Distribuciones de probabilidad continuas

Responden al tipo continuo tanto variables típicamente continuas (estaturas, longitudes de piezas fabricadas, temperaturas, etc.) como variables discretas en las que las posibilidades de tomar valores son muy amplias (edades, sueldos de trabajadores, calificaciones, etc.).

Hablando en rigor, este segundo tipo de variable continua es el único que se presenta en la práctica. Si tomamos la estatura, el peso o la edad de una persona, aunque los posibles resultados son infinitos, los instrumentos de medida sólo nos van a dar valores discretos. Tendrá 1'72 ó 1'73 cm. de estatura, y esto sólo porque el instrumento con el que medimos no puede precisar más. Existe una diferencia esencial con el caso discreto.

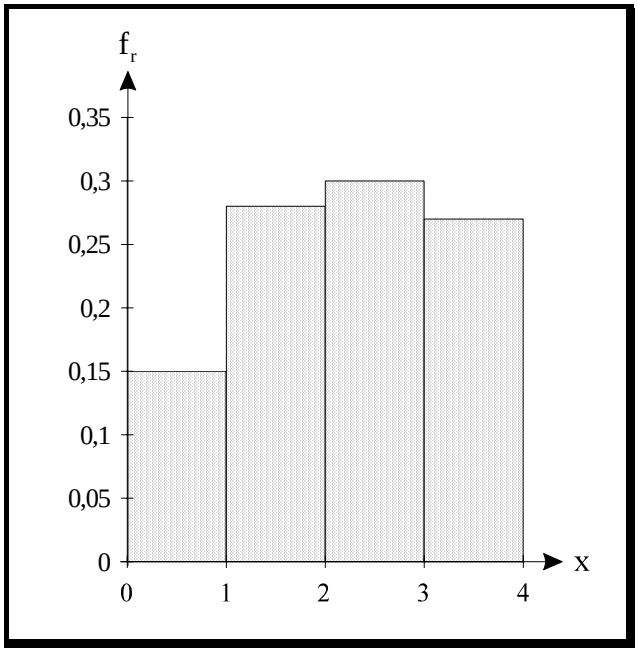
Consideremos el conjunto de números comprendidos entre 0 y 1 y tomemos uno al azar de entre ellos (selección que puede realizar el más sencillo ordenador o algunas calculadoras científicas). ¿Cuál es la probabilidad de que el número elegido sea, por ejemplo, 0'7? Siguiendo la definición de Laplace, los casos favorables se reducen a uno solo (el número 0'7) y los posibles, los infinitos números reales de ese intervalo y, por tanto, habrá que concluir que $p(X = 0'7) = 0$, pese a que elegir el número 0'7, no es, obviamente, un suceso imposible.

Puesto que una función de probabilidad de una variable aleatoria discreta asigna a cada valor la probabilidad de que éste ocurra, lo que acabamos de ver es que este concepto no tiene sentido en el caso continuo, dado que habría que asignar a cada valor de la variable la probabilidad nula.

Consideremos la distribución de la variable X continua, que se muestra en la tabla siguiente, así como el histograma correspondiente.

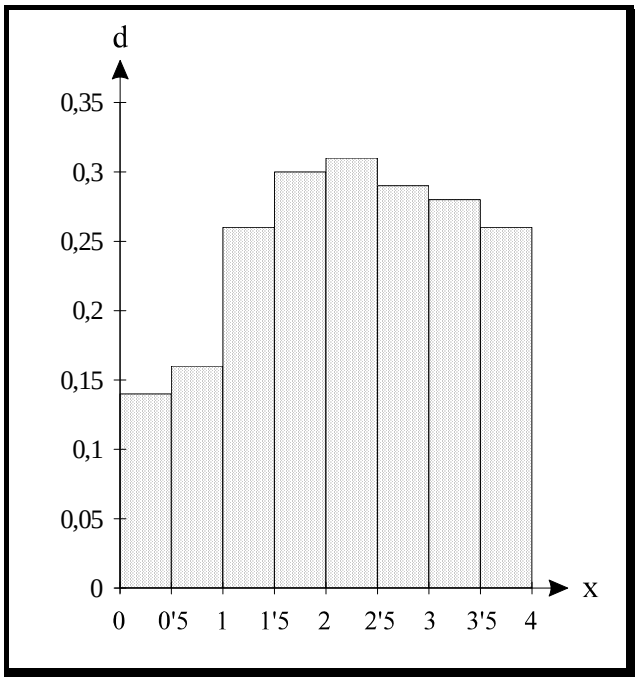
X	f_r
$[0,1)$	0'15
$[1,2)$	0'28
$[2,3)$	0'30
$[3,4)$	0'27
Total	1

Conviene recordar que en un histograma es el área de cada rectángulo levantado sobre el intervalo correspondiente, lo que nos indica las frecuencias relativas f_r y no la altura del mismo, que corresponde a la densidad de frecuencia. Así, la suma de las áreas ha de ser la unidad.



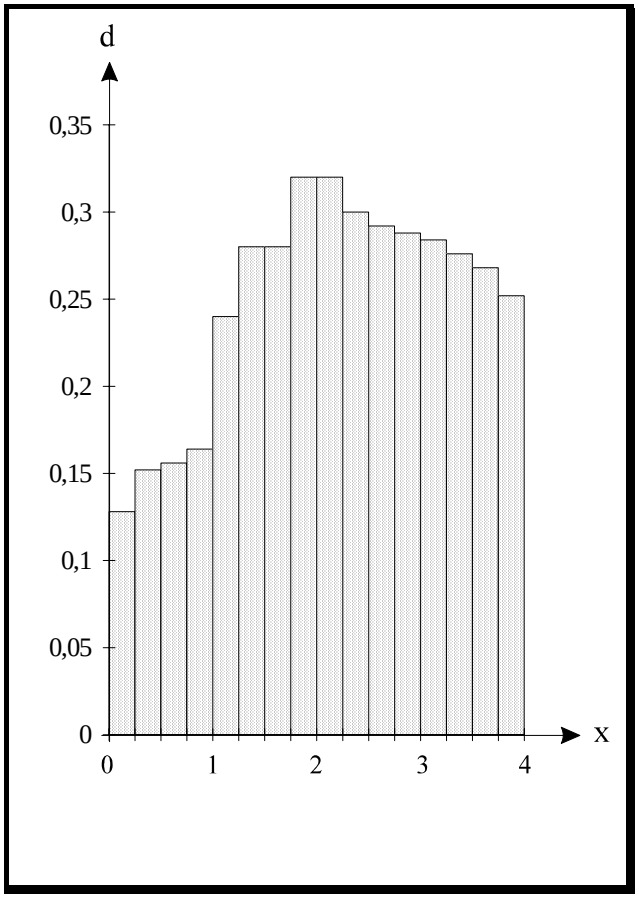
Si en la distribución presentada se reduce a la mitad la amplitud de los intervalos de clase y se aumenta el tamaño de la muestra a fin de considerar frecuencias significativas, se obtiene lo siguiente:

X	f_r	densidad
$[0,0'5)$	0'07	0'14
$[0'5,1)$	0'08	0'16
$[1,1'5)$	0'13	0'26
$[1'5,2)$	0'15	0'30
$[2,2'5)$	0'155	0'31
$[2'5,3)$	0'145	0'29
$[3,3'5)$	0'14	0'28
$[3'5,4)$	0'13	0'26
Total	1	

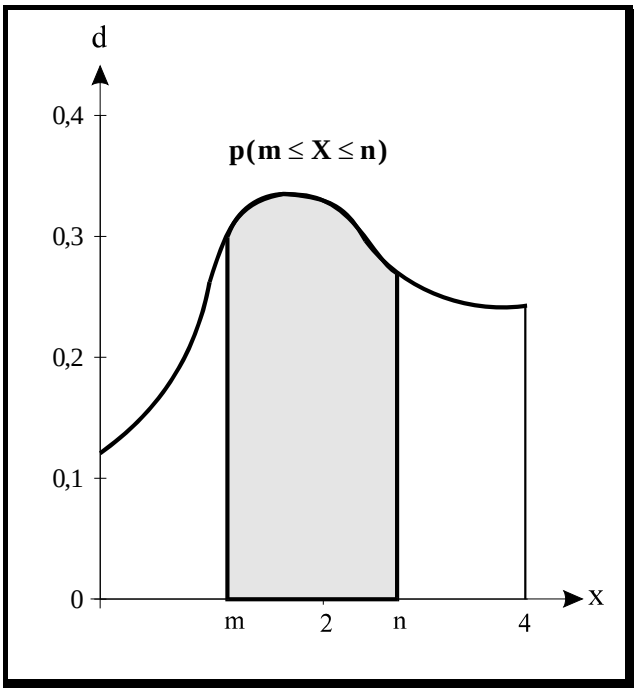


Si procedemos como anteriormente, reduciendo a la mitad la amplitud de los intervalos e incrementando la muestra se tiene:

X	f_r	densidad
$[0,0'25)$	0'032	0'128
$[0'25,0'5)$	0'038	0,152
$[0'5,0'75)$	0'039	0'156
$[0'75,1)$	0'041	0'164
$[1,1'25)$	0'06	0'24
$[1'25,1'50)$	0'07	0'28
$[1'50,1'75)$	0'07	0'28
$[1'75,2)$	0'08	0'32
$[2,2'25)$	0'08	0'32
$[2'25,2'5)$	0'075	0'3
$[2'5,2'75)$	0'073	0'292
$[2'75,3)$	0'072	0'288
$[3,3'25)$	0'071	0'284
$[3'25,3'5)$	0'069	0'276
$[3'5,3'75)$	0'067	0'268
$[3'75,4)$	0'063	0'252
	1	



Idealizando el proceso de reducción progresiva de la longitud de los intervalos y la extensión del número de datos al total de la población (a fin de considerar las frecuencias relativas de las clases como probabilidades de las mismas), el polígono superior que delimita los rectángulos se transforma en una línea continua bajo la cual el área encerrada es la unidad, y la probabilidad de que la variable esté comprendida entre dos valores determinados es el área del rectángulo mixtilíneo definido por esos valores y la propia curva, como se indica en el gráfico adjunto.



Tal curva, es la representación gráfica de la función de probabilidad de la variable continua X, pues con ella podemos determinar probabilidades del tipo:

$$p(m \le X \le n)$$

y se conoce con el nombre de **función de densidad** de la variable aleatoria (o de la distribución de probabilidad) continua.

Función de densidad

La probabilidad en las distribuciones continuas se determina mediante funciones **f(x)**, ***que llamamos*** **funciones densidad**, caracterizadas por:

1. $\mathbf{f(x) \ge 0}$, *para todo* $a \le x \le b$ *siendo a y b los extremos del campo de existencia de la variable X.*

   Esta condición indica que la probabilidad siempre es un número positivo o nulo.

2. *El área bajo la curva determinada por* $f(x)$, *entre a y b, es 1. Esta condición impone que:*

$$\int_a^b \mathbf{f(x)\,dx = 1}$$

   Esta condición indica que la probabilidad del suceso seguro es 1.

3. *La probabilidad de que la variable continua X esté en el intervalo* $[m, n]$, *viene dada por el área bajo la función* $f(x)$ *y los límites m y n. Es decir*:

$$\mathbf{p(m \le X \le n) = \int_m^n f(x)\,dx}$$

Es necesario destacar que las ordenadas de la función densidad $f(x)$ no tienen ningún significado especial, pues las probabilidades, como reiteradamente se ha señalado, vienen determinadas por áreas bajo la curva y no por valores de $f(x)$.

En todo caso, este hecho nos informa de que en las distribuciones continuas **la probabilidad de que la variable tome un valor concreto, $\mathbf{p(X = m)}$, es 0**, ***como corresponde al área de un rectángulo de base un punto y altura*** $\mathbf{f(m)}$. Tal resultado es coherente con la conocida regla de Laplace, pues frente a un caso favorable (el valor $X = m$) tenemos infinitos posibles (todo el recorrido de X).

