19	46	5
	Probablemente anormal	45	13	44	7
	Seguramente anormal	56	2	33	18

- A modo de ejemplo de interpretación, cogemos los pacientes clasificados como sanos en la primera fila: en este caso un resultado negativo de la prueba equivaldría únicamente a estar clasificado en el grupo "Seguramente normal"), así tendríamos 33 pacientes bien clasificados (dan negativo) y 25 (6+6+11+2) mal clasificados (dan positivo).
- También a modo de ejemplo de lectura de la tabla anterior observemos la segunda fila, entre los pacientes enfermos: eligiendo el punto de corte "Dudosa", tendríamos 46 (33+11+2) individuos bien clasificados (es decir que dan positivo en la prueba índice, ya que están clasificados en una de las categorías "Normales") y 5 individuos (3+2) mal clasificados (que dan negativo).
- El siguiente paso es construir una tabla con las proporciones de positivos, es decir, los valores de sensibilidad y (1-especificidad) para los distintos puntos de corte. Hay dos puntos de corte, al principio y al final que corresponden a las situaciones extremas en que todos los pacientes son o bien clasificados como positivos o, todo lo contrario, como sanos. La tabla completa que obtendríamos añadiendo también la columna de especificidad sería:

		Especificidad	1-Especificidad	Sensibilidad
Punto de corte	-	0.00	1.00	1.00
	Probablemente normal	33/58 = 0.57	0.43	48/51 = 0.94
	Dudosa	39/58 = 0.67	0.33	46/51 = 0.90
	Probablemente anormal	45/58 = 0.94	0.22	44/51 = 0.86
	Seguramente anormal	56/58 = 0.97	0.03	33/51 = 0.65
	-	1.00	0.00	0.00

La primera y la última fila corresponden a los valores extremos comentados.

2.4. Su dibujo debería parecerse a la Figura 2.4 que proporciona R con la ayuda de 2 paquetes adicionales: *epitools* (para pasar de la tabla a un data.frame) y *pROC* (para dibujar la curva)

```
> install.packages('epitools')
> library(epitools)
> install.packages('pROC')
> library(pROC)
> a <- matrix(c(33,6,6,11,2,
              3,2,2,11,33),nrow=2,byrow=TRUE,
              dimnames=list(c("Sano","Enfermo"),1:5))
> b <- expand.table(a)
> response <- b[,1]
> test <- as.numeric(b[,2])
> r <- roc(response,test,plot=TRUE)
```

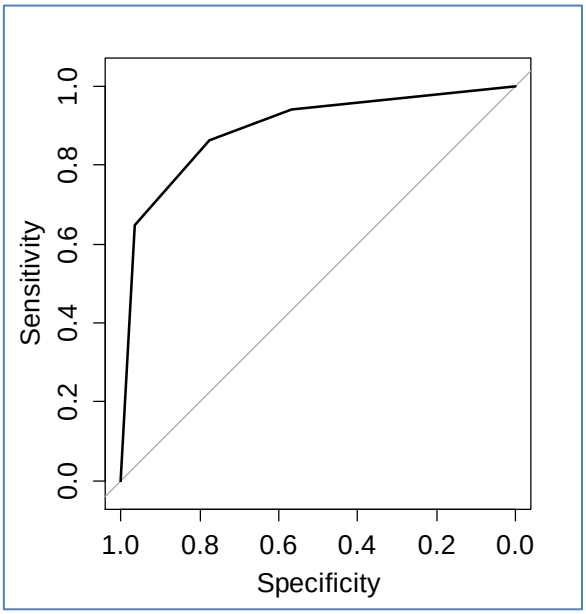


Figura 2.4 Curva ROC

Se pueden consultar las sensibilidades y las especifidades correspondientes a los puntos de la curva con `r$sensitivities` y `r$specificities`, respectivamente. El área bajo la curva ABC se obtiene con `r$auc` (Area Under the Curve). Para ver todo lo que puede obtener, haga `names(r)`