

# Capítulo 4:

## Uso de la distribución normal

### Presentación

Hemos visto que, en Ciencias de la Vida, la variabilidad es la norma. Eso implica saber aceptar que ciertas distancias con el valor central son, por definición, «normales», en el sentido de no patológicas. Y por tanto, hay que aprender a distinguir qué distancias, por su magnitud, ya no deben ser aceptadas y pueden ser sospechosas de patológicas. La distribución del Gauss-Laplace, llamada normal, es muy útil para una gran cantidad de variables. En este capítulo, con la ayuda de ejercicios de dificultad progresiva, el lector se habituará al uso de la tabla de la distribución normal y aprenderá a distinguir hasta qué punto una observación puede considerarse, o no, «normal».

### Objetivos

#### Al terminar este capítulo, un lector que haya realizado los ejercicios:

- Sabrá manejar las tablas de la distribución normal para convertir valores observados en percentiles.
- Sabrá manejar las tablas de la distribución normal para convertir percentiles en valores observados.
- Sabrá reconocer si una variable puede ser representada por la distribución normal.
- Interpretará un valor fuera de bandas como poco frecuente.
- A partir de pares de valores de sensibilidad y especificidad, dibujará la curva característica (ROC) para evaluar un indicador diagnóstico cuantitativo.

## DISTRIBUCIÓN NORMAL

La probabilidad no sólo aparece en variables con dos posibles valores como las estudiadas hasta ahora. A continuación se expone cómo el modelo *normal* de Gauss-Laplace permite representar la distribución de variables cuantitativas.

La distribución normal (fig. 4-1) tiene la conocida forma de campana o montaña, simétrica alrededor de la media ( $\mu$ ) y con la desviación típica ( $\sigma$ ) marcando la distancia que separa la media del punto de máxima pendiente o de inflexión de la curva.

Recuerde:  $\mu$  (mu) y  $\sigma$  (sigma) representan los parámetros **media** y **desviación típica**;  $\sigma^2$  representa la **varianza**.

### Nota técnica



**Interpretación física.** La media representa el centro de gravedad, es decir, aquel punto que permitirá aguantar en equilibrio, la distribución. La varianza representa la inercia, es decir, la resistencia en hacer girar la distribución alrededor de la media.

### Recuerde



*En la distribución normal, la media  $\mu$  (centro) y la desviación típica  $\sigma$  (distancia con la media del punto de máxima pendiente) tienen pleno sentido.*

Es sorprendente cómo este modelo matemático consigue reproducir con bastante exactitud la distribución empírica de un buen número de variables biológicas.

### Nota técnica



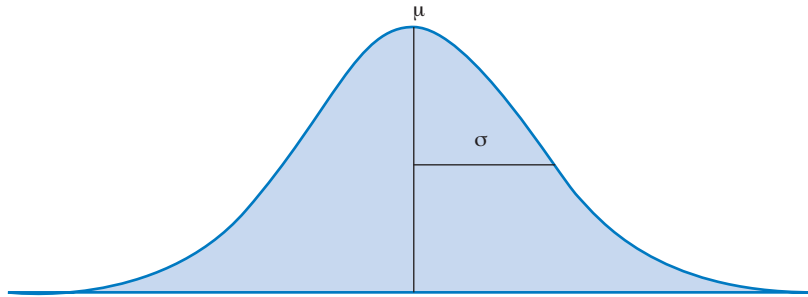
Decir que «una variable biológica sigue la distribución normal» es un abuso de lenguaje. Sería más correcto decir que, utilizando el modelo normal, se consigue reproducir de forma muy notable el comportamiento empírico de dicha variable. La distribución normal es un modelo, no una verdad absoluta.

La distribución normal asume que la variable en estudio es el resultado de la actuación de muchos fenómenos independientes y con igual influencia.

### Ejemplo 4.1



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 errores que acompañan a ciertos procedimientos de medida como la balanza de fiel.



**Figura 4-1** Representación de la distribución normal con media  $\mu$  y desviación estándar  $\sigma$ .

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 que representa los efectos más «sólidos», de mayor envergadura.

#### Ejemplo 4.2



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 primero especificar la edad, el género y la población, ya que éstas son variables que podrían originar diferencias notables, remarcables. Si, por ejemplo, se mezclan ambos géneros, la distribución resultante tendrá dos montañitas o jorobas que definen los intervalos modales de hombres y mujeres.

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». Su media, en cambio, representa cierta tendencia que puede ser el resultado de otros fenómenos de mayor envergadura.

#### Recuerde



La media  $\mu$  de la distribución normal representa la señal «relevante»; y la desviación típica, el ruido «irreproducible».

**Ejercicio de Navegación**

Entre en la página que se indica a continuación y observe, con la ayuda de la simulación que realiza la aplicación («La máquina de Galton»), que la distribución resultante de dejar caer unas bolas sobre clavos separadores (que las van distribuyendo al azar) es la distribución normal.

<http://www.rand.org/methodology/stat/applets/clt.html>

**Recuerde**

La notación  $N(\mu, \sigma)$  indica que una variable sigue la distribución normal con media  $\mu$  y desviación típica  $\sigma$ .

**Ejemplo 4.3**

La altura de los varones adultos sanos es  $N(170 \text{ cm}, 8 \text{ cm})$ .

