Bioestadística – Grado en Biología
Curso 2022-2023

Tema 3. Variables aleatorias

Profesores: Mª Ángeles Casares de Cal, Fernando Castro Prado, Laura Davila Pena y Pedro Faraldo Roca

Índice

1 Variables aleatorias

En el tema de Estadística descriptiva hemos estudiado variables, entendiéndolas como mediciones que se efectúan sobre los individuos de una muestra. Así, la Estadística descriptiva nos permitía analizar los distintos valores que tomaban las variables sobre una muestra ya observada. Se trataba, pues, de un estudio posterior a la realización del experimento aleatorio.

En este tema trataremos las variables situándonos antes de la realización del experimento aleatorio. En cualquier caso, muchos desarrollos serán análogos a los del tema de Estadística descriptiva.

1.1 Concepto de variable aleatoria

Definición 1
*Llamamos **variable aleatoria** a una aplicación que a cada resultado de un experimento aleatorio le asigna un número real, obtenido por la medición de cierta característica. Lo podríamos escribir así:*

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longrightarrow & X(\omega) \end{array}$$

donde Ω es un conjunto en el que situamos todos los resultados posibles del experimento aleatorio. Realmente no vamos a prestar mucha atención al conjunto Ω, sino que lo que nos importa son los valores de la variable que se está midiendo y sus probabilidades asociadas.

Denotamos la variable aleatoria por una letra mayúscula, en este caso X. Los valores que pueda tomar serán denotados por letras minúsculas.

Ejemplo 1
Proponemos dos ejemplos de experimentos aleatorios clásicos (dados y monedas) sobre los que medimos características muy comunes:

(a) Lanzamos dos monedas y contamos el número de caras.

(b) Lanzamos dos dados y sumamos sus puntuaciones.

Solución. Vamos a obtener los valores posibles de la variable aleatoria en cada apartado, y sus probabilidades.

En el apartado (a) la variable aleatoria X ="número de caras" toma los siguientes valores, con las probabilidades que figuran debajo de cada uno de ellos:

x_i	0	1	2
$p_i = \mathbf{P}(X = x_i)$	0,25	0,5	0,25

En el apartado (b) la variable aleatoria X ="suma de puntuaciones" toma los valores que figuran en la tabla, con las probabilidades que se indican:

x_i	2	3	4	5	6	7	8	9	10	11	12
$p_i = \mathbf{P}(X = x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

□

Los ejemplos clásicos, de monedas o dados, o más en general de los juegos de azar, han sido fuente de inspiración para el desarrollo del Cálculo de Probabilidades, que es una disciplina que aporta la herramienta de cálculo y los modelos para el estudio teórico de los experimentos aleatorios. Así, concebimos como experimento aleatorio todo aquello cuyo resultado futuro es incierto, pero cuyos valores posibles conocemos y les podemos atribuir unas probabilidades. Y eso no sólo es posible con dados y monedas, sino que, por ejemplo, un error de medición en laboratorio también lo concebimos como aleatorio y podemos pensar en cuáles son sus posibles valores, sus probabilidades, su media y su dispersión.

Podríamos pensar, por ejemplo, que estamos midiendo la masa de un objeto y la balanza presenta un error de medida de media cero y con ciertas probabilidades que conocemos o que nos indica el fabricante, de manera que nos garantizan que sería poco probable un error superior a un gramo tanto por exceso como por defecto.

Las variables del Ejemplo 1 eran discretas, mientras que el error de medida sería una variable continua. Y es que de igual modo a lo dicho en el tema de Estadística descriptiva, podemos clasificar las variables aleatorias en **discretas** y **continuas** en función del conjunto de valores que pueden tomar. Así, será discreta si dichos valores se encuentran separados entre sí. Por tanto será representable por conjuntos discretos, como $\mathbb{Z}$ o $\mathbb{N}$. Será continua cuando el conjunto de valores que puede tomar es un intervalo.

2 Variable aleatoria discreta

Como ya hemos dicho, la variable aleatoria discreta toma un conjunto de valores separados entre sí y, por tanto, enumerables. Podemos entonces considerar la sucesión ordenada de valores $x_1 < x_2 < \ldots < x_k < \ldots$. La siguiente tarea consiste en asignar probabilidades a esos valores. Esa asignación la efectúa lo que denominaremos como función de probabilidad o función de masa.