Ejemplo: **En un fábrica de automóviles una de las piezas del motor de un modelo determinado es producida automáticamente por una máquina. Por muy ajustada que esté la máquina no todas las piezas son iguales. ¿Cuál es la probabilidad de que una pieza mida 280 dm exactamente?**

La respuesta es 0 ya que la pregunta está mal formulada porque toda medida lleva consigo una cierta imprecisión. Si la imprecisión de la medida llega a las milésimas de milímetro, una pregunta como la anterior equivale a preguntar por la probabilidad de que la variable se encuentre en el intervalo $[279'995, 280'005]$. Esta probabilidad es muy pequeña pero sí que puede calcularse a través de la función densidad.

$$p(279'95 \le X \le 280'005) = \int_{279'995}^{280'005} f(x)\,dx$$

Esto no quiere decir que no pueda haber una pieza de 280 dm, sino que únicamente afirmamos que la probabilidad de encontrarla es nula, ya que:

$$p(X = 280\,dm) = \int_{280}^{280} f(x)\,dx = 0$$

Esta situación es nueva, ya que en el caso de las variables aleatorias discretas hay sucesos x para los cuales $p(x) = 0$ (valores imposibles) y otros x_i para los cuales $p(x_i) = p_i \neq 0$ (valores posibles) estos últimos en modo finito n, de modo que:

$$p\sum_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} p_i = 1$$

Es decir, hasta ahora si A era un suceso se verificaba la equivalencia:

$$\text{"A es imposible"} \quad \Leftrightarrow \quad p(A) = 0$$

Pero ahora, para las variables continuas solo podemos poner la implicación:

$$\text{"A es imposible"} \quad \Rightarrow \quad p(A) = 0$$

Como consecuencia de lo anterior, deben señalarse como iguales las probabilidades de los intervalos:

$$\mathbf{p(m \le X \le n) = p(m \le X < n) = p(m < X \le n) = p(m < X < n)}$$

Ejemplo: **Comprueba que la función definida por** $\mathbf{f(x) = \begin{cases} \dfrac{15-x}{98} & \text{si} \quad 1 \le x \le 15 \\ 0 & \text{si} \quad x < 1 \ \text{ó} \ x > 15 \end{cases}}$

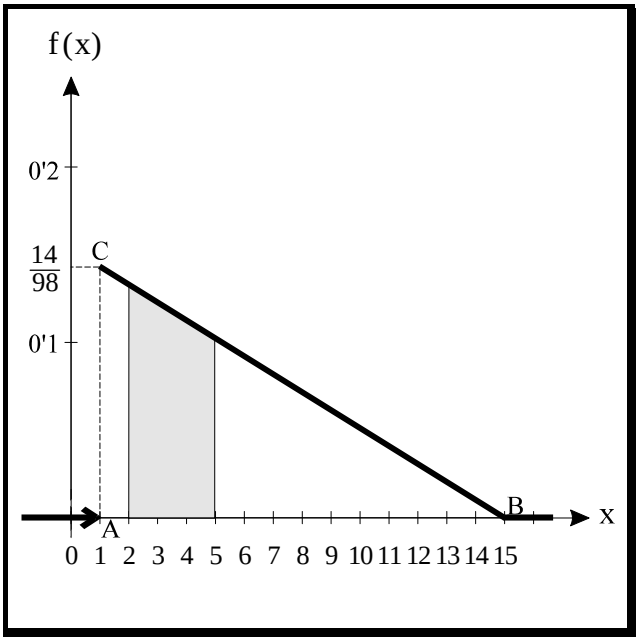
es función de densidad de cierta variable X. Halla la probabilidad de que X esté en el intervalo $[2,5]$.

Evidentemente, $f(x) \ge 0 \ \forall x$ y el área determinada por la función en el intervalo $[1,15]$ es (obsérvese que es el área del triángulo ABC)

$$\int_1^{15} \frac{15-x}{98}\,dx = \left[\frac{15}{98}x - \frac{x^2}{196}\right]_1^{15} = 1$$

Además:

$$p(2 \le x \le 5) = \int_2^5 \frac{15-x}{98}\,dx = 0'352$$



Ejemplo: **Se conoce como distribución uniforme aquella cuya función de densidad toma un valor constante en su intervalo de existencia $[a,b]$ y 0 en el resto. Halla dicha función de densidad.**

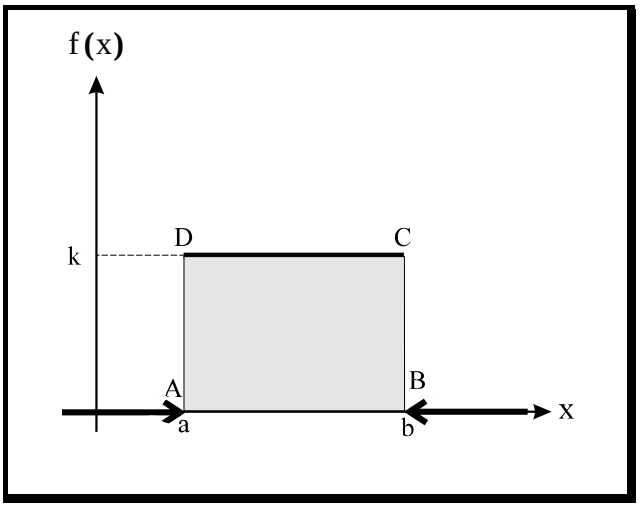
La función citada es:

$$f(x) = \begin{cases} k & \text{si} \quad a \le x \le b \\ 0 & \text{si} \quad x < a \ \text{ó} \ x > b \end{cases}$$

Puesto que el área del rectángulo ABCD debe ser la unidad:

$$\text{Área} = k \cdot (b-a) = 1 \Rightarrow$$

$$k = \frac{1}{b-a}$$



Función de distribución

La **función de distribución,** como en el caso de la variable aleatoria discreta, proporciona la probabilidad acumulada hasta un determinado valor de la variable.

$$\mathbf{F(x) = p(X \le x) = \int_a^x f(t)\,dt \quad \forall x \quad \Rightarrow \quad F'(x) = f(x)}$$

Obsérvese que la función de distribución es el límite del polígono de frecuencias relativas acumulado, cuando el tamaño del intervalo de clase tiende a cero.

Propiedades de la función de distribución

Pueden destacarse las siguientes características de la función de distribución $F(x)$ de una variable X, cuyo campo de definición se circunscribe al intervalo $[a, b]$:

1. Puesto que $F(x)$ es el valor de una probabilidad, se verifica $\mathbf{0 \le F(x) \le 1}$

2. La función de distribución $F(x)$ es nula para todo valor de x anterior al menor valor de la variable aleatoria y es igual a la unidad para todo valor de x posterior al mayor valor de la variable aleatoria, es decir:

$$\mathbf{F(x) = 0 \quad si \quad x \le a \quad y \quad F(x) = 1 \quad si \quad x \ge b}$$

3. $F(x)$ es monótona creciente.

4. $\mathbf{F'(x) = f(x)}$, con $f(x)$ función de densidad de X.

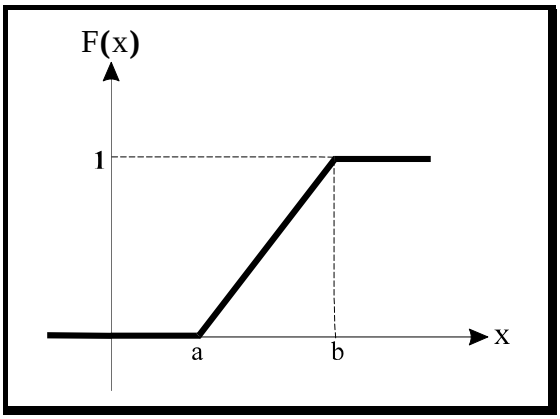
Ejemplo: Calcular la función de distribución de la llamada distribución uniforme.

La función de densidad se estableció como: $$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad a \le x \le b \\ 0 & \text{si} \quad x < a \ \text{ó} \ x > b \end{cases}$$

La función de distribución es:

$$F(x)=\int_a^x \frac{1}{b-a}dt=\left[\frac{t}{b-a}\right]_a^x=$$

$$\frac{x}{b-a}-\frac{a}{b-a}=\frac{x-a}{b-a} \quad a\le x\le b$$



Ejemplo: Sea la función de distribución de una variable X

$$\mathbf{F(x)}=\begin{cases}\mathbf{0} & \mathbf{si} \quad \mathbf{x<0}\\ \mathbf{x^3} & \mathbf{si} \quad \mathbf{0\le x\le 1}\\ \mathbf{1} & \mathbf{si} \quad \mathbf{x>1}\end{cases}$$

Halla su función de densidad y las probabilidades: $\mathbf{p\left(X\le\frac{2}{3}\right)}$ **y** $\mathbf{p\left(\frac{1}{5}<X\le\frac{2}{3}\right)}$

Sabemos por la propiedad 4 que la derivada de la función de distribución F(x) nos proporciona la función de densidad, por tanto:

$$f(x)=F'(x)=\begin{cases}0 & si \quad x\le 0\\ 3x^2 & si \quad 0<x<1\\ 0 & si \quad x>1\end{cases}$$

Por otra parte tenemos:

$$p\left(X\le\frac{2}{3}\right)=F\left(\frac{2}{3}\right)=\left(\frac{2}{3}\right)^3=\frac{8}{27} \qquad p\left(\frac{1}{5}<X\le\frac{2}{3}\right)=F\left(\frac{2}{3}\right)-F\left(\frac{1}{5}\right)=\frac{8}{27}-\frac{1}{125}=0'288$$

Media de una variable aleatoria continua

En el caso de la variables probabilísticas discretas teníamos $\mu=\sum_{i=1}^{n}x_ip_i$.

Si pasamos a una variable continua, al tener que efectuar la suma para todos los posibles valores de x, los dx correspondientes se hacen más y más pequeños. Nos enfrentamos ahora con un problema análogo al del cálculo de áreas, que nos condujo a la definición de integral definida, lo que nos sugiere dar la siguiente definición para la media:

$$E(X) = \mu = \int_a^b x \cdot f(x)\, dx$$

siendo $[a,b]$ el intervalo fuera del cual $f(x)$ es nula y siendo $f(x)$ la función "densidad de probabilidad" que sustituye a la probabilidad p_i en el caso de las discretas.

La media de una variable aleatoria continua también recibe el nombre de **esperanza matemática** o **valor esperado**.

Varianza de una variable aleatoria continua

La varianza de una variable aleatoria discreta es $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot p_i = \sum_{i=1}^{n} x_i^2 \cdot p_i - \mu^2$.

La dispersión de los valores de una variable aleatoria respecto a la media μ, se cuantifica mediante la varianza, cuya definición en el caso continuo es**:**

$$\sigma^2 = \int_a^b (x-\mu)^2 \cdot f(x)\, dx = \int_a^b x^2 \cdot f(x)\, dx - \mu^2$$

expresando, cuanto mayor sea su valor, una mayor variablidad de X respecto a μ.

Una medida de la dispersión en las mismas unidades que la variable X nos la da la raíz cuadrada positiva de la varianza, que se conoce, por la **desviación típica** de X. Se representa por σ.