**Uso de las tablas de la distribución normal**

La utilidad de la distribución normal reside en que permite cuantificar la proporción de observaciones que se encuentran a cierta distancia de la media (fig. 4-2). Por ejemplo, si se toma una vez hacia arriba y una vez hacia abajo el valor de la desviación típica, se incluye 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, se incluye el 95% de las observaciones.

Por ello, es muy útil para construir intervalos en los que cabe esperar que se encuentre un determinado porcentaje de las unidades.

Así, la distribución normal permite establecer una correspondencia entre los valores de una variable y el porcentaje de unidades comprendidas entre estos dos valores. Lo que permite dos usos recíprocos:

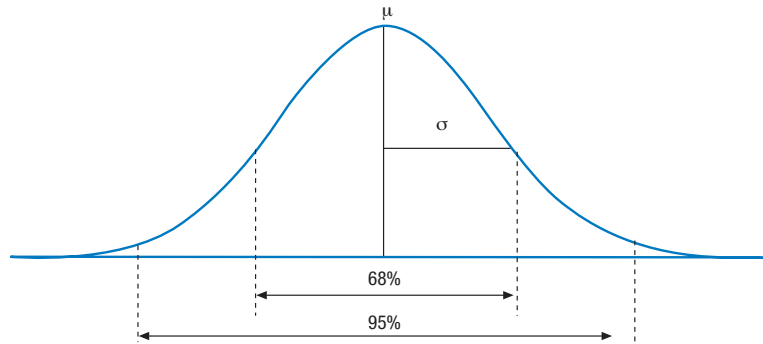
1) Dada una probabilidad, buscar un valor: **ir de los porcentajes a los valores**; cierto porcentaje (p. ej., 95%) viene delimitado por ¿qué valores de la variable? (p. ej., 150 y 170 cm)

2) Dado un valor, buscar una probabilidad: **ir de los valores a los porcentajes**; ciertos valores de la variable (p. ej., 150 y 170 cm) ¿qué porcentaje de unidades comprenden? (p. ej., 95%)

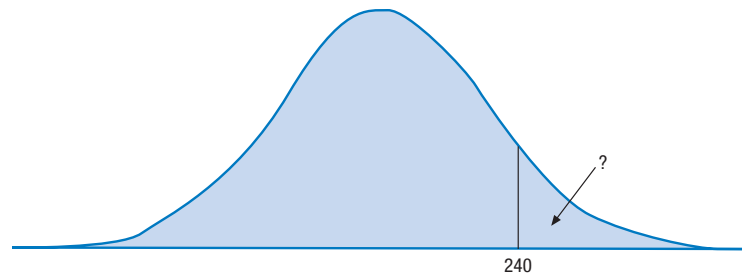
**Ejemplo 4.4**

De la utilidad 1 (fig. 4-3): Podríamos desear conocer la proporción de MIR que sacan más de 240 puntos.

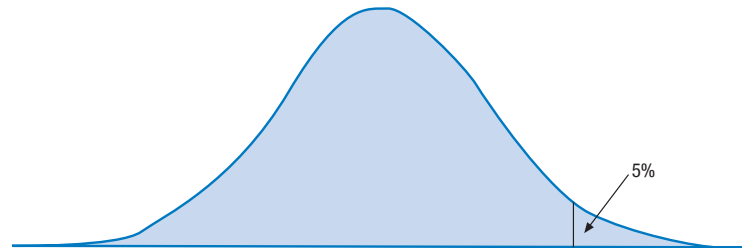
De la utilidad 2 (fig. 4-4): ¿Cuál es límite de la glucemia que deja por encima el 5% de los individuos sanos?



**Figura 4-2** Representación de la distribución normal con media  $\mu$  y desviación estándar  $\sigma$ , con las regiones que contienen el 68 y el 95% de las observaciones.



**Figura 4-3** ¿Cuál es la proporción de casos que obtiene más de 240 puntos?



**Figura 4-4** ¿Qué valor deja por encima el 5% de la distribución?

### Ejemplo 4.5



De la utilidad 1: Un paciente tiene en cierta prueba, índice o escala de medida 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 casos tienen puntuaciones inferiores.

De la utilidad 2: Conocidos los valores de la distribución de los individuos sanos de un cierto indicador bioquímico, se pueden calcular los valores de referencia que delimitan el 95% de los individuos sanos.

Para resolver estos ejemplos, los estadísticos matemáticos han realizado los cálculos necesarios y los han puesto en una tabla. Dado que diferentes valores de la media y de la desviación típica resultan en diferentes valores de los intervalos y de los porcentajes, deberían hacerse tantas tablas como posibles combinaciones de valores de la media y de la desviación típica. Para poder usar una única tabla, se puede recurrir al desvío tipificado descrito en el apartado «Descripción de los participantes en un estudio» del capítulo 2.

$$z = \text{desvío tipificado} = \frac{\text{Valor} - \text{Media}}{\text{desviación típica}} = \frac{x - \mu}{\sigma}$$

Recuérdese que esta nueva variable tiene media 0 (valores negativos representarán observaciones por debajo de la media) y desviación típica 1 (una observación prototípica se aleja de la media, por arriba o por debajo, en una unidad). Asimismo, se necesita relacionar los valores de esta variable con los porcentajes. Esto es lo que hace la tabla: proporciona el valor por el que debe ser multiplicada la desviación típica para obtener el porcentaje deseado.