También presentaremos la función de distribución de una variable aleatoria, y que nos permite calcular probabilidades acumuladas.

2.1 Función de masa o de probabilidad y función de distribución

Definición 2
*Se define la **función de masa de probabilidad** de una variable aleatoria discreta, X, como la función que asigna a cada valor su probabilidad, $p_i = \mathsf{P}(X = x_i)$.*

Los valores de la función de probabilidades o función de masa se suelen escribir en una tabla de probabilidades con este formato:

x_i	x_1	x_2	$\ldots$	x_k	$\ldots$
$p_i = \mathsf{P}(X = x_i)$	p_1	p_2	$\ldots$	p_k	$\ldots$

Debemos tener en cuenta que si sumamos estas probabilidades sobre todos los posibles valores de X, la suma deberá ser 1: $\sum_i p_i = 1$

La función de probabilidades se puede representar mediante un diagrama de barras, de manera análoga a la representación de las frecuencias que veíamos en el tema de Estadística descriptiva. Para el Ejemplo 1(a) el diagrama de barras asociado a la función de masa sería:

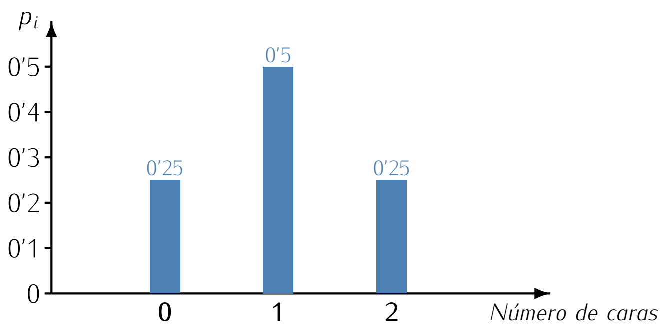


Figura 1. Función de masa del número de caras al lanzar dos monedas.

Observemos que, conociendo la función de masa de una variable aleatoria discreta, se puede determinar la probabilidad de cualquier suceso relativo a la variable. Por ejemplo, la probabilidad de que salga alguna cara al lanzar dos monedas, se puede identificar como la probabilidad de que la variable $X =$"número de caras" valga 1 ó 2. Sumando las probabilidades correspondientes resulta:

$\mathsf{P}(\text{"que salga alguna cara"}) = \mathsf{P}(X \in \{1, 2\}) = \mathsf{P}(X = 1) + \mathsf{P}(X = 2) = 0'5 + 0'25 = 0'75$

Ejemplo 2
Lanzamos dos dados y se considera la variable aleatoria que suma las puntuaciones. Calcula la probabilidad de que dicha variable aleatoria valga a lo sumo 4 (la suma de las puntuaciones sea a lo sumo 4).

Solución. Sea Y la variable aleatoria "suma de las puntuaciones en el lanzamiento de dos dados".

Consideramos el suceso:

$A = \{Y \leq 4\}$

entonces:

$$\mathbf{P}(A) = \mathbf{P}(Y \leq 4) = \mathbf{P}(Y \in \{2,3,4\}) = \mathbf{P}(Y = 2) + \mathbf{P}(Y = 3) + \mathbf{P}(Y = 4) =$$
$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{1}{6}$$

□

Igual que en Estadística descriptiva se definían las frecuencias acumuladas y se representaban mediante el diagrama escalonado, ahora se puede construir un diagrama escalonado con las probabilidades acumuladas. Lo llamaremos función de distribución y en el Ejemplo 1(a) adoptaría esta forma:

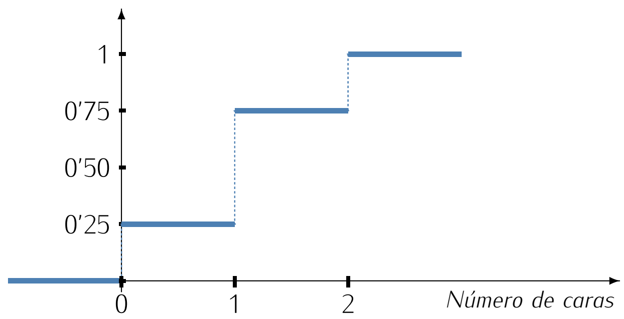


Figura 2. Función de distribución del número de caras al lanzar dos monedas.