Ejemplo: Calcular la media y la varianza de la distribución uniforme.

$$\mu = \int_a^b x \cdot \frac{1}{b-a} \cdot dx = \left[\frac{x^2}{2(b-a)} \right]_a^b = \frac{a+b}{2}$$

$$\sigma^2 = \int_a^b x^2 \cdot \frac{1}{b-a} \cdot dx - \left(\frac{a+b}{2} \right)^2 = \left[\frac{x^3}{3(b-a)} \right]_a^b - \left(\frac{a+b}{2} \right)^2 = \frac{(a-b)^2}{12}$$

Ejemplo: Hallar la media y la desviación típica de una variable aleatoria que tiene la siguiente función de densidad:

$$\mathbf{f(x)} = \begin{cases} \mathbf{0} & \mathbf{si} \quad \mathbf{x < 0} \\ \dfrac{\mathbf{1}}{\mathbf{3}} & \mathbf{si} \quad \mathbf{0 \le x \le 3} \\ \mathbf{0} & \mathbf{si} \quad \mathbf{3 < x} \end{cases}$$

$$\mu = \int_0^3 x \cdot \frac{1}{3} \cdot dx = \left[\frac{1}{6} x^2 \right]_0^3 = \frac{9}{6} = 1'5$$

$$\sigma^2 = \int_0^3 x^2 \cdot \frac{1}{3} \cdot dx - 1'5^2 = \left[\frac{x^3}{9} \right]_0^3 - 1'5^2 = 3 - 1'5^2 = 0'75 \quad \Rightarrow \quad \sigma = \sqrt{0'75} = 0'86$$

Ejemplo: La función de densidad de una variable aleatoria continua viene definida por:

$$\mathbf{f(x)} = \begin{cases} \mathbf{2x} & \mathbf{si} \quad \mathbf{0 \le x \le 1} \\ \mathbf{0} & \mathbf{si} \quad \mathbf{x < 0} \ \ \mathbf{ó} \ \ \mathbf{x > 1} \end{cases}$$

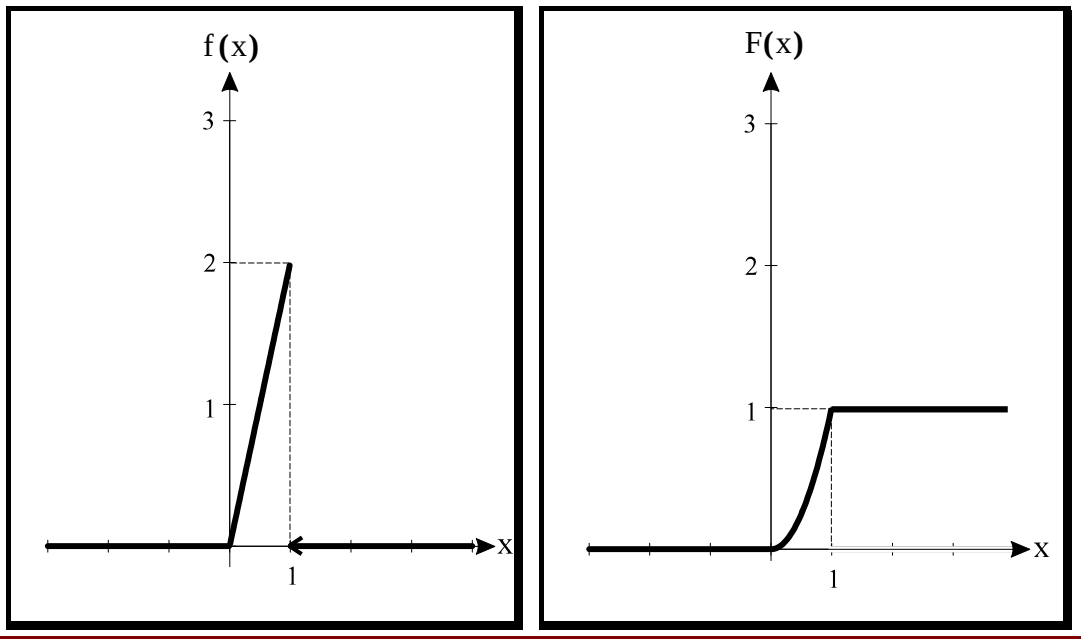
a) Hallar la función de distribución F(x). b) Representar f(x) y F(x).

c) Hallar la media, la varianza y la desviación típica.

a) Partiendo de la igualdad $F'(x) = f(x)$ obtenemos la función de distribución, que será:

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ \int_0^x 2t\,dt & si \quad 0 \le x \le 1 \\ 1 & si \quad x > 1 \end{cases} \quad \Rightarrow \quad F(x) = \begin{cases} 0 & si \quad x < 0 \\ x^2 & si \quad 0 \le x \le 1 \\ 1 & si \quad x > 1 \end{cases}$$

b)



c)

$$\mu = \int_0^1 x \cdot 2x\,dx = \left[\frac{2x^3}{3}\right]_0^1 = \frac{2}{3}$$

$$\sigma^2 = \int_0^1 x^2 \cdot 2x\,dx - \left(\frac{2}{3}\right)^2 = \left[\frac{x^4}{2}\right]_0^1 - \left(\frac{2}{3}\right)^2 = 0'055 \quad \Rightarrow \quad \sigma = 0'235$$

Distribución Normal de probabilidad

Existen diversas distribuciones continuas de interés y varias de ellas se han interpretado como modelos ideales de distribución de probabilidad. Pero de todas, la que merece especial atención es la que se conoce como distribución **normal**. Entre la cantidad de variables que se distribuyen normalmente podemos citar:

- ***Caracteres morfológicos de individuos*** (personas, animales, plantas) de una misma raza. Por ejemplo, tallas, pesos, envergaduras, longitud de las hojas de un árbol o de las vainas de una plantación de guisantes, etc.

- ***Caracteres fisiológicos***. Por ejemplo, efecto de una misma dosis de un fármaco, o de una misma cantidad de abono.

- ***Caracteres sociológicos***. Por ejemplo, consumo de ciertos productos por individuos de un mismo grupo humano, aceptación de una norma o costumbre por los miembros de una comunidad, etc.

- ***Caracteres psicológicos***. Por ejemplo, coeficiente intelectual de los individuos, grado de adaptación a un medio,...

- ***Caracteres físicos***. Por ejemplo, resistencia a la rotura de piezas, distribución de los errores en una serie de medidas.

- ***Económicos***. Por ejemplo, el impacto de un nuevo producto en el mercado o el consumo de un bien no básico cuyo precio se mantiene estable.

Existen también muchos fenómenos que responden a una distribución simétrica y que se pueden considerar como suma de una serie de efectos parciales independientes. Y puede ocurrir que esos efectos no se ajusten a la normal; pero, sin embargo, el fenómeno resultante tiende asintóticamente a la distribución normal. Así, en el peso de los individuos de una población influyen una serie de efectos causados por la herencia genética, por el tipo de alimentación, por el clima, por el tipo de trabajo, etc. Algunos de estos efectos, o todos, puede ser que no se distribuyan normalmente; sin embargo, la concurrencia de todos ellos hace que el peso tienda asintóticamente a la distribución normal. Éste es un resultado importantísimo que se conoce con el nombre de **teorema central del límite** que fue demostrado en 1901 por el matemático ruso **Alexander Michailowicz Liapunov** (1857-1918), y se suele enunciar de formas distintas, todas ellas en términos de gran complejidad.

Este es el comportamiento que describe la ley normal y su "caldo de cultivo" se encontrará, entonces, en todos aquellos fenómenos que están sometidos a complejos y variados factores, donde la influencia de unos sobre otros eliminará los casos extremos, haciéndoles infrecuentes frente al tipo medio, que será el de más corriente aparición.

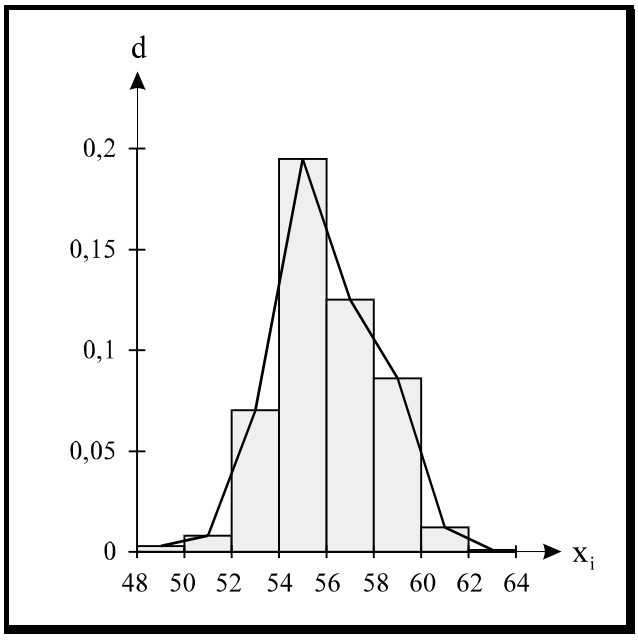
Aunque de lo expuesto anteriormente parece deducirse que, en efecto, todas las distribuciones se comportan según el modelo normal esto no es cierto. Consideremos un ejemplo que está bien presente en nuestra sociedad. Clasifiquemos a los ciudadanos españoles según su nivel de renta. Evidentemente, son muy pocos los españoles que poseen niveles de rentas altas y en cambio son muchos los que poseen niveles de rentas bajas. Por tanto, la distribución no es simétrica y, en consecuencia, no se adapta al modelo normal.

No es este el único ejemplo claro de distribuciones que no se ajustan al modelo normal, por lo que podemos concluir que por supuesto no todas las distribuciones son normales, pero sí un gran número de ellas. De hecho, ***la distribución normal se llama así porque durante mucho tiempo se pensó que ése era el comportamiento normal de todos los fenómenos.***

Variable aleatoria de la distribución normal

La distribución de 500 estudiantes de COU, respecto del perímetro craneal, es la siguiente:

x_i	f_i	f_r	densidad
$[48,50)$	3	0'006	0'003
$[50,52)$	8	0'016	0'008
$[52,54)$	70	0'14	0'070
$[54,56)$	195	0'39	0'195
$[56,58)$	125	0'25	0'125
$[58,60)$	86	0'172	0'086
$[60,62)$	12	0'024	0'012
$[62,64)$	1	0'002	0'001
	500	1	



Si el número de estudiantes crece indefinidamente y vamos haciendo cada vez más pequeña la amplitud de los intervalos, el polígono de densidad de frecuencias toma entonces la forma de la distribución normal. La variable, en este caso, sería continua y su recorrido sería, en principio el intervalo $[48,64$.

Para definir una variable aleatoria continua es preciso conocer dos datos:

1º El recorrido o intervalo de variabilidad de la variable.

2º Su función de densidad o la ecuación de la curva de probabilidad.

Para el caso de una distribución normal teórica, se dice que una variable aleatoria continua X sigue una distribución normal de media μ y desviación típica σ, y se designa por $\mathbf{N(\mu,\sigma)}$, si se cumplen las siguientes condiciones:

1º La variable recorre toda la recta real, es decir $(-\infty,+\infty)$.

2º La función de densidad, que es la expresión en términos de ecuación matemática de la curva de Gauss, es:

$$\mathbf{f(x)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}}$$

e = **2'7182……** es la base de los logaritmos neperianos.

π = **3'1415…..** es la relación de la longitud de una circunferencia a su diámetro

x = **abscisa** cualquier valor de un punto del intervalo.

μ **= media** de la variable aleatoria X. (parámetro)

σ **= desviación típica** de la variable aleatoria X .

A los valores μ y σ se los denomina **parámetros de la distribución normal**.

Hay que probar que esta función verifica las condiciones para las funciones de densidad de probabilidad.

a) $\mathbf{f(x)\geq 0}$

Es evidente puesto que la función exponencial toma siempre valores positivos cualquiera que sea el exponente.

b) $\mathbf{\int_{-\infty}^{+\infty} f(x)\,dx = 1}$

En nuestro caso será: $$\int_{-\infty}^{\infty}\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)}\,dx = 1$$

Lo admitimos, puesto que es una integral no estudiada hasta ahora y muy complicada de demostrar.

Propiedades de la función de densidad

1. La curva tiene forma de campana, de ahí que se la denomine **campana de Gauss**, y es simétrica respecto a la recta vertical $\mathbf{x = \mu}$.

$$f(\mu - \varepsilon) = f(\mu + \varepsilon) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

2. La ordenada máxima se obtiene precisamente para el valor de $x = \mu$, donde se sitúan también la moda y la mediana.

$$\text{Máximo en } p\left(\mu, \frac{1}{\sigma \cdot \sqrt{2\pi}}\right)$$

$$f'(x) = -\frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} \cdot \frac{2(x-\mu)}{2\sigma^2} = 0 \quad \rightarrow \quad \frac{x-\mu}{\sigma^2} = 0 \quad \Rightarrow \quad x = \mu$$

Como $f''(x) < 0 \quad \Rightarrow \quad x = \mu$ es máximo relativo.

Para $-\infty < x < \mu \quad \Rightarrow \quad f'(x) > 0 \quad \Rightarrow \quad f(x)$ es creciente.

Para $\mu < x < +\infty \quad \Rightarrow \quad f'(x) < 0 \quad \Rightarrow \quad f(x)$ es decreciente.

Si $f''(x) = 0 \quad \Rightarrow \quad x_1 = \mu + \sigma$ y $x_2 = \mu - \sigma$ son puntos de inflexión.

3. El área del recinto encerrado bajo la curva, el eje de abscisas y las ordenadas en los puntos $(\mu - \sigma, \mu + \sigma)$ (es decir, la probabilidad de que la variable se encuentre en el intervalo $[\mu - \sigma, \mu + \sigma]$) es 0'6826.