Un dilema es qué porcentajes poner en la tabla: ¿los centrales que quedan dentro?, ¿los que quedan fuera?, ¿por debajo? o ¿por encima? En función del uso que se hará de las tablas convendrá poner unos u otros. En la tabla 4-1,  $\alpha$  representa el porcentaje de casos que quedan fuera.

$\alpha$	0,001	0,01	0,05	0,10	0,20	0,32
$\alpha/2$	0,0005	0,005	0,025	0,05	0,10	0,16
<b>Z</b>	3,29	2,58	1,96	1,64	1,28	1

**Tabla 4-1** Valores seleccionados de la distribución normal tipificada (Z)

#### Ejercicio 4.1



En la distribución normal tipificada, Z, ¿qué proporción de casos quedan por encima de  $-1,96$  y por debajo de  $+1,96$ ?

#### Ejemplo 4.6



Por debajo de  $-1,96$  y por encima de  $+1,96$  queda un 5% de unidades.

#### Recuerde



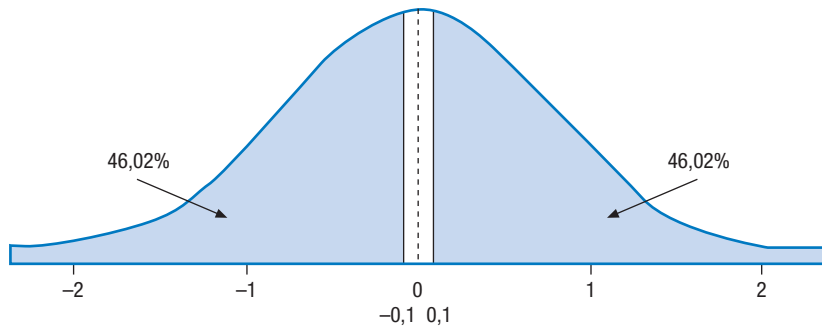
En la normal tipificada, Z, «1,96» (o redondeado: «2») es el valor que contiene el 95% de las observaciones.

### Ejercicio 4.2



- ¿Entre qué valores de la distribución normal tipificada se encuentran el 99% de las observaciones?
- ¿Qué valores contienen el 90%?
- ¿Qué valor deja por encima el 5%? ¿Y por debajo?

La tabla 4-2 es más completa que la tabla 4-1 y sirve para encontrar más valores. Para poderlos contener, necesita abarcar más de una fila. El primer valor de la tabla, 0,5000, se corresponde con la fila 0,0\_ y la columna \_,0, indicando que, por encima de  $z = 0,0_ + _,0 = 0,00$  se encuentra el 50% de la distribución. En la última columna de la primera fila se comprueba que, por encima de  $z = 0,0_ + _,9 = 0,09$  se encuentra el 46,41% de los casos. El valor siguiente a  $z = 0,09$  que muestra la tabla 4-2 es 0,10, en la primera columna de la segunda fila  $z = 0,1_ + _,0 = 0,10$  que deja por encima el 46,02% de los casos (fig. 4-5).



**Figura 4-5** En la tabla 4-2 puede verse, en la 1.ª columna y 2.ª fila, que por encima de  $Z = 0,10$  queda un 46,02% de los casos.

### Comentario

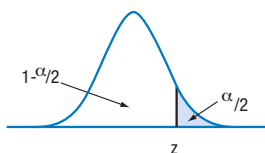


*Dado que para hacer corresponder este valor con su porcentaje se necesitarían dos filas larguísimas, por cuestiones de edición, al llegar al 0 se parten las filas, se ponen por debajo y así queda en forma de tabla, más fácil de imprimir en un libro.*

### Ejemplo 4.7



¿Qué proporción de casos están por encima de  $z = 1,66$ ? Es decir, ¿cuál es la probabilidad de que  $Z > 1,66$ ? Se descompone el número en  $1,6 + 0,06$ , y se busca en la celda que une la fila del 1,6 y la columna del 0,06: el resultado es 0,0485.