La acumulación de probabilidades es una idea que se podrá aplicar a cualquier variable aleatoria, discreta o continua, y por eso definimos la función de distribución de la siguiente manera.

Definición 3
*La **función de distribución** de una variable aleatoria se define como:*

$$\begin{array}{rccl} F: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x_0 & \longrightarrow & F(x_0) = \mathbf{P}(X \leq x_0) \end{array}$$

Como ya hemos visto en el Ejemplo 1(a), para el caso de una variable discreta tenemos que
$F(x_0) = \mathbf{P}(X \leq x_0) = \sum_{x_i \leq x_0} p_i$
y esto se reduce a sumar las probabilidades de los valores menores o iguales que x_0, y juega un papel análogo a las frecuencias relativas acumuladas.
Además, su representación da lugar a un diagrama escalonado con saltos en los puntos x_i, siendo la amplitud de cada salto $p_i = \mathbf{P}(X = x_i)$.

Propiedades. Cualquier función de distribución, de una variable discreta o continua, verifica:

1. $0 \leq F(x) \leq 1 \qquad \forall x \in \mathbb{R}$
2. $\lim_{x \to -\infty} F(x) = 0$

3. $\lim_{x\to+\infty} F(x) = 1$

4. F es creciente
$$x_1 \leq x_2 \Longrightarrow F(x_1) \leq F(x_2)$$

5. F es continua por la derecha
$$\lim_{h>0, h\to 0} F(x+h) = F(x) \qquad \forall x \in \mathbb{R}$$

6. $\mathbf{P}(a < X \leq b) = F(b) - F(a) \qquad \forall a, b \in \mathbb{R},\ a < b$

2.2 Medidas de posición y dispersión calculadas sobre una variable aleatoria discreta

Al igual que en el tema de Estadística descriptiva definíamos medidas sobre las variables observadas, como la media y la varianza, a partir de los valores y sus frecuencias, ahora haremos lo mismo con las variables aleatorias sustituyendo las frecuencias relativas por las probabilidades. Las interpretaciones de estas medidas como indicadores de la posición o dispersión de la distribución son las mismas que para el tema de Estadística descriptiva.

Empezaremos dando las expresiones que permiten calcular todas estas medidas en el caso discreto.

Media o Esperanza. La denotamos por $E(X)$ o por μ. Para variables aleatorias se emplea el término esperanza, que no se usa en Descriptiva, dando idea de que la variable, por ser aleatoria, ocurrirá en un futuro, y si hay que aventurar un valor "esperado", ése sería la media (o esperanza). Para variables aleatorias discretas se calcula así:

$$E(X) = \mu = \sum_i x_i \mathbf{P}[X = x_i] = \sum_i x_i p_i$$

Igual que en Descriptiva, la media de variables aleatorias se ve afectada por cambios de localización y escala, esto es, $E(a + bX) = a + bE(X)$ siendo a y b constantes.

Mediana. La mediana de X es el menor valor "m" de la variable aleatoria que deja a su izquierda al menos una probabilidad 0'5, es decir, $P[X \leq m] \geq 0'5$.
Una vez definida la función de distribución, que denotamos por F, la mediana sería la abscisa en la cual F alcanza el valor 0,5. Por tanto,

$$m = F^{-1}(0'5)$$

En el caso discreto podemos tener dificultades con los saltos, que resolveremos igual que en el tema 1 (Estadística descriptiva), tomando el primer salto que supera el valor 0'5.

Moda. Es el valor más probable en una variable discreta.

Cuantiles. Sea $p \in (0, 1)$. El cuantil p de una variable aleatoria (que denotaremos por Q_p) será el menor valor de la variable aleatoria que deja al menos una probabilidad p a su izquierda. En consecuencia,
$$\mathrm{Q}_p = F^{-1}(p)$$

De nuevo en el caso discreto sería el primer salto en el que se supera la probabilidad p.

Varianza. La denotamos por $Var(X)$ o por σ^2. Es la esperanza de las desviaciones cuadráticas respecto de la media. Para variables discretas se calcula así:

$$\mathrm{Var}(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_i (x_i - \mu)^2\, \mathbf{P}[X = x_i] = \sum_i (x_i - \mu)^2\, p_i$$

Los cambios de localización o escala producen los mismos efectos sobre la varianza, que ya hemos observado en Descriptiva. En concreto: $\mathrm{Var}\,(a + bX) = b^2\mathrm{Var}(X)$ siendo a y b constantes.