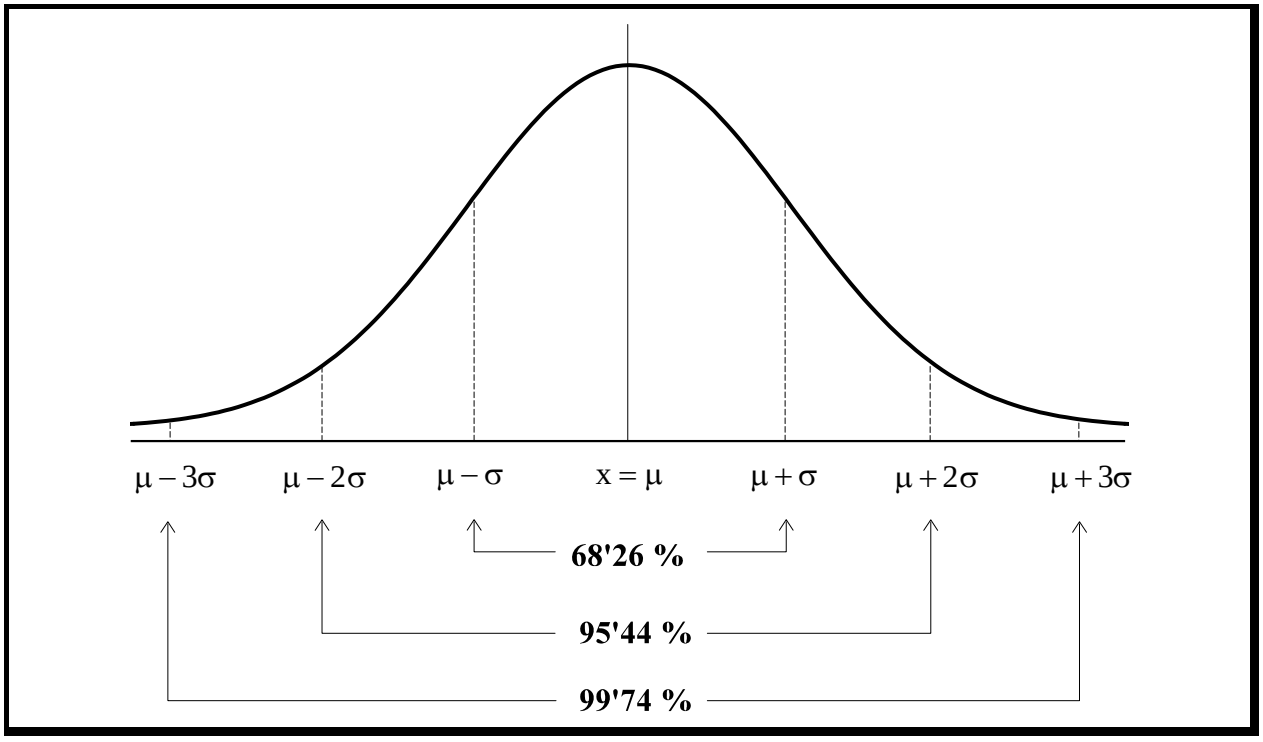
La correspondiente entre $(\mu - 2\sigma, \mu + 2\sigma)$ es de 0'9544 y entre $(\mu - 3\sigma, \mu + 3\sigma)$ es prácticamente la unidad: 0'9974.

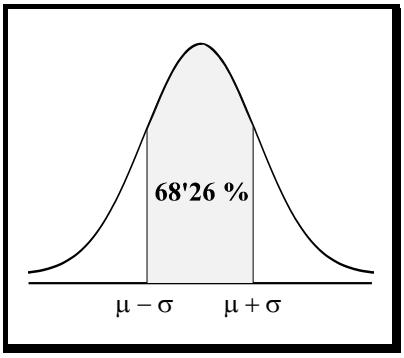
Al ser simétrica respecto al eje vertical que pasa por μ, dicho eje de simetría deja un área igual a 0'5 a la izquierda y otra igual a 0'5 a la derecha.

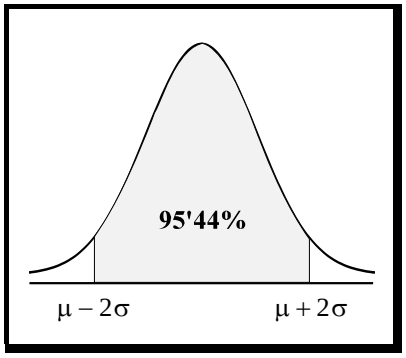
4. A ambos lados del valor $x = \mu$ las ordenadas de la curva decrecen, primero lentamente, y después con mayor rapidez, hasta hacerse su ordenada nula a medida que la abscisa se aleja hacia la derecha o hacia la izquierda del valor central. Se dice que la curva se acerca asintóticamente al eje de abscisas.

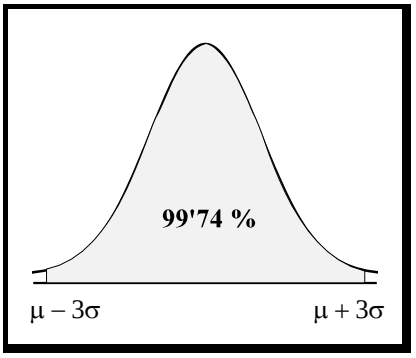
$$\lim_{x \to +\infty} f(x) = 0 \qquad \lim_{x \to -\infty} f(x) = 0$$

Las siguientes gráficas reflejan todo lo expuesto anteriormente:

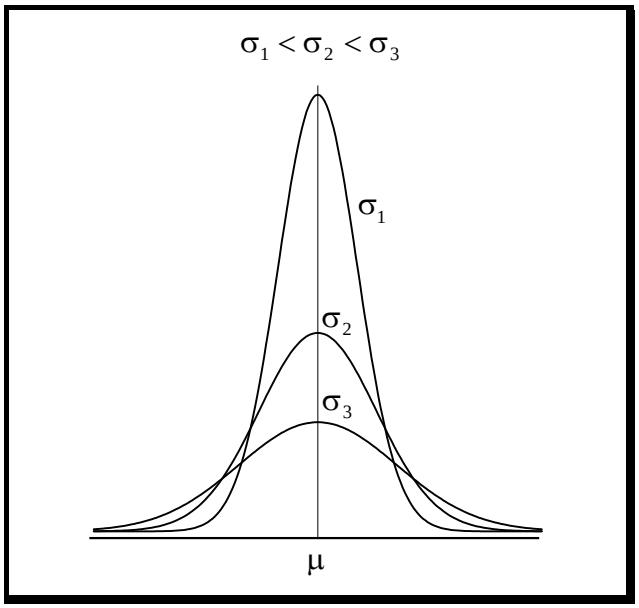








En las siguientes gráficas se observa que cuando la desviación típica es elevada aumenta la dispersión y, en consecuencia, la gráfica es menos estilizada. Por el contrario, para valores de σ muy pequeños la dispersión disminuye y, en consecuencia, la gráfica de la función es mucho más estilizada y concentrada en torno a la media. En cualquier caso, el área encerrada bajo cualquiera de las curvas anteriores es la unidad.



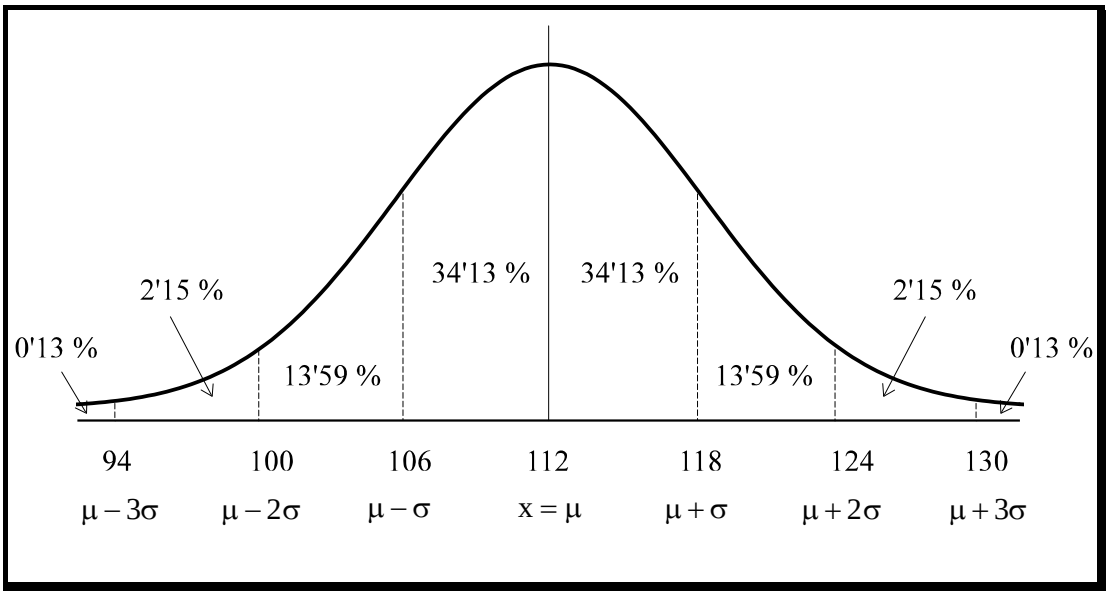
Una curva normal queda completamente identificada conociendo el valor de los parámetros μ y σ, es decir, su media y su desviación típica, por lo que las funciones de densidad de las variables normales constituyen una familia de curvas determinadas por esos valores, designándose cada una de ellas por $N(\mu,\sigma)$.

Ejemplo: **El coeficiente intelectual de los 5600 alumnos de COU de una provincia se distribuyen según una distribución normal $N(112,6)$. Calcular aproximadamente cuántos de ellos tienen :**

a) Más de 112; b) Entre 106 y 118; c) Entre 106 y 112; d) Menos de 100

e) Más de 130; f) Entre 118 y 124.

Con los datos que hemos obtenido teóricamente podemos construir la siguiente distribución:



Con ella es fácil responder a las seis preguntas:

a) $p[X \geq 112] = 50\%$ pues 112 es la media. Por tanto, habrá unos 2800 alumnos que cumplan esta condición.

b) $p[106 \leq X \leq 118] = 68'26\%$. Habrá $0'6826 \cdot 5600 \cong 3823$ alumnos.

c) Es la mitad de la anterior $\cong 1911$ alumnos.

d) $p[X \leq 100] = 0'13\% + 2'15\% = 2'28\%$. Habrá $0'0228 \cdot 5600 \cong 128$ alumnos.

e) $p[130 \leq X] = 0'13\%$. Habrá $0'0013 \cdot 5600 \cong 7$ alumnos.

f) $p[118 \le X \le 124] = 13'59\%$. Habrá $0'1359 \cdot 5600 \cong 761$

Hasta ahora, todas las probabilidades que hemos estudiado en distribuciones normales se han hecho cuidando que los extremos de los intervalos distaran de la media un número entero de desviaciones típicas, porque solo sabemos hasta ahí. Necesitamos conocer la distribución de probabilidades para intervalos más pequeños, cuyas amplitudes sean fracciones de σ.

Distribución normal estándar

El cálculo de las diferentes áreas bajo la curva normal, que nos muestran las diversas probabilidades de la distribución, exigiría resolver distintas integrales definidas de la función de densidad, variables según el valor de μ y σ considerado.

Ahora bien, si X es una variable aleatoria de parámetros μ y σ podemos realizar las siguientes transformaciones:

$$\mu - \sigma \le X \le \mu + \sigma$$

$$-\sigma \le X - \mu \le \sigma$$

$$-1 \le \frac{X-\mu}{\sigma} \le +1$$

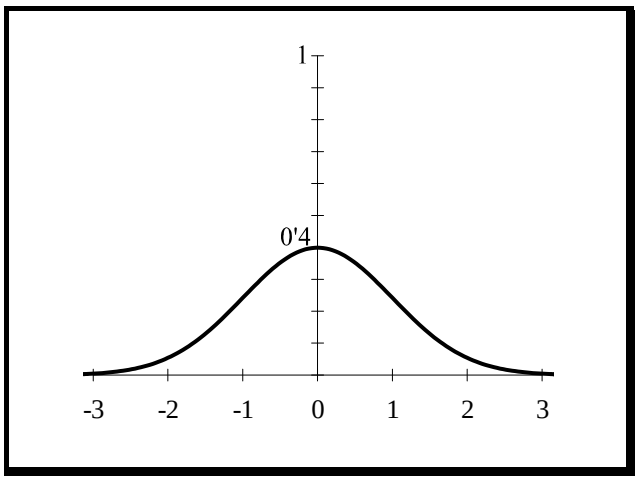
resultando que la variable aleatoria $Z = \dfrac{X-\mu}{\sigma}$ tiene de media 0 y desviación típica 1 y se llama ***variable aleatoria tipificada***.