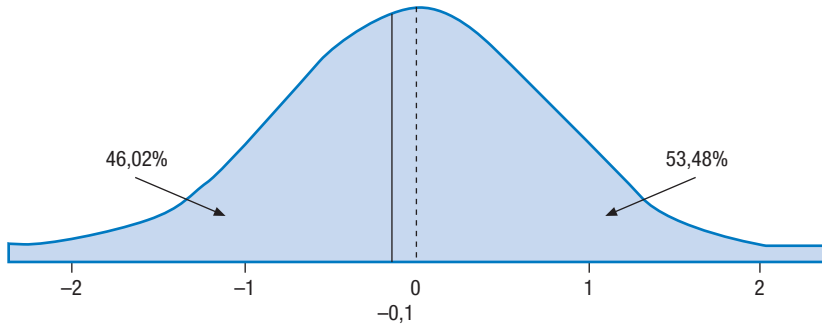


z	._0	._1	._2	._3	._4	._5	._6	._7	._8	._9
0,0_	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1_	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2_	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3_	0,3821	0,3783	0,3745	0,3707	0,3669	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
3,9_	4,8E-05	4,6E-05	4,4E-05	4,2E-05	4,1E-05	3,9E-05	3,7E-05	3,6E-05	3,4E-05	3,3E-05
4,0_	3,2E-05	3,0E-05	2,9E-05	2,8E-05	2,7E-05	2,6E-05	2,5E-05	2,4E-05	2,3E-05	2,2E-05
4,5_	3,4E-06	3,2E-06	3,1E-06	3,0E-06	2,8E-06	2,7E-06	2,6E-06	2,4E-06	2,3E-06	2,2E-06
5,0_	2,9E-07	2,7E-07	2,6E-07	2,5E-07	2,3E-07	2,2E-07	2,1E-07	2,0E-07	1,9E-07	1,8E-07
5,5_	1,9E-08	1,8E-08	1,7E-08	1,6E-08	1,5E-08	1,4E-08	1,4E-08	1,3E-08	1,2E-08	1,1E-08
6,0_	9,9E-10	9,3E-10	8,8E-10	8,2E-10	7,7E-10	7,3E-10	6,8E-10	6,4E-10	6,0E-10	5,7E-10

Tabla 4-2 Distribución NORMAL estandarizada. Áreas de cola hacia la derecha

En la fila 1,9\_ y columna \_\_,6 puede leerse que, por encima de  $Z = 1,9_ + __,6 = 1,96$  queda el 2,5% de la distribución.

Dada la simetría de la distribución normal, la tabla 4-2 también proporciona los límites por la izquierda. Así, por debajo de  $-0,10$  también hay el 46,02% de los casos.



**Figura 4-6** Por debajo de 0,1 hay el 46,02% de las observaciones.

#### Ejemplo 4.7 (Cont.)



¿Qué proporción de casos están por debajo de  $-1,38$ ? Es decir, ¿cuál es la probabilidad de que  $Z$  sea menor que  $-1,38$ ? Según la tabla 4-2, a  $1,38$  (fila 1,3\_ y columna \_\_,8) le corresponde 0,0838.

Asimismo, pueden obtenerse los valores complementarios restando el observado de 100 (fig. 4-6).

#### Ejemplo 4.8



Así, por encima de  $-0,1$ , se encuentra el 53,48% de los casos.

$$100 - 46,02 = 53,48$$

#### Ejercicio 4.3



Compruebe que sabe reproducir con la tabla 4-2 los resultados de los ejercicios 4.1 y 4.2, que se obtuvieron con la tabla 4-1.

#### Nota técnica



Los textos de estadística suelen expresar la frase «por encima de 0,1 se encuentra el 46,02% de la distribución normal tipificada  $Z$ » de manera más compacta, como, por ejemplo:  $P(Z > 0,1) = 0,4602$ .

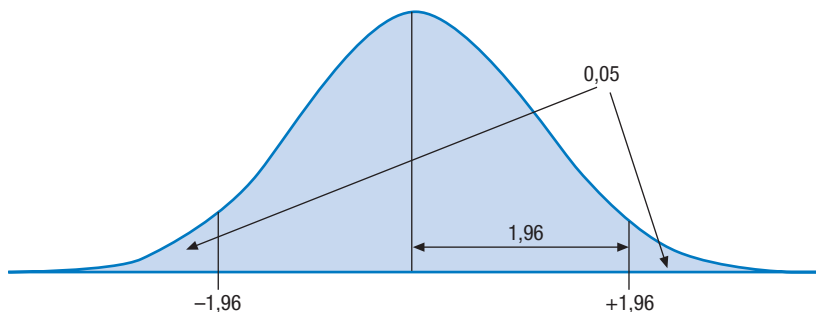
**Ejemplo 4.9**

¿Qué valor  $Z$  de la distribución normal deja por encima el 29% de los casos? En la tabla 4-2, las celdas que más se parecen a este valor son 0,2912 y 0,2877. Si la diferencia entre 29 y 29,12% puede despreciarse, la respuesta es que por encima de  $Z = 0,55$  se encuentra el 29% de los casos.

**Recuerde**

*Siempre que utilice unas nuevas tablas observe detalladamente el gráfico y el ejemplo para concretar qué valor se representa en la tabla.*

Un «truco» que funciona bastante bien es recordar el valor «mágico» que define el 95% central de las observaciones, que es el 1,96 y buscarlo en las tablas para comprobar que se están interpretando correctamente (fig. 4-7). En la tabla puede verse que el valor 1,96 se encuentra en la fila y columna encabezadas por 0,00 y 0,05, la suma de ambos valores es el valor representado por  $\alpha$  (proporción de casos que queda por debajo de  $-1,96$  y por encima de  $+1,96$ ).



**Figura 4-7**  $-1,96$  y  $+1,96$  dejan por fuera el  $0,05 = 5\%$  de los casos.

**Ejemplo 4.10**

De la utilidad 1: dado un valor, buscar la probabilidad que delimita: la puntuación del examen MIR sigue una DN de media  $\mu = 200$  puntos y desviación típica  $\sigma = 20$  puntos ¿Qué proporción de casos se sitúan por encima de 240 puntos?

$$Z = \frac{X - \mu}{\sigma} = \frac{240 - 200}{20} = 2$$

Según las tablas, la proporción de alumnos que sacan más de 240 puntos es del 2,28%.