Desviación típica. La definimos como la raíz cuadrada de la varianza y, por tanto, la denotamos por σ.

Ejemplo 3
Calcula la media, varianza y desviación típica de la variable X ="número de caras al lanzar dos monedas".
Solución.
$\mu = E(X) = \sum_{x=0}^{2} x \cdot \mathbf{P}[X = x] = 0 \cdot \mathbf{P}[X = 0] + 1 \cdot \mathbf{P}[X = 1] + 2 \cdot \mathbf{P}[X = 2] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = \frac{4}{4} = 1$

$\sigma^2 = Var(X) = E[(X-E(X))^2] = \sum_{x=0}^{2}(x-\mu)^2 \cdot \mathbf{P}[X = x] = (0-1)^2 \cdot \mathbf{P}[X = 0] + (1-1)^2 \cdot \mathbf{P}[X = 1] + (2-1)^2 \cdot \mathbf{P}[X = 2] = 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = \frac{2}{4} = 0'5$

Por último la desviación típica resulta de efectuar la raíz cuadrada de la varianza:
$\sigma = D.t.(X) = \sqrt{0'5} = 0'707$

□

Ejemplo 4
Sea X una variable aleatoria discreta cuya media vale 7 y cuya varianza vale 5. Calcula la media y la varianza de la variable $Y = 2 + 3X$.
Solución. Sabemos que:
$E(a + bX) = a + bE(X)$
$Var(a + bX) = b^2 Var(X)$
por lo tanto:
$E(Y) = E(2 + 3X) = 2 + 3 \cdot E(X) = 2 + 3 \cdot 7 = 23$
$Var(Y) = Var(2 + 3X) = 3^2 Var(X) = 9 \cdot 5 = 45$

□

Ejemplo 5
Los animales atendidos en una clínica veterinaria pueden requerir el internamiento en dicho centro durante 0, 1, 2, 3 ó 4 días, con probabilidades respectivas 0'4, 0'3, 0'05, 0'2 y 0'05.

El precio que cobra la clínica por la atención de un animal es de $300 + cX$ euros, donde c es una cantidad constante y X representa el número de días de estancia en la clínica.

Determinar el valor de c de manera que el precio medio cobrado por la clínica por la atención de un animal sea de 540 euros.

Solución. En el enunciado se pide el valor de c tal que $E(300 + cX) = 540$.

Ahora bien, por las propiedades de la esperanza, sabemos que $E(300 + cX) = 300 + cE(X) = 540$. Por tanto, basta con calcular $E(X)$, sustituir su valor en la ecuación y despejar c.

A partir de sus valores y probabilidades, calculamos
$E(X) = 0 \cdot 0'4 + 1 \cdot 0'3 + 2 \cdot 0'05 + 3 \cdot 0'2 + 4 \cdot 0'05 = 1'2$ días.

Sustituyendo en la ecuación, tenemos $300 + c \cdot 1'2 = 540$, de lo cual resulta
$c = (540 - 300)/1'2 = 200$ euros/día.

□

2.3 Variable aleatoria multidimensional discreta

En Descriptiva estudiamos una variable multidimensional desde el punto de vista descriptivo, esto es, analizando unos datos observados sobre dos o más variables simultáneamente. En esta sección extenderemos los conceptos que surgían allí, como la distribución conjunta, las distribuciones marginales y las distribuciones condicionadas, al caso en que el vector no haya sido observado sino que presente ciertos valores con ciertas probabilidades.

Definición 4
*Una **variable aleatoria multidimensional** o **vector aleatorio** es una colección de variables aleatorias*

$(X_1, \ldots, X_k)$ medidas simultáneamente sobre el mismo individuo o sobre el mismo resultado de un experimento aleatorio.

Formalmente, se puede representar como una aplicación:

$$X = (X_1, \ldots, X_k) : \Omega \longrightarrow \mathbb{R}^k$$

donde Ω representa los resultados del experimento aleatorio y se calculan k variables en cada realización del experimento, que por tanto forman un vector en $\mathbb{R}^k$.