De las infinitas distribuciones $N(\mu,\sigma)$ tiene especial interés la distribución **$N(0,1)$**; es decir, aquella que tiene por media el valor cero $(\mu = 0)$ y por desviación típica la unidad $(\sigma = 1)$. Esta distribución se llama **distribución normal estándar**, o bien **distribución normal reducida**.

La función de densidad para $\mu = 0$ y $\sigma = 1$ es:

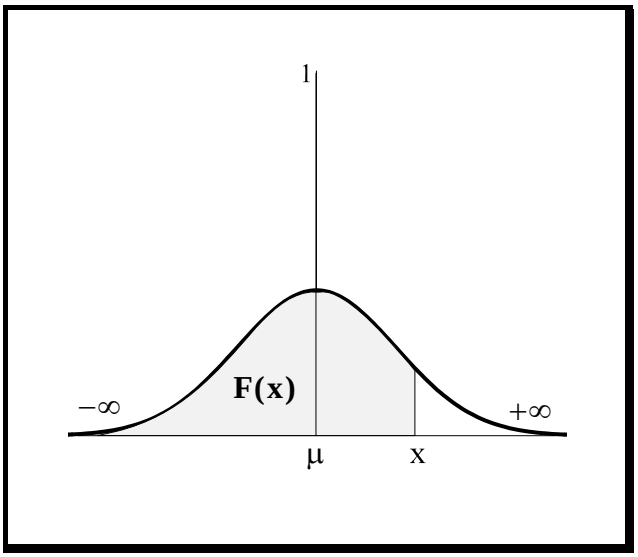
$$\mathbf{f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}}$$

cuya representación gráfica se muestra al margen.



La función de distribución de la ley normal estándar proporciona el área del recinto sombreado de la figura. Dependiendo del valor que tome en cada caso la variable X, se obtendrá un área distinta que, por tratarse de una probabilidad, será menor o igual a la unidad.

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \, dt$$



Con el fin de facilitar el cálculo de esta superficie y no tener que utilizar en todo momento el cálculo integral, se han elaborado una tablas de muy fácil uso que permiten calcular las probabilidades correspondientes a las más diversas situaciones sin más que tener en cuenta la simetría de la distribución, y que el área desde $-\infty$ a 0 es 0'5.

Tipificación de la variable

En realidad, no hay ningún fenómeno que siga exactamente una distribución $N(0,1)$, y por otra parte es obvio que no se pueden construir tablas para todos los tipos posibles de distribuciones $N(\mu,\sigma)$, pues μ y σ toman infinitos valores.

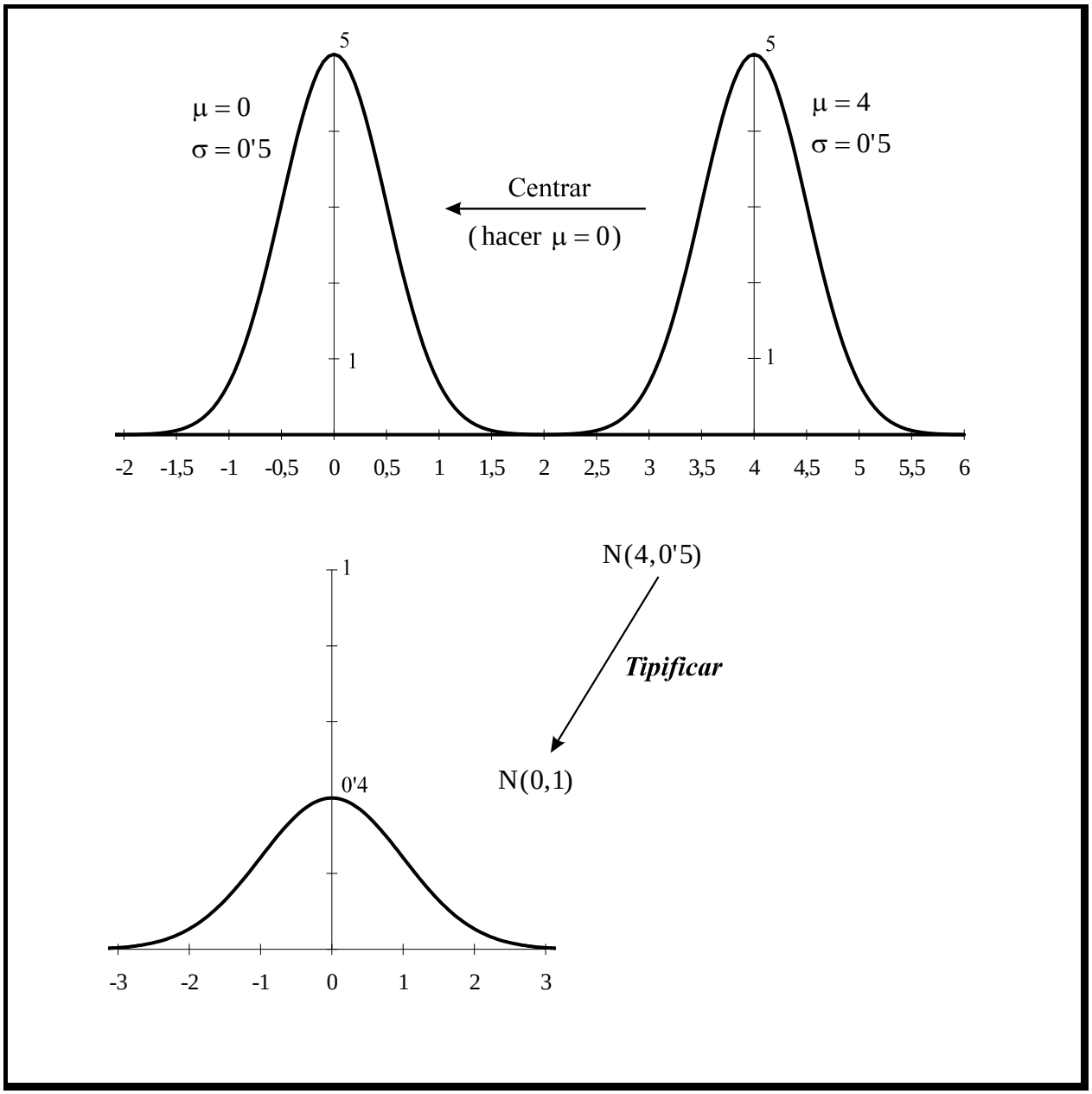
Lo más aconsejable sería poder transformar la variable **X** que sigue una distribución $N(\mu,\sigma)$ en otra variable **Z** que siga una distribución $N(0,1)$. Esta transformación se conoce con el nombre de **tipificación de la variable**.

Para llevar a cabo esta transformación. es obvio que hay que realizar dos pasos:

1º **Centrar**, es decir, trasladar la media de la distribución al origen de coordenadas. esto equivale a hacer $\mu = 0$.

2º **Reducir** la desviación estándar a 1 $(\sigma = 1)$. Esto equivale a **dilatar** o **contraer** la gráfica de la distribución para que coincida con la de la ley estándar.

Podemos observar estos dos pasos en el siguiente esquema:



Estos dos pasos se consiguen simultáneamente sin más que hacer el siguiente cambio de variable:

$$Z = \frac{X - \mu}{\sigma}$$

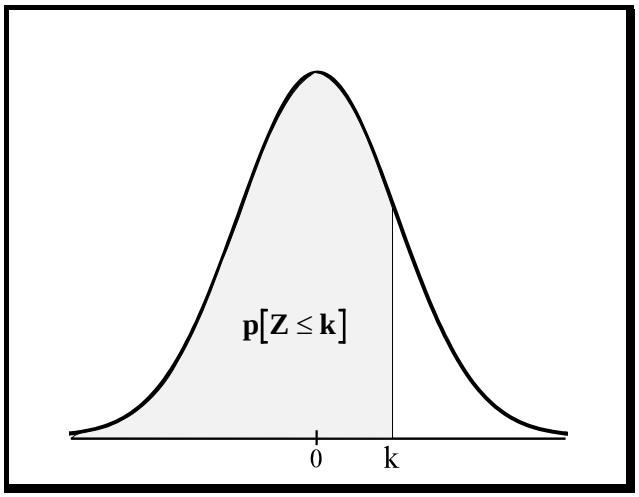
Manejo de las tablas

Sea Z una variable que sigue una distribución normal $N(0,1)$. Existe una tabla de valores obtenidos por el cálculo numérico de las probabilidades del tipo $p(Z \le k)$.

En general, para calcular las probabilidades $\mathbf{p(Z \le k)}$***, en las tablas únicamente se dan los valores para*** $\mathbf{k > 0}$***. Esto es suficiente, ya que el área total bajo la curva de densidad es la unidad y esta curva es simétrica respecto al eje de ordenadas***.

Distribución Normal

$$F(k) = p(Z \le k) = \int_{-\infty}^{k} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}\, dt$$



Para cualquier valor de k positivo tenemos las siguientes gráficas, teniendo en cuenta que los valores de las probabilidades del tipo $p(Z \le k)$ vienen tabulados para k entre 0 y 3'99:

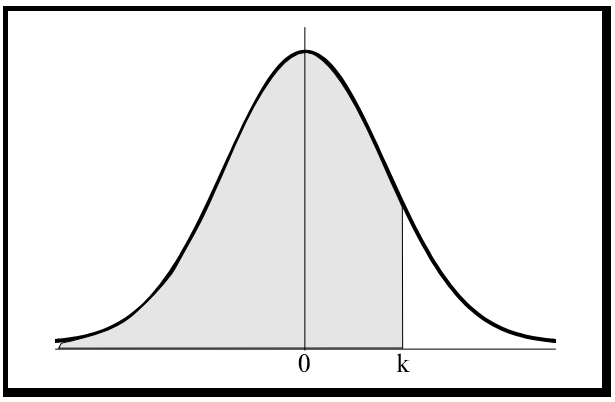
❒ $p(Z \le k)$

Representa el área encerrada entre la curva y el eje de abscisas en el intervalo $[-\infty, k]$ **como se indica en la gráfica.**

Para calcular el valor de $p(Z \le 1'45)$ en la tabla, buscaremos 1'4 en la primera columna y la intersección de su fila con la columna 5 nos da la probabilidad:

$$p(Z \le 1'45) = 0'9265$$

Esto quiere decir que el 92'65 % de las observaciones se distribuyen entre $-\infty$ y 1'45



❒ $p(Z \le -k)$

Representa el área encerrada entre la curva y el eje de abscisas en el intervalo $[-\infty, -k]$ **como se indica en la gráfica.**

La tabla sólo proporciona probabilidades para valores de Z positivos. Pero teniendo en cuenta la simetría de la función de densidad, y que el área encerrada por toda la curva es igual a la unidad resulta:

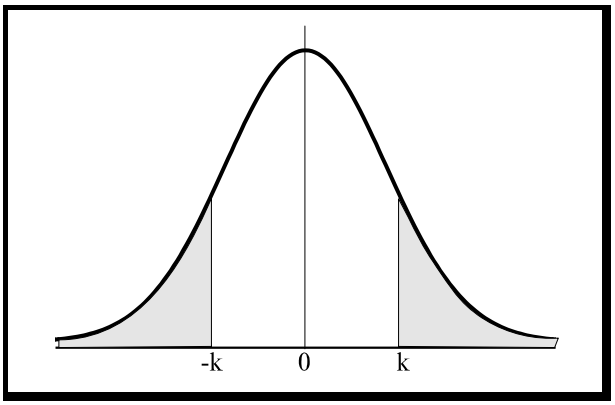
$$p(Z \leq -k) = p(Z \geq k) = 1 - p(Z \leq k)$$

Para calcular el valor de $p(Z \leq -1'45)$ tenemos:

$$p(Z \leq -1'45) = p(Z \geq 1'45) =$$

$$1 - p(Z \leq 1'45) = 1 - 0'9265 = 0'0735$$

Este resultado indica que únicamente un 7'35% de las observaciones se distribuyen entre $-\infty$ y $-1'45$.