### Ejemplo 4.11



De la utilidad 2, dada una probabilidad, buscar el valor que la delimita: La HCM (hemoglobina corpuscular media) en sangre sigue una DN de media  $\mu = 30$  y desviación típica  $\sigma = 2$ . ¿Qué límites de normalidad que incluyan el 95% de los individuos sanos se pueden proponer? En tablas, se vuelve a encontrar el valor  $Z = 1,96$ , pero antes de usarla hay que darle la vuelta a la fórmula anterior

$$Z = \frac{X - \mu}{\sigma} : \text{se convierte, como } z \text{ puede ser positiva o negativa, en:}$$

$$X = \mu + Z\sigma \quad \text{y} \quad X = \mu - Z\sigma$$

$$\text{por lo que: } X = \mu + Z\sigma = 30 + 1,96 \cdot 2 = 33,92 \approx 34$$

$$X = \mu - Z\sigma = 30 - 1,96 \cdot 2 = 26,08 \approx 26$$

y los límites serán 26 y 34.

### Ejercicio 4.4



Cierto estimulador tiene un umbral que varía de un voluntario sano a otro. Su distribución es aproximadamente normal con una media de 5 voltios y una desviación típica de 0,5.

- El 95% de los voluntarios tienen un umbral que se sitúa entre \_\_\_ y \_\_\_ voltios.
- En el 95% de los voluntarios, el umbral se sitúa por encima de \_\_\_ voltios.
- En el 95% de los voluntarios, el umbral se sitúa por debajo de \_\_\_ voltios.
- El 90% de los voluntarios tienen un umbral que se sitúa entre \_ y \_ voltios.
- En el 84% de los voluntarios, el umbral se sitúa por encima de \_\_\_ voltios.
- En el 84% de los voluntarios, el umbral se sitúa por debajo de \_\_\_ voltios.
- ¿Cuál es la probabilidad de que el umbral supere 6,3 voltios?
- ¿Cuál es la probabilidad de que un voluntario tenga un umbral entre 4,5 y 5,5?

## Uso de la distribución normal con aplicaciones informáticas

Por cuestiones de espacio, la tabla 4-2 sólo incluye algunos valores seleccionados, lo que puede ocasionar pequeños errores de aproximación. Muchas aplicaciones informáticas permiten obtener muchos más valores. Una hoja de cálculo, Excel por ejemplo, permite obtener directamente la proporción de casos por debajo de un cierto valor. Por ejemplo, la función:

= DISTR. NORM. ESTAND. (z)

devuelve la proporción de casos por debajo de z en el caso de una distribución normal tipificada. Así, si introducimos  $Z = 1,96$ , dará 0,9750021.

A su vez, la función:

= DISTR. NORM. ESTAND. INV. (probabilidad)

hace la función inversa: devuelve el valor de Z a partir de la probabilidad. Si se le introduce probabilidad = 0,975, proporciona  $Z = 1,95996398$ .

La función:

= DISTR. NORM. (x; media; desviación estándar; 1)

devuelve la proporción de casos por debajo de X en el caso de una distribución normal con la media y la desviación típica especificadas. Así, si introducimos  $X = 1,96$ , media = 0 y desviación estándar = 1 proporciona el valor anterior de 0,9750021.

Finalmente:

= DISTR. NORM. INV. (probabilidad; media; desviación estándar)

devuelve la inversa de la función anterior. Si se introduce probabilidad = 0,975, media = 0; desviación estándar = 1, proporciona  $x = 1,95996398$ .

### Ejercicio 4.5



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

- ¿Es posible que una persona sana supere estos límites?
- ¿Cuál cree usted que es el valor de la media y de la desviación típica de esta variable en los «normales»?
- ¿Existe alguna condición (premisa) para este cálculo?
- Para la ferritina, estos límites son 15-200 mg/l ¿Cómo se imagina su distribución?

### Ejercicio 4.6\*

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

### Ejercicio 4.7\*

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.

\*No incluye solución al final del capítulo.

## Aplicaciones al diagnóstico

Hasta este momento se ha hecho hincapié en pruebas diagnósticas llamadas cualitativas, es decir, aquellas que sólo admiten dos posibles resultados: positivo y negativo. En este caso, la definición de sensibilidad y especificidad, así como sus complementarias, es inmediata y unívoca.

En los test donde hay varios resultados numéricos posibles, llamados cuantitativos, la definición de los términos anteriores no es inmediata sino convencional, ya que dicha definición requiere establecer un límite o umbral que separe el conjunto de resultados en dos grupos, positivo y negativo.

### Ejemplo 4.12



Por ejemplo, los resultados de un test que mide la concentración de glucosa en plasma, en condiciones basales. Dichos resultados, expresados en mg/dl, pueden ser muy variados: 50, 75, 110, 128, 165, 192, etc. Ninguna cifra de éstas es, por sí misma, positiva ni negativa. Ahora bien, si en virtud de conocimientos fisiológicos y epidemiológicos, entre otros, se considera que las cifras inferiores a 100 definen un resultado negativo, y las superiores, positivo, entonces la situación se ha hecho similar a los test cualitativos.