Por supuesto, cada una de las componentes del vector aleatorio, $X_1, \ldots, X_k$, es una variable aleatoria, y por tanto se puede calcular su media, su varianza y su distribución, que hemos estudiado en las secciones anteriores. Como las k variables se miden sobre el mismo experimento aleatorio, podemos estudiar también las relaciones entre las variables.

Igual que hicimos en Descriptiva, cuando hay dos variables, $k = 2$, las denotaremos como (X, Y) y hablaremos de una variable bidimensional. En el caso en que ambas sean **discretas**, si $x_1 < \cdots < x_I$ son los valores posibles de X e $y_1 < \cdots y_J$ son los valores posibles de Y, podremos definir la **distribución de probabilidad conjunta** como

$$p_{ij} = \mathbf{P}(X = x_i, Y = y_j) \qquad i \in \{1, \ldots, I\} \quad j \in \{1, \ldots, J\}$$

y formar una tabla con estas probabilidades

	Y				
X	y_1	y_2	$\cdots$	y_J	
x_1	p_{11}	p_{12}	$\cdots$	p_{1J}	$p_{1\bullet}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2J}	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	$p_{I\bullet}$
	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet J}$	1

La última columna contiene las sumas de cada fila, y forma la **distribución marginal** de X, que de hecho es la distribución de probabilidad de la variable X si se considerara por separado.

$$p_{i\bullet} = \mathbf{P}(X = x_i) = \sum_{j=1}^{J} p_{ij} \qquad i \in \{1, \ldots, I\}$$

De igual modo, la última fila de la tabla contiene las sumas de cada columna, y constituye la distribución marginal de Y:

$$p_{\bullet j} = \mathbf{P}(Y = y_j) = \sum_{i=1}^{I} p_{ij} \qquad j \in \{1, \ldots, J\}$$

Las distribuciones condicionadas se definen de manera similar a como se hizo en Descriptiva,

$Y \mid X = x_i$	y_1	y_2	$\cdots$	y_J	
	$p_{1\vert i}$	$p_{2\vert i}$	$\cdots$	$p_{J\vert i}$	1

donde

$$p_{j|i} = \mathbf{P}(Y = y_j \mid X = x_i) = \frac{\mathbf{P}(X = x_i, Y = y_j)}{\mathbf{P}(X = x_i)} = \frac{p_{ij}}{p_{i\bullet}} \qquad j \in \{1, \ldots, J\}$$

Ejemplo 6

Una moneda se lanza tres veces. Se consideran las variables aleatorias X ="número de caras en los dos primeros lanzamientos" y Y ="número de caras en los tres lanzamientos". Vamos a calcular la distribución de probabilidad conjunta, las distribuciones marginales y las distribuciones condicionadas.

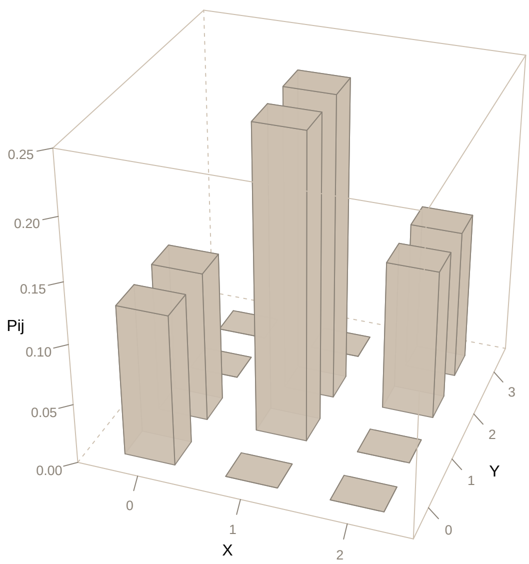
Solución.

El espacio muestral original asociado al experimento aleatorio de lanzar tres veces la moneda tendría la forma $\Omega = \{ccc, cc+, c+c, +cc, c++, +c+, ++c, +++\}$ y estos ocho resultados posibles son equiprobables.

El vector aleatorio (X, Y) toma el valor $(2, 3)$ si hubo tres caras (resultado ccc), el valor $(2, 2)$ si salió cc+, y así sucesivamente podríamos ir obteniendo los valores posibles del vector aleatorio (X, Y). Juntando los valores posibles y sus probabilidades, llegamos a la siguiente tabla con la distribución conjunta de (X, Y) y las distribuciones marginales de X y de Y.