❒ $p(Z \geq k)$

Representa el área encerrada entre la curva y el eje de abscisas en el intervalo $[k, \infty]$ como se indica en la gráfica.

Teniendo en cuenta la simetría de la función de densidad, y que el área encerrada por toda la curva es igual a la unidad resulta:

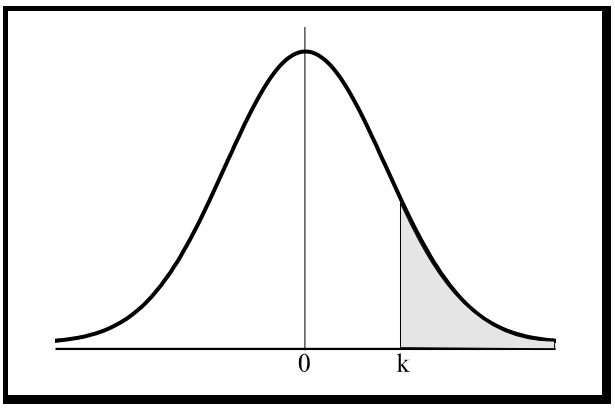
$$p(Z \geq k) = 1 - p(Z \leq k)$$

Para calcular el valor de $p(Z \geq 1'45)$ tenemos:

$$p(Z \geq 1'45) = 1 - p(Z \leq 1'45) =$$

$$1 - 0'9265 = 0'0735$$

Este resultado indica que únicamente un 7'35% de las observaciones se distribuyen entre $1'45$ e ∞.



❒ $p(Z \geq -k)$

Representa el área encerrada entre la curva y el eje de abscisas en el intervalo $[-k, \infty]$ como se indica en la gráfica.

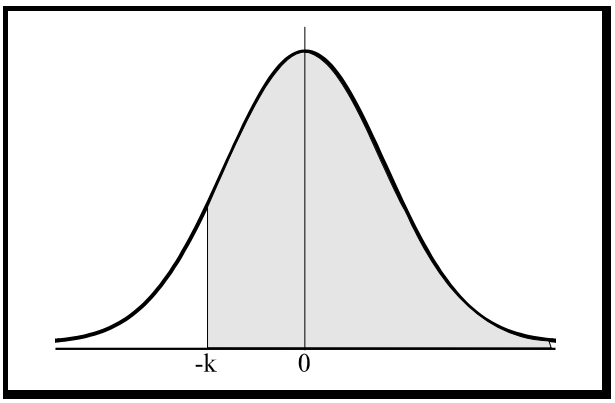
Teniendo en cuenta la simetría de la función de densidad, y que el área encerrada por toda la curva es igual a la unidad resulta:

$$p(Z \geq -k) = p(Z \leq k)$$

Para calcular el valor de $p(Z \geq -1'45)$ tenemos:

$$p(Z \geq -1'45) = p(Z \leq 1'45) = 0'9265$$

Este resultado indica que el 92'65 % de las observaciones se distribuyen entre $-1'45$ e ∞.



❒ $p(k_1 \leq Z \leq k_2)$

Representa el área encerrada entre la curva y el eje de abscisas en el intervalo $[k_1, k_2]$ como se indica en la gráfica.

De la figura deducimos:

$$p(k_1 \leq Z \leq k_2) = p(Z \leq k_2) - p(Z \leq k_1)$$

y las probabilidades de la diferencia se reducen a un caso de los anteriores.

Se verifican las siguientes igualdades:

$$p(k_1 \leq Z \leq k_2) = p(k_1 < Z \leq k_2) = p(k_1 \leq Z < k_2)$$

- Supongamos que queremos calcular el valor de $p(-0'53 \leq Z < 2'46)$. Entonces podemos establecer la siguiente igualdad:

$$p(-0'53 \leq Z < 2'46) =$$

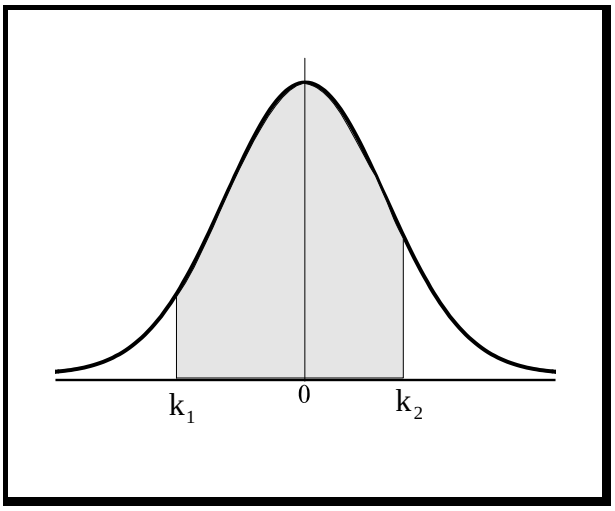
$$p(Z \leq 2'46) - p(Z \leq -0'53) =$$

$$p(Z \leq 2'46) - p(Z \geq 0'53) =$$

$$p(Z \leq 2'46) - (1 - p(Z \leq 0'53)) =$$

$$0'9931 - (1 - 0'7019) = 0'695$$

Es decir, que el 69'5 % de las observaciones se encuentran entre $-0'53$ y $2'46$.



- Supongamos que queremos calcular el valor de $p(-2'57 \le Z < -1'25)$. Como consecuencia de la simetría de la función de densidad se tiene:

$$p(-2'57 \le Z < -1'25) = p(1'25 \le Z < 2'57) = 0'1005$$

Es decir, únicamente el 10'05 % de las observaciones se distribuyen entre 1'25 y 2'57, o bien entre $-2'57$ y $-1'25$.

Caso particular de áreas bajo la curva

A veces tiene interés saber qué proporción de individuos se distribuyen en intervalos de la forma $(\mu - k\sigma, \mu + k\sigma)$, siendo k un número entero.

Supongamos que tenemos una variable aleatoria X que sigue una distribución Normal $N(\mu, \sigma)$, ¿qué proporción de individuos se encontrará en el intervalo $(\mu - \sigma, \mu + \sigma)$?

En realidad, este es un caso particular de manejo de tablas; únicamente hemos de tener en cuenta que hemos de tipificar previamente. Así pues, se tiene:

$$p(\mu - \sigma < X < \mu + \sigma) = p\left(\frac{\mu - \sigma - \mu}{\sigma} < Z \le \frac{\mu + \sigma - \mu}{\sigma}\right) = p(-1 < Z \le 1) = p(Z \le 1) - p(Z < -1) =$$

$$p(Z \le 1) - p(Z > 1) = p(Z \le 1) - \left[1 - p(Z < 1)\right] = 2p(Z \le 1) - 1 = 2 \cdot 0'8413 - 1 = 0'6826$$

Es decir, el 68'26 % de las observaciones (o de los individuos) se encuentra en el intervalo $(\mu - \sigma, \mu + \sigma)$.

Siguiendo el mismo razonamiento se puede calcular que:

El 95'4 % de las observaciones está en $(\mu - 2\sigma, \mu + 2\sigma)$.

El 99'7 % de las observaciones está en $(\mu - 3\sigma, \mu + 3\sigma)$.

Modo de proceder inverso

A veces nos interesa saber entre qué valores se encuentra el 50 % de los individuos. Para ello hay que manejar las tablas en forma inversa, ya que en los casos anteriores se conocía el valor de la abscisa y se trataba de hallar el valor de la probabilidad, pero ***ahora se conoce el valor de la probabilidad y se trata de hallar el valor de la abscisa.***

Supongamos que deseamos saber cuál es el valor de la abscisa que deja a su izquierda el 75 % del área, o por simetría, el valor de la abscisa que deja a su derecha el 25 % del área.

Si X es una variable aleatoria $N(\mu, \sigma)$, se verifica $p\left(Z \le \frac{X - \mu}{\sigma}\right) = 0'75$. ***Buscando en las tablas una probabilidad lo más próxima a 0'75 se obtiene para z el valor 0'67,*** por tanto:

$$\frac{X-\mu}{\sigma}=0'67 \quad \Rightarrow \quad X-\mu=0'67\sigma \quad \Rightarrow \quad X=\mu+0'68\sigma$$

Supongamos que deseamos saber cuál es el valor de la abscisa que deja a su izquierda el 30% del área, es decir $p\left(Z\le\frac{x-\mu}{\sigma}\right)=0'30$.

Pero 0'30 es un valor de la probabilidad que no está en las tablas, ya que en ellas sólo aparecen los valores de la probabilidad comprendidos entre 0'5 y 1. El problema se soluciona efectuando el cálculo del valor de la abscisa que deja a su derecha el 30% del área, teniendo en cuenta que ésta tiene signo contrario, es decir:

$$p\left(Z\ge-\frac{x-\mu}{\sigma}\right)=0'30 \quad \rightarrow \quad 1-p\left(Z\le-\frac{x-\mu}{\sigma}\right)=0'30 \quad \rightarrow \quad 0'70=p\left(Z\le-\frac{x-\mu}{\sigma}\right)$$

$$-\frac{x-\mu}{\sigma}=0'52 \quad \rightarrow \quad -x+\mu=0'52\sigma \quad \rightarrow \quad x=-0'52\sigma+\mu$$

Ejemplo: Los 600 soldados de un cuartel poseen una altura que se distribuye según una normal $\mu=166\,cm.$ y $\sigma=12cm.$

a) Halla el número aproximado de soldados cuya altura esté comprendida entre los 165 y 182 cm.

b) ¿Cuántos medirán más de 190 cm.? ¿Cuántos medirán exactamente 168 cm.?

c) Si los mandos del cuartel deben formar un batallón de "gastadores" con el 4% de los soldados más altos ¿a partir de qué altura deben seleccionarse éstos?

a) Sea **X** la variable $N(166,12)$. Nos piden calcular $p(165\le X\le 182)$.

$$p(165\le X\le 182)=p\left(\frac{165-166}{12}\le Z\le\frac{182-166}{12}\right)=p(-0'083\le Z\le 1'333)=$$

$$p(Z\le 1'333)-p(Z\le -0'083)=p(Z\le 1'333)-p(Z\ge 0'083)=$$

$$p(Z\le 1'333)-\left[1-p(Z\le 0'083)\right]=0'9082-1+0'5319=0'4401$$

El número de soldados entre esas alturas será $0'4401\cdot 600\cong 264$

b) Nos piden calcular $p(X\ge 190)$.

$$p(Z\ge 190)=p\left(Z\ge\frac{190-166}{12}\right)=p(Z\ge 2)=1-p(Z\le 2)=1-0'9772=0'0228$$

El número de soldados cuya altura sea superior a 190 cm. será $0'0228 \cdot 600 \cong 14$

Para contestar a la pregunta de cuántos medirán exactamente 168 cm., si la contestamos en sentido estricto la respuesta es ninguno.