### Recuerde



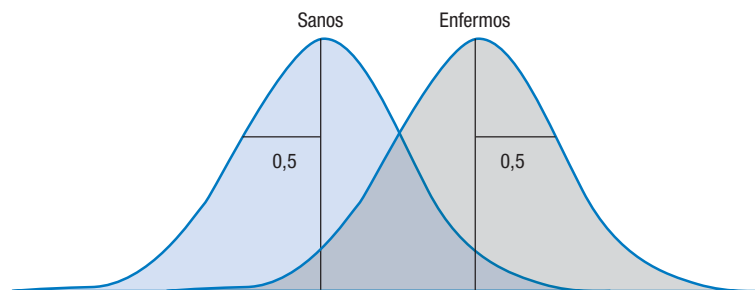
*Si el resultado del indicador diagnóstico es un número, es necesario establecer un punto de corte.*

## Las curvas ROC (Receiver Operating Characteristic curves)

### Ejemplo 4.13



El ejercicio 4.4 dice que el umbral de estimulación de los voluntarios sanos sigue una  $N(5, 0,5)$ . Supongamos, además, que en cierto tipo de enfermos sigue una  $N(6, 0,5)$  (fig. 4-8).



**Figura 4-8** Distribución del umbral de estimulación en sanos y enfermos.

**Ejemplo 4.13 (Cont.)**

Si el criterio diagnóstico se establece en 5,5 (fig. 4-9), los valores de sensibilidad y especificidad serán:

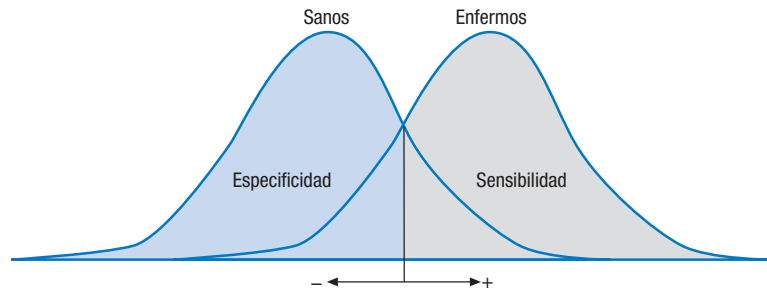
$$\begin{aligned}\text{sens} &= P(+|E) = P(y > 5,5 \mid \text{Enfermo}) = P\left(z > \frac{5,5 - 6}{0,5}\right) \\ &= P(z > -1) \approx 84,13\%\end{aligned}$$

$$\begin{aligned}\text{esp} &= P(-|S) = P(y < 5,5 \mid \text{Sano}) = P\left(z < \frac{5,5 - 5}{0,5}\right) \\ &= P(z < 1) \approx 84,13\%\end{aligned}$$

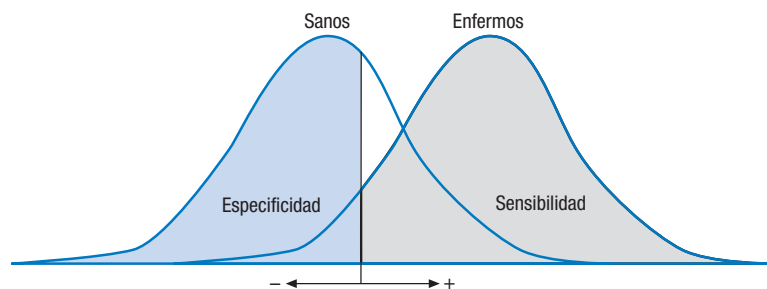
En cambio, si el criterio se hubiera establecido en 5,2 (fig. 4-10), serían:

$$\begin{aligned}\text{sens} &= P(+|E) = P(y > 5,2 \mid \text{Enfermo}) = P\left(z > \frac{5,2 - 6}{0,5}\right) \\ &= P(z > -1,6) \approx 94,52\%\end{aligned}$$

$$\begin{aligned}\text{esp} &= P(-|S) = P(y < 5,2 \mid \text{Sano}) = P\left(z < \frac{5,2 - 5}{0,5}\right) \\ &= P(z < 0,4) \approx 65,54\%\end{aligned}$$



**Figura 4-9** 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.



**Figura 4-10** Al desplazar el criterio diagnóstico 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».

#### Ejemplo 4.14



Everitt (29). «Examine los siguientes valores de 1 (claramente sano) a 5 (claramente enfermo) para 50 sujetos sanos y 50 enfermos [tabla 4-3]. Si el valor 5 se usa como punto de corte para identificar casos de enfermedad, entonces la sensibilidad se calcula como  $8/50 = 0,16$ ; y la especificidad, como  $49/50 = 0,98$ . En cambio, usar el valor 4 como punto de corte supone una sensibilidad de  $27/50 = 0,54$  y una especificidad de  $41/50 = 0,82$ .»

Valor	1	2	3	4	5	TOTAL
Sanos	4	17	20	8	1	50
Enfermos	3	3	17	19	8	50

**Tabla 4-3** Número de casos, sanos y enfermos, con cada valor de la escala

#### Ejercicio 4.8



Calcule la sensibilidad y la especificidad si el punto de corte fuera 3, 2 o 1.