	Y				
X	0	1	2	3	
0	1/8	1/8	0	0	1/4
1	0	1/4	1/4	0	1/2
2	0	0	1/8	1/8	1/4
	1/8	3/8	3/8	1/8	1

Diagrama de barras de (X,Y)



Las distribuciones condicionadas de Y a cada valor de X se obtienen dividiendo cada fila por el total de fila, con el resultado siguiente:

$Y \mid X = 0$	0	1	2	3	
$P(Y = y_j \mid X = 0)$	1/2	1/2	0	0	1

$Y \mid X = 1$	0	1	2	3	
$P(Y = y_j \mid X = 1)$	0	1/2	1/2	0	1

$Y \mid X = 2$	0	1	2	3	
$P(Y = y_j \mid X = 2)$	0	0	1/2	1/2	1

De manera similar se obtienen las distribuciones condicionadas de X a cada valor de Y, que figuran a continuación:

$X \mid Y = 0$	0	1	2	
$P(X = x_i \mid Y = 0)$	1	0	0	1

$X \mid Y = 1$	0	1	2	
$P(X = x_i \mid Y = 1)$	1/3	2/3	0	1

$X \mid Y = 2$	0	1	2	
$P(X = x_i \mid Y = 2)$	0	2/3	1/3	1

$X \mid Y = 3$	0	1	2	
$P(X = x_i \mid Y = 3)$	0	0	0	1

□

Definición 5 (Variables aleatorias independientes)
*Las distribuciones condicionadas permiten constatar la relación que hay entre dos variables aleatorias. Pues bien, cuando dos variables no mantienen ningún tipo de relación diremos que son **independientes**. Si las dos variables X e Y de una variable bidimensional (X, Y) son discretas, entonces su independencia consiste en que las distribuciones condicionadas (por ejemplo, de Y sobre X) coincidan con la distribución marginal (de Y).*

Como definición formal, se define la independencia de X e Y así:

$$\mathbf{P}(X = x_i, Y = y_j) = \mathbf{P}(X = x_i) \cdot \mathbf{P}(Y = y_j) \qquad i \in \{1, \ldots, I\} \quad j \in \{1, \ldots, J\}$$

esto es, son independientes si la distribución conjunta coincide con el producto de las marginales.

(Obsérvese que esta definición es similar a la de sucesos independientes en el tema 2.)

En el ejemplo anterior, X e Y no son independientes porque $\mathbf{P}(X = 0, Y = 2) = 0$, mientras que $\mathbf{P}(X = 0) \cdot \mathbf{P}(Y = 2) = (1/4) \cdot (3/8) \neq 0$. En todo caso, parece lógico que el número de caras en los dos primeros lanzamientos, que es la variable X, tenga relación o dependencia con el número de caras en los tres lanzamientos, que es la variable Y.

Como ejemplos de variables independientes, podemos pensar en las variables $X =$"puntuación obtenida en el primer lanzamiento de un dado" e $Y =$"puntuación obtenida en el segundo lanzamiento del dado". Se entiende que el dado se lanza dos veces consecutivas de manera independiente, esto es, el primer lanzamiento no influye en el resultado del segundo.

Definición 6 (Medidas de posición y dispersión de (X, Y))
*Podemos seguir definiendo las medidas de posición o dispersión de cada una de las variables X e Y que componen un vector aleatorio (X, Y). Para ello, basta emplear las distribuciones marginales de X e Y. En el caso **discreto** se pueden definir*

$$E(X) = \sum_{i=1}^{I} x_i \cdot \mathbf{P}(X = x_i), \qquad Var(X) = E\left[\left(X - E(X)\right)^2\right] = \sum_{i=1}^{I} \left(x_i - E(X)\right)^2 \cdot \mathbf{P}(X = x_i)$$
$$E(Y) = \sum_{j=1}^{J} y_j \cdot \mathbf{P}(Y = y_j), \qquad Var(Y) = E\left[\left(Y - E(Y)\right)^2\right] = \sum_{j=1}^{J} \left(y_j - E(Y)\right)^2 \cdot \mathbf{P}(Y = y_j)$$

El vector $(E(X), E(Y))$ recibirá el nombre de vector de medias, igual que en Descriptiva.