Interpretamos, pues, que medir 168 cm. significa medir entre 167'5 y 168'5 cm. Por tanto:

$$p(167'5 \le X \le 168'5) = p\left(\frac{167'5 - 166}{12} \le Z \le \frac{168'5 - 166}{12}\right) = p(0'125 \le Z \le 0'208) =$$

$$p(Z \le 0'208) - p(Z \le 0'125) = 0'5793 - 0'5478 = 0'0315$$

El número de soldados cuya altura es aproximadamente 168 cm. es:

$$0'0315 \cdot 600 \cong 19$$

c) Ha de cumplirse que $p(X \ge x) = p\left(Z \ge \frac{x - 166}{12}\right) = 0'04 = 1 - p\left(Z \le \frac{x - 166}{12}\right)$

$$p\left(Z \le \frac{x - 166}{12}\right) = 1 - 0'04 = 0'96 \quad \Rightarrow \quad \frac{x - 166}{12} = 1'75$$

Las tablas dan para el valor de z de 1'75 el valor de 0'9599 de probabilidad. La diferencia tan poco significativa, nos induce a despreciarla.

$$x = 1'75 \cdot 12 + 166 = 187 \text{ cm.}$$

Áreas limitadas por la curva N(0,1) desde $-\infty$ hasta k

k	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5754
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7518	0,7549
0,7	0,7580	0,7612	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9734	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

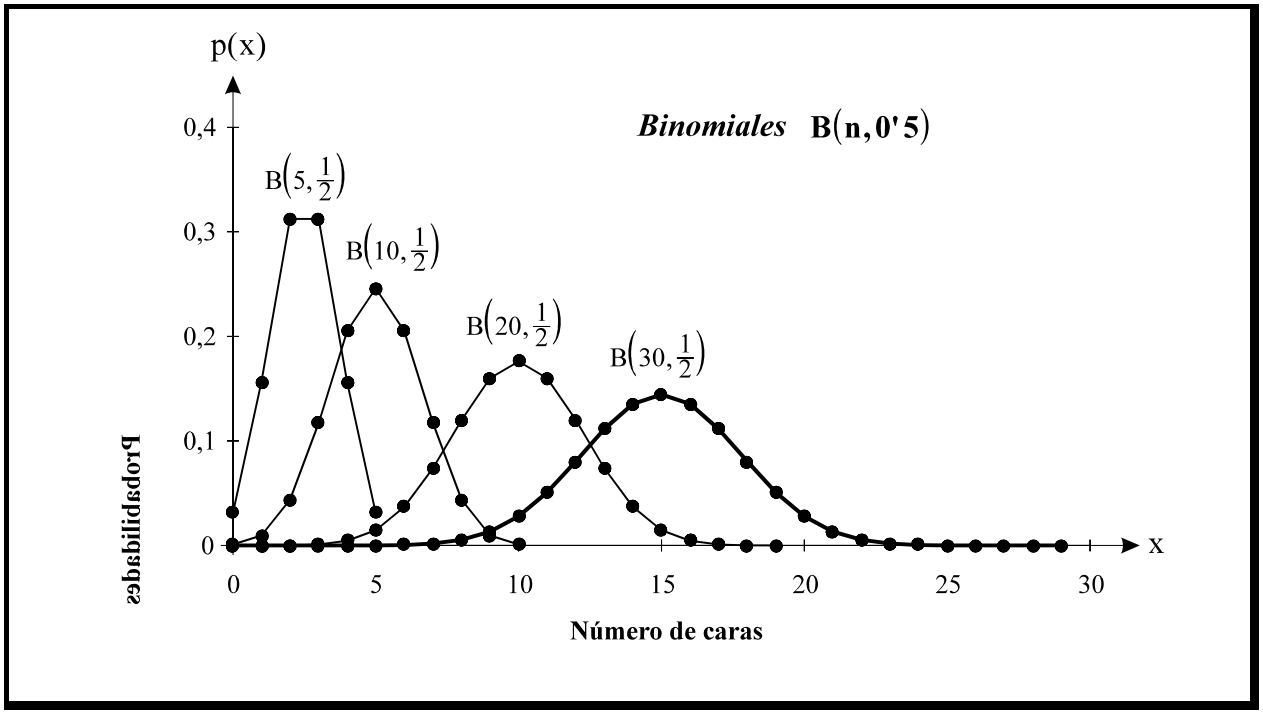
La Distribución Normal como aproximación de la Binomial

El uso práctico de la distribución binomial suele necesitar la realización de cálculos y operaciones en buena medida engorrosos y desde luego, en absoluto triviales. Si queremos calcular probabilidades en una $B(n,p)$ con n grande, las operaciones pueden ser muy laboriosas. Por ejemplo, para una $B(200,0'3)$ el cálculo de $p(X \geq 70)$ supone obtener $\sum_{k=70}^{200} \binom{200}{k} \cdot 0'3^{k} \cdot 0'7^{200-k}$, es decir, sumar 131 sumandos tremendos. La tarea es prácticamente imposible.

Estos inconvenientes pueden obviarse estudiando la evolución de las distribuciones binomiales $B(n,p)$ cuando el parámetro n se hace muy grande.

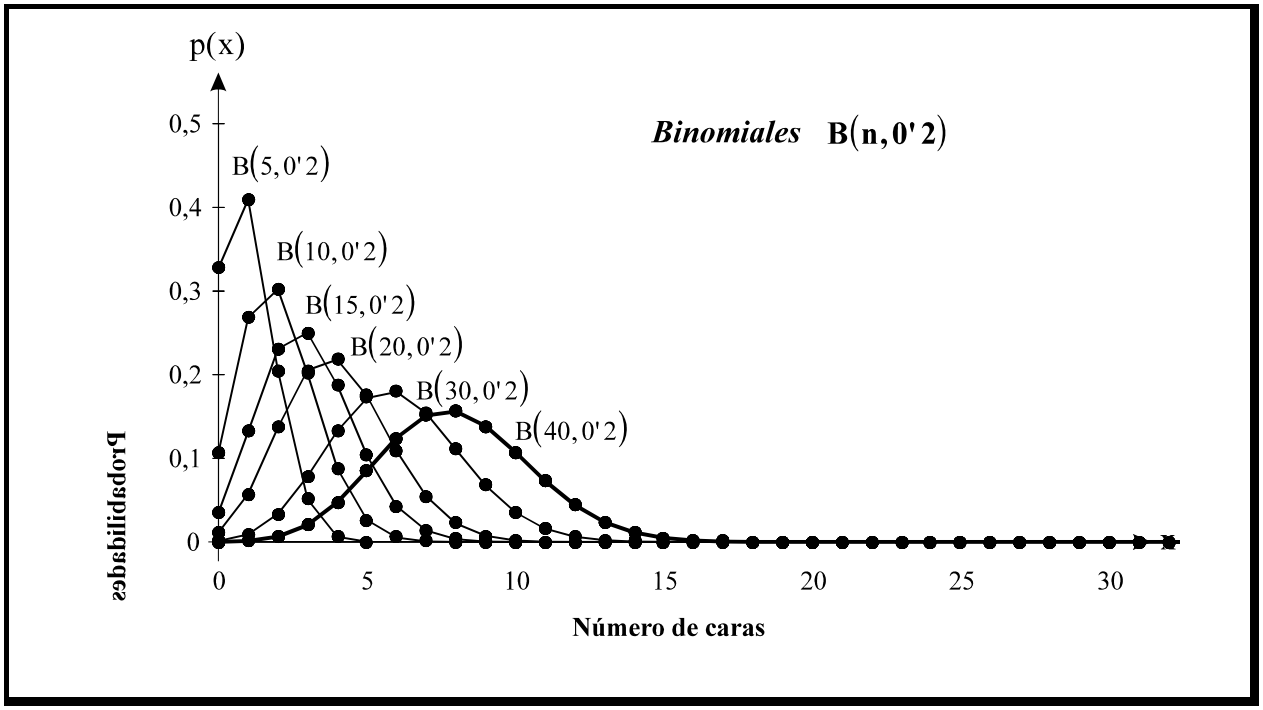
Si recordamos la distribución de las canicas en las casillas del aparato de Galton, es indudable que imitan perfectamente a la curva normal. Pero el aparato de Galton es una distribución binomial $B(n,0'5)$, es decir, una distribución binomial en la que $p = 0'5$.

Otra distribución de este tipo es el ejemplo siguiente. Si lanzamos una moneda 5 veces, 10 veces, 20 veces, 30 veces, etc. y analizamos la variable aleatoria que da el número de caras observamos que sigue una distribución binomial.



Es evidente que, a medida que crece el número de pruebas realizadas, el polígono de probabilidad tiende a la curva de la distribución normal.

¿Ocurrirá lo mismo con otras distribuciones $B(n,p)$ con $p \neq 0'5$? Observemos lo que ocurre con $p = 0'2$.



Si observamos bien estas distribuciones binomiales, la primera se parece muy poco a una curva normal. Poco a poco se van pareciendo más y más. La última, una binomial $B(40,0'2)$, es casi exactamente una curva normal.

¿Cuáles son los parámetros de esta distribución normal?

De Moivre demostró que siempre que se verifique $\mathbf{np} \geq \mathbf{5}$ y $\mathbf{nq} \geq \mathbf{5}$ la distribución binomial $B(n,p)$ se puede aproximar mediante la distribución normal $N\left(np, \sqrt{npq}\right)$ sin pérdida de exactitud significativa. Obsérvese que cuanto mayor sea el valor de n y cuanto más próximo sea p a 0'5, tanto mejor será la aproximación realizada.

¡OJO! ***La distribución binomial es de variable aleatoria discreta y, por tanto, tiene sentido calcular probabilidades puntuales; por ejemplo,*** $\mathbf{p(X=2)}$***, en cambio la distribución normal es de variable aleatoria continua y, por tanto, no tiene sentido calcular probabilidades puntuales, pues son todas nulas.***

¿Cómo proceder entonces para calcular la probabilidad en la distribución binomial cuando se aproxima por la normal? Basta considerar los valores de la variable aleatoria discreta como marcas de clase de intervalos del siguiente modo:

$$p(X=2) = p(1'5 \leq X' \leq 2'5)$$
$$p(X \leq 4) = p(X' \leq 4'5)$$
$$p(X < 4) = p(X' \leq 3'5)$$

Se toman intervalos con medio punto a izquierda y derecha para que la probabilidad total sea 1, que corresponde al área bajo la curva de la función de densidad. En caso contrario quedarían áreas sin sumar y la suma total no daría 1.

En general, si en una distribución binomial $B(n,p)$ tanto $n\cdot p$ como $n\cdot q$ son mayores o iguales que 5, tenemos:

$$X \text{ es } B(n,p) \;\rightarrow\; X' \text{ es } N\left(np, \sqrt{npq}\right)$$
$$p[X=k] = p[k-0'5 \le X' \le k+0'5]$$

Ejemplo: **Se ha encuestado a la población masculina de cierto municipio, encontrándose que un 34% son hinchas del equipo de fútbol local. Elegidos 50 ciudadanos al azar:**

a) ¿Cuál será la probabilidad de que haya exactamente 18 aficionados?

b) ¿Cuál será la probabilidad de que haya más de 20 aficionados? ¿Y de haya 25 o menos? ¿Y de que haya menos de 40?

a) La probabilidad de ser hincha del equipo local es $p = 0'34$. Se trata, pues, de una binomial $B(50, 0'34)$.

Puesto que $50\cdot 0'34 = 17 \ge 5$ y $50\cdot 0'66 = 33 \ge 5$ podemos aproximar la distribución binomial por una normal, de tal manera que:

$$X \text{ es } B(50,0'34) \;\rightarrow\; X' \text{ es } N(17,3'34) \quad \text{ya que} \begin{cases} \mu = 50\cdot 0'34 = 17 \\ \sigma = \sqrt{50\cdot 0'34\cdot 0'66} = 3'34 \end{cases}$$

$$p(X=18) = p(17'5 \le X' \le 18'5) = p\left(\frac{17'5-17}{3'34} \le Z \le \frac{18'5-17}{3'34}\right) =$$

$$p(0'14 \le Z \le 0'44) = p(Z \le 0'44) - p(Z \le 0'14) = 0'6700 - 0'5557 = 0'1143$$

b) $$p(X>20) = p(X' \ge 20'5) = p\left(Z \ge \frac{20'5-17}{3'34}\right) = p(Z \ge 1'04) = 1 - p(Z \le 1'04) =$$

$$1 - 0'8508 = 0'1492$$

$$p(X \le 25) = p(X' \le 25'5) = p\left(Z \le \frac{25'5-17}{3'34}\right) = p(Z \le 2'54) = 0'9945$$

$$p(X<40) = p(X' \le 39'5) = p\left(Z \le \frac{39'5-17}{3'34}\right) = p(Z \le 6'73) \cong 1$$

Aproximación de una distribución empírica por una normal

Consideremos el siguiente ejemplo:

Se ha estudiado la distribución de una variable antropométrica, que representaremos por x, de 50 hombres normales comprendidos entre los 20 y los 30 años, obteniéndose los resultados que muestra la tabla de la derecha.