#### Definición



La curva característica (ROC: Receiver Operating Characteristic Curve) contiene todos los posibles pares de sensibilidad y especificidad de un indicador diagnóstico.

#### Nota técnica



El término «*receiver operating characteristic*» proviene de las telecomunicaciones y analiza la capacidad de un receptor de señales para discriminarlas correctamente. Una posible traducción sería curva característica de la operatividad del receptor.

## Lectura



Everitt (29). «Curvas características (curvas ROC): grafico de la sensibilidad de un test de diagnóstico frente al complementario de la especificidad según varía el punto de corte que indica que un test es positivo. A menudo se usa para elegir entre varios test en competencia, aunque el procedimiento no tiene en cuenta la prevalencia de la enfermedad que se estudia.»

## Nota técnica



El área bajo la curva ROC se interpreta como la probabilidad de, seleccionados al azar 1 sano y 1 enfermo, que el primero tenga valores menos patológicos que el segundo.

## Ejemplo 4.15



El doctor Manuel Callis, en su tesis doctoral, predice el grado de afectación abdominal en enfermos de Hodking sin necesidad de recurrir a la cirugía (laparoscopia). Para ello, ha obtenido la siguiente regresión logística:

$$\log(p/(1-p)) = 0,85 + 0,04 \cdot X_1 - 13,9 \cdot X_2 + 1,14 \cdot X_3 + 1,93 \cdot X_4 + 1,53 \cdot X_5$$

en la que las diferentes X indican variables que pueden ser obtenidas en una visita clínica habitual (análisis, radiografías, etc.). La tabla 4-4 indica los valores de sensibilidad y especificidad que se obtienen escogiendo diferentes puntos de corte mediante esta regresión logística (p. ej., escogiendo como punto de corte una predicción de afectación abdominal de 0,10, se obtiene una sensibilidad de 0,95 y una especificidad de 0,22).

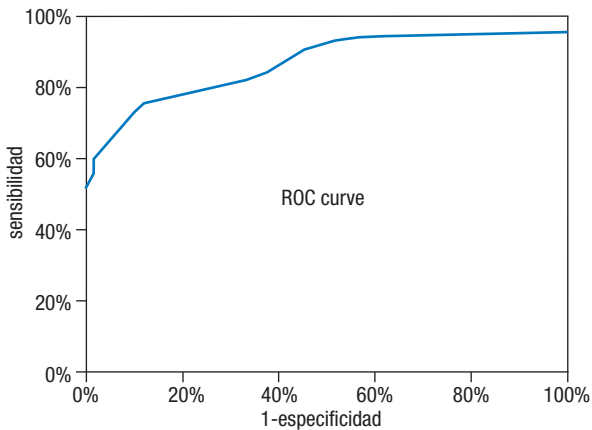
Utilizando Excel, se genera el gráfico de la curva ROC (fig. 4-11) con las instrucciones del cuadro 4-1.

Punto de corte	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
Sensibilidad	0,95	0,93	0,90	0,83	0,76	0,73	0,59	0,59	0,49
Especificidad	0,22	0,49	0,65	0,86	0,86	0,90	0,98	0,98	1,00

**Tabla 4-4** Pares de sensibilidad y especificidad del ejemplo 4-15

En la barra de herramientas seleccionar ->Insertar->Gráfico  
 Seleccionar la opción XY (Dispersión)  
 Elegir el gráfico  
 Dispersión con líneas suavizadas y sin marcadores de datos  
 Seleccionar la opción para introducir una serie  
 En *Agregar* introducimos los datos correspondientes al gráfico:  
*Nombre:* nombre del gráfico, ROC curve  
*Valores eje X:* Seleccionar el rango de valores (sensibilidad) de la hoja de Excel  
*Valores eje Y:* Seleccionar el rango de valores (especificidad) de la hoja de Excel  
 Seguir avanzando con *Siguiente* hasta finalizar la construcción del gráfico.

**Cuadro 4-1** Instrucciones de Excel para generar una curva ROC.



Especificidad	Sensibilidad
22%	95%
49%	93%
65%	83%
86%	76%
90%	73%
98%	59%
100%	49%

**Figura 4-11** Curva ROC.

#### Nota técnica



Este clínico le pide a usted asesoría para escoger el punto de corte más adecuado para decidir si existe o no existe afectación abdominal, ¿qué nuevo análisis le recomendaría Vd. para responder a esta pregunta? Debería realizarse un análisis de decisión y considerar la matriz de beneficios/pérdidas (¿qué «cuesta» un enfermo al que se declara sano?, ¿y un sano al que se le declara enfermo? ¿Qué «vale» un sano al que se le declara como tal? ¿Y un enfermo?). El gran reto de la economía de la salud es precisamente convertir en una misma «moneda» las consecuencias de no detectar a un enfermo y de tratar a un sano. Una vez establecidos estos «valores» para una cierta población, es fácil, usando sensibilidad y especificidad, decidir la estrategia «óptima».

## Soluciones a los ejercicios

**4.1** 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. Tal y como puede verse en la tabla 4-1, el valor  $1,96$  deja fuera el 5% de los casos, así que, por encima de  $-1,96$  se encuentran el 97,5% de los casos, así como por debajo de  $1,96$ .