En el Ejemplo 6 tenemos las medidas univariantes:

$$E(X) = 1, \qquad E(Y) = 1'5, \qquad Var(X) = 0'5, \qquad Var(Y) = 0'75.$$

Propiedades
Recogemos dos propiedades de aditividad de la media y la varianza. Respecto de la media, para cualquier variable multidimensional $(X_1, \ldots, X_k)$,

$$E(X_1 + \cdots + X_k) = E(X_1) + \cdots + E(X_k)$$

Sobre la varianza, si $X_1, \ldots, X_k$ son independientes, entonces

$$Var(X_1 + \cdots + X_k) = Var(X_1) + \cdots + Var(X_k)$$

Nótese que esta igualdad no se cumple si las variables no son independientes.

A continuación veremos dos de los modelos de distribución de probabilidad más importantes en variables aleatorias discretas. Serán la distribución "de Bernoulli"y "Binomial".

2.4 La distribución de Bernoulli

Realizamos un experimento aleatorio y nos preocupamos únicamente de si cierto suceso A ocurre. Por ejemplo, observamos si cierta sustancia está presente en un material. Diremos que se ha producido un

éxito si ha ocurrido el suceso A. Para indicar la ocurrencia del suceso A (éxito) definimos una variable aleatoria

$$X(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in A \\ 0 & \text{si} \quad \omega \notin A \end{cases}$$

Definición 7
*Si denotamos por $p = \mathsf{P}(A)$ a la probabilidad de éxito, entonces diremos que la variable X tiene distribución de **Bernoulli** de parámetro p, y lo denotamos "$X \sim$ Bernoulli(p)".*

La función de probabilidad o función de masa de $X \sim$ Bernoulli(p) viene dada por

x_i	0	1
$\mathsf{P}(X = x_i)$	$1 - p$	p

y se puede representar mediante el siguiente diagrama de barras:

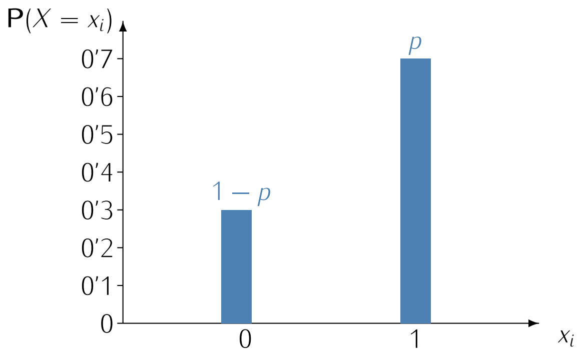


Figura 3. Función de masa de una distribución de Bernoulli(p), con $p = 0'7$.

Por tanto, la probabilidad de éxito p determina plenamente la distribución de Bernoulli. Llamaremos **parámetro** de un modelo de distribución a un número (o números, caso de ser más de uno) que determinan plenamente la distribución. De hecho, en el caso de la distribución de Bernoulli(p), la media y la varianza son función del parámetro p y son:

$$E(X) = p \qquad\qquad Var(X) = p(1 - p)$$

2.5 La distribución Binomial

Realizamos m veces el experimento aleatorio en idénticas condiciones y de forma independiente, y contamos el número de veces que ocurre el suceso A, esto es, el **número de éxitos**. A cada realización del experimento aleatorio le llamamos **intento**.

Definición 8
*La variable aleatoria X que representa el número de éxitos en n intentos independientes, siendo la probabilidad de éxito en cada intento $p = \mathsf{P}(A)$, diremos que tiene distribución **Binomial** de parámetros* **m** *y* **p**. *Lo denotamos "$X \sim$ Binomial(**m**,**p**)".*

La distribución binomial es discreta y toma los valores $0, 1, 2, 3, \ldots, m$. Las probabilidades asociadas se pueden obtener mediante cálculos de combinatoria, y resultan:

$$\mathsf{P}(X = k) = \binom{\mathsf{m}}{k} \mathsf{p}^k (1 - \mathsf{p})^{\mathsf{m}-k} \qquad \text{si} \quad k \in \{0, 1, 2, \ldots, m\}$$

Los programas de cálculo más comunes, como las hojas de cálculo y software matemático o estadístico como el lenguaje R, ya tienen implementados todos estos cálculos y ofrecen funciones para obtener las probabilidades bajo la distribución binomial. En ausencia de ordenador, para el cálculo de estas probabilidades dispondremos de tablas para distintos valores de los parámetros m y p.

Si definimos X_i como la variable –con distribución de Bernoulli(p)– asociada al i-ésimo intento, donde $i \in \{1, \dots, m\}$, entonces podemos expresar la distribución *Binomial(m,p)* como suma de las m variables *Bernoulli(p)*:

$$X = X_1 + X_2 + \cdots + X_m$$

Dado que esas m variables de Bernoulli son independientes, podemos calcular la media y la varianza de la Binomial a través de esa suma, para obtener:

$$E(X) = mp \qquad Var(X) = mp(1-p)$$

Ejemplo 7

En una población porcina hay una enfermedad que está determinada genéticamente por un solo gen. Este gen tiene dos alelos: "B" y "b". La enfermedad está asociada con el alelo recesivo "b", es decir, los animales con el genotipo "bb" tendrán la enfermedad, mientras que los animales con "Bb" serán sólo portadores. La probabilidad del alelo "b" es 0'5.

Una pareja de cerdos, ambos con genotipo "Bb", tiene diez lechones. Sea X ="número de lechones que tienen la enfermedad". Se propone:

(a) Representar la función de masa de X.

(b) Calcula el número de lechones que se espera tengan la enfermedad.

(c) Calcula la probabilidad de que al menos un lechón tenga la enfermedad.

(d) Calcula la probabilidad de que exactamente la mitad de la camada tenga la enfermedad.

Solución. La variable de este ejemplo tiene distribución binomial de parámetros $m = 10$ y $p = 0'25$*, esto es, X ∼ Binomial($m = 10$, $p = 0'25$). Contestamos a cada una de las cuestiones:

(a) En la siguiente figura se representa su función de probabilidad o función de masa. Los valores posibles van de 0 a 10, y el valor más probable es el 2. Las probabilidades decrecen según nos desviamos de 2 tanto a la izquierda como a la derecha, aunque hay cierta asimetría.

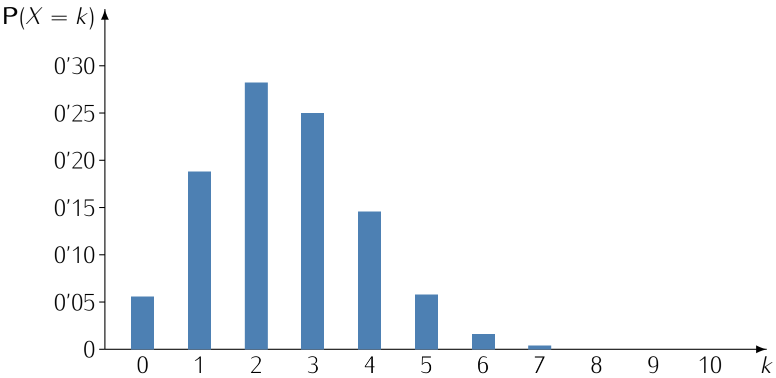


Figura 4. Función de masa de X ∼ Binomial(m,p), con m = 10, p = 0'25.

(b) $E(X) = mp = 2'5$ (Esperamos que tengan la enfermedad entre dos y tres lechones.)

Observemos que la distribución binomial tiene una única moda en uno de los dos valores más próximos (o en el valor más próximo) a la media, y las probabilidades decrecen a ambos lados según nos alejamos de la media. Esto ocurre en este ejemplo y también en general para otros parámetros.

(c) P("Que al menos uno esté enfermo") = $P(X \geq 1) = 1 - P(X = 0) = 1 - 0'0563 = 0'9437$.

(d) P("Que enfermen cinco") = $P(X = 5) = 0'0584$.

*p es la probabilidad de "que un lechón elegido al azar esté enfermo", es decir, "que tenga genotipo *bb*".

En ausencia de ordenador vamos a emplear una tabla que proporcione los valores de la función de masa, $P[X = x_0]$ de una distribución **Binomial(m,p)** para algunos valores de p desde $0'01$ hasta $0'50$ y para los valores de m desde 2 hasta 10. Tendremos así una rejilla de valores de las probabilidades $P[X = x_0]$, donde x_0 varía desde 0 hasta m.

Recurriendo al ejemplo, si tenemos una distribución Binomial(10,0'25), para encontrar $\mathbf{P}(X = 5)$ tenemos que buscar en la columna de