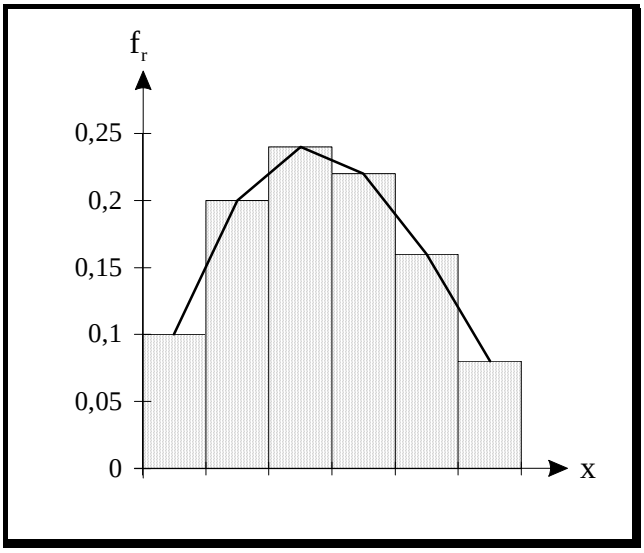
x	f_i	f_r
$[99'5,109'5)$	5	0'1
$[109'5,119'5)$	10	0'2
$[119'5,129'5)$	12	0'24
$[129'5,139'5)$	11	0'22
$[139'5,149'5)$	8	0'16
$[149'5,159'5)$	4	0'08
	50	1

a) Hallar la probabilidad de que un individuo elegido al azar tenga su medida antropométrica menor o igual a 100 cm.

b) Hallar la probabilidad de que mida más de 150 cm.

Este es un caso típico de ajuste de una distribución empírica mediante una teórica. Procedamos de forma ordenada:

1º Se representa el histograma y el polígono de frecuencias relativas de la distribución empírica dada. Si adopta una forma que nos recuerda a la curva de probabilidad de la distribución normal, haremos uso de este modelo teórico. En caso contrario, no tiene sentido el ajuste mediante una distribución normal.



2º Calculamos la media aritmética $\overline{x}$ y la desviación típica s de la distribución empírica.

$$\overline{x} = 128'30 \quad y \quad s = 14'2674$$

3º Ajustamos la distribución empírica mediante una distribución normal que tenga por parámetros $\overline{x}$ y s, es decir, $N(\overline{x},s)$, ya que está demostrados que la distribución teórica que mejor se aproxima a una empírica es aquella que tiene la misma media y la misma desviación típica.

Por tanto la aproximación correcta es mediante la distribución $N(128'3,14'26)$.

4º A partir del modelo teórico hacemos los cálculos:

a) $$p(X \le 100) = p\left(Z \le \frac{100 - 128'3}{14'2}\right) = p(Z \le -1'99) = p(Z \ge 1'99) = 1 - p(Z \le 1'99) = 0'024$$

b) $p(X>150)=p\left(Z>\dfrac{150-128'3}{14'2}\right)=p(Z>1'52)=1-p(Z<1'52)=1-0'9357=0'0643$

5º Por último, es conveniente realizar un estudio sobre si el ajuste realizado es bueno o no. Este proceso se llama ***estudio de la bondad del ajuste***.

Test de normalidad

Hay ocasiones en las que empíricamente se han obtenido unos datos:

$$x_1, x_2, x_3, \ldots\ldots\ldots, x_n \qquad \text{(Muestra)}$$

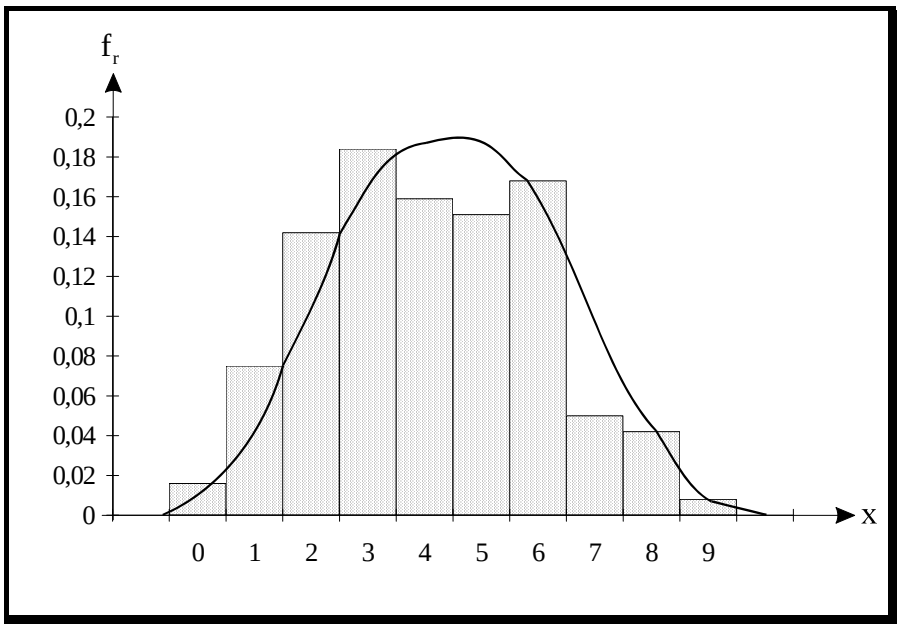
cuya distribución tiene un cierto parecido con la curva normal. Y nos preguntamos si la población de partida será normal.

Esta pregunta, que en muchos casos es de gran importancia, no puede tener una respuesta absoluta (sí o no), pero podemos aproximarnos a ella en términos de probabilidad (es "así" de probable que sea cierto). Esto es lo que se consigue con los test de normalidad. Vamos a estudiar dos de ellos.

Prueba de Kolmogorov. Estudio de un ejemplo

Vamos a suponer que contamos el número de veces que aparece la letra "o" en cada uno de los 119 renglones de un artículo periodístico. Los resultados son:

x_i	f_i	f_r
0	2	0'016
1	9	0'075
2	17	0'142
3	22	0'184
4	19	0'159
5	18	0'151
6	20	0'168
7	6	0'050
8	5	0'042
9	1	0'008
	119	1



La distribución es sólo ligeramente similar a una curva normal. Las diferencias que se aprecian pueden ser debidas al azar, o quizá no. Para decidir científicamente si es lo uno o lo otro, procederemos del siguiente modo:

Empezaremos calculando la media, $\overline{x}$, y la desviación típica, σ, de los datos que se tienen. En este caso: $\overline{x} = 4'08$ y $\sigma = 1'96$

A continuación comparamos nuestra distribución empírica con la distribución $N(\overline{x}, \sigma)$ es decir $N(4'08, 1'96)$.

En lugar de comparar frecuencias relativas por intervalos se comparan las frecuencias relativas acumuladas: ¿qué porcentaje de renglones hay con 3 oes o menos?, ¿cuántos renglones con 4 oes o menos?, etc.

Por ejemplo, hay 50 renglones con 3 oes o menos, lo que supone un porcentaje de $50 \cdot \frac{110}{119} = 42'02\%$.

Sobre la curva normal se tiene:

$$p(X \le 3) = p(X' \le 3'5) = p\left(Z \le \frac{3'5 - 4'08}{1'96}\right) = p(Z \le -0'30) = p(Z \ge 0'30) = 1 - p(Z \le 0'30) =$$

$$1 - 0'6179 = 0'3821. \text{ Es el } 38'21\%$$

Así procederemos para cada valor x_i.De este modo obtendremos en nuestro caso, dos listas de 10 números cada una:

(1) l_k	(2) F_i	(3) F_{r_k}	(4) P_k	(5) D_k
0'5	2	1'68	3'44	1'76
1'5	11	9'24	9'51	0'27
2'5	28	23'53	21'19	2'34
3'5	50	42'02	38'21	3'81
4'5	69	57'98	58'32	0'34
5'5	87	73'11	76'42	3'31
6'5	107	89'92	89'07	0'85
7'5	113	94'96	95'91	0'95
8'5	118	99'16	98'78	0'38
9'5	119	100	99'71	0'28

(1) Límite superior de cada intervalo.

(2) Frecuencias acumuladas.

(3) Frecuencias relativas acumuladas en la muestra hasta l_k, expresadas en tanto por ciento.

(4) Probabilidad acumulada hasta l_k en la teórica población normal, expresada en tantos por ciento.

(5) Diferencias entre las dos columnas anteriores, tomadas en valor absoluto.

Restando los pares de números correlativos de las dos últimas columnas, obtendremos 10 diferencias. La mayor de ellas en valor absoluto, $D_{máx}$, representa la máxima diferencia entre *lo que*

se tiene y lo que se habría obtenido si los resultados fueran **exactamente** *normales*. (En nuestro caso se obtiene $D_{máx} = 3'81\%$ correspondiente a los porcentajes acumulados hasta 3'5).

Pues bien, está demostrado que si $D_{máx}$ es menor que $\frac{136}{\sqrt{n}}$ se puede admitir que las diferencias son debidas al azar y, por tanto, que los datos obtenidos provienen de una distribución normal.

La afirmación se hace con un margen de error del 5%, que quiere decir que, por término medio, de cada 100 veces que apliquemos este método nos equivocaremos cinco. En nuestro caso, $\frac{136}{\sqrt{119}} = 12'46$ es el máximo error que se le puede atribuir al azar. Como el valor obtenido para $D_{máx}$ es 3'81, consideraremos que la *variable número de veces que aparece la letra "o" en cada renglón*, se distribuye normalmente.

Prueba de Kolmogorov: exposición general

¿Es razonable suponer que los valores $x_1, x_2, x_3, ...x_n$ han sido extraídos de una población normal? Para responder a esta pregunta se compara la distribución de estos valores con una curva normal cuyos parámetros, $\overline{x}$ y σ coincidan con los de la muestra.

La comparación se efectúa contrastando:

- La proporción en % de valores de la muestra menores que l_k.
- Con el porcentaje de área bajo la curva normal que hay desde $-\infty$ a l_k.

para todos los l_k, límites superiores de los intervalos en los que hemos clasificado los elementos de la muestra.

De este modo, para cada valor de K hay una diferencia

$$D_k = |\%\text{ hasta } l_k \text{ en la muestra} - \%\text{ hasta } l_k \text{ en la curva normal}|$$

La mayor de las diferencias obtenida, $D_{máx}$, se compara con el número que se obtiene al efectuar el cociente $\frac{136}{\sqrt{n}}$, siendo n el tamaño de la muestra.

Si $D_{máx} < \frac{136}{\sqrt{n}}$ entonces no se rechaza la hipótesis de que la muestra está extraída de una población normal.

Si $D_{máx} > \frac{136}{\sqrt{n}}$ se rechaza la hipótesis. Ésta se haría con un riesgo de error del 5%, es decir, de cada 100 veces que rechacemos la hipótesis en estas circunstancias, nos exponemos a equivocarnos en 5 de ellas.

La fórmula $\frac{136}{\sqrt{n}}$ es válida sólo para muestras de 20 o más individuos, y debe interpretarse como la máxima diferencia que se puede atribuir al azar si extraemos una muestra de n elementos de una población normal.

Ejemplo: La siguiente distribución corresponde a las estaturas en cm., de 1400 mujeres:

141	**146**	**151**	**156**	**161**	**166**	**171**	**176**	**181**
2	**25**	**146**	**327**	**424**	**314**	**128**	**29**	**5**

a) Calcula su media $\bar{x}$, su desviación típica, las frecuencias acumuladas y obtén, dividiendo las anteriores por 14, los porcentajes acumulados.

b) Pon el extremo superior de los intervalos de clase y tipifica dichos valores respecto a los valores $\bar{x}$ y σ.

c) Calcula en % la probabilidad acumulada hasta cada uno de los extremos de los intervalos. Compara restándolos, los porcentajes acumulados en la distribución empírica y en la teórica.

d) Selecciona la mayor de las diferencias y prueba si supera o no el número $\frac{136}{\sqrt{1400}} = 3'63$. Decide, en consecuencia, si se rechaza o no la hipótesis de normalidad de la distribución de partida.

x_i	f_i	F_i	$\%\, a_i$	l_i	z_i	P_i	D_i
141	2	2	0'14	143'5	–2'69	0'36	0'22
146	25	27	1'93	148'5	–1'92	2'74	0'81
151	146	173	12'36	153'5	–1'15	12'51	0'15
156	327	500	35'71	158'5	–0'38	35'20	0'51
161	424	924	66	163'5	0'40	65'54	0'46
166	314	1238	88'43	168'5	1'17	87'90	0'53
171	128	1366	97'57	173'5	1'95	97'44	0'13
176	29	1395	99'64	178'5	2'72	99'67	0'03
181	5	1400	100	183'5	3'49	99'98	0'02

$$\bar{x} = 160'9 \qquad \sigma = 6'46$$

La mayor diferencia, 0'81 % es muy inferior a $\frac{136}{\sqrt{1400}} = 3'63$. Por tanto, no se rechaza la hipótesis de normalidad.