**4.2 a)** Para acotar el 99% de las observaciones, debemos mirar la fila  $\alpha/2$  de la tabla 4.1, ya que en este caso repartimos la probabilidad del 1% entre las dos colas dada la simetría de la distribución. Marcando el límite en  $0,005 = 0,5\%$ , se obtiene que el 99% de las observaciones se encuentran entre los valores  $-2,58$  y  $2,58$ .

**b)** Análogamente, marcando el límite en  $0,05 = 5\%$ , se obtiene que el 90% de las observaciones se encuentran entre los valores  $-1,64$  y  $1,64$ .

**c)** En esta ocasión, debemos fijarnos en la fila  $\alpha/2$  de la tabla 4-1 para descubrir qué valor deja por encima el 5% y qué valor deja por debajo el 5%; en el segundo caso, la tabla 4-1 nos muestra que el valor es  $1,64$ , así que para el primer caso será  $-1,64$ .

**4.3** La fila que empieza en «1,9\_» se une a la columna encabezada por «\_,\_6» en el valor  $0,025$ , lo que significa que por encima de  $1,96$  hay el 2,5% de la distribución. La fila que empieza en «2,5\_» y la columna «\_,\_8» se unen en  $0,0049$ , aproximadamente 0,5%.

La fila que empieza en «1,6\_» y la columna «\_,\_4» proporcionan  $0,0505$ , aproximadamente 5%.

**4.4 a)** En las tablas de la distribución normal se encuentra que el valor  $1,96$  deja fuera el 5% y delimita el 95% de los casos. Por tanto, se debe multiplicar la desviación típica por este número y el resultado sumarlo y restarlo de la media:

Valores = media  $\pm 1,96$  desviación típica =  $5 \pm 1,96 \cdot 0,5 \approx 5 \pm 1 = [4, 6]$

**b)** Ahora se trata de dejar en un extremo el 5%, lo que equivale a dejar en dos extremos iguales al 10% y a contener el 90%. El valor encontrado en las tablas es  $1,645$ .

Valor = media  $- 1,64 \cdot$  desviación típica =  $5 - 1,64 \cdot 0,5 = 5 - 0,82 = 4,18$

**c)** Valor = media  $+ 1,64 \cdot$  desviación típica =  $5 + 1,64 \cdot 0,5 = 5 + 0,82 = 5,82$

**d)** Valores = media  $\pm 1,64 \cdot$  desviación típica =  $5 \pm 1,64 \cdot 0,5 = 5 \pm 0,82 = [4,18, 5,82]$

**e)** Ahora se trata de dejar en un extremo el 16%, lo que equivale a dejar en dos extremos iguales al 32% y a contener el 68%. El valor encontrado en las tablas es  $0,99$ .

Valor = media  $- 0,99$  desviación típica =  $5 - 0,99 \cdot 0,5 \approx 5 - 0,5 = 4,5$

**f)** y, por simetría,

Valor = media  $+ 0,99$  desviación típica =  $5 + 0,99 \cdot 0,5 \approx 5 + 0,5 = 5,5$

**g)** Ahora el problema se resuelve en el orden inverso; primero se tipifica la variable:

$$z = \text{desvío tipificado} = \frac{\text{valor} - \text{media}}{\text{desviación típica}} = \frac{6,3 - 5}{0,5} = 2,6$$

En la tabla 4-2, el valor  $2,6$  deja por encima  $0,0047 \approx 0,005 = 0,5\%$ . Por tanto, la probabilidad de que un caso supere la cifra  $6,3$  es menor del 0,5%.

**h)** Calculemos primero las probabilidades de no alcanzar  $5,5$  y  $4,5$  voltios y luego las restaremos entre ellas.

$$P(X > 5,5) = P\left(Z > \frac{5,5 - 5}{0,5}\right) = P(K > 1) = 0,1587$$

$$P(X < 5,5) = 1 - P(X > 5,5) = 1 - 0,1587 = 0,8413$$

$$P(X < 4,5) = P(X > 5,5) = 0,1587 \text{ (por simetría)}$$

$$P(4,5 < X < 5,5) = P(X < 5,5) - P(X < 4,5) = 0,8413 - 0,1587 = 0,6826$$

Por lo tanto, tenemos una probabilidad del 68% de que el umbral se sitúe entre los 4,5 y los 5,5 voltios.

**4.5 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, es posible que una persona sana supere estos límites.

**b)** A partir de estas cifras, si se asume la forma de montaña 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, transformar logarítmicamente estas variables positivas permite descubrir detrás una forma de ¡montaña simétrica!

**4.8** Si se define como positivo con valores iguales o superiores a 3, habrá  $17 + 19 + 8 = 44$  enfermos que dan positivo y  $1 + 8 + 20 = 29$  sanos. Por tanto, la sensibilidad será de  $44/50 = 88\%$  y la especificidad será  $(50 - 29)/50 = 42\%$ .

Para el valor 2 serán  $\text{sens} = 47/50 = 94\%$  y  $\text{esp} = 4/50 = 8\%$ .

Para el valor 1, dado que todos los casos se declaran como positivos, la sensibilidad será del 100% y la especificidad del 0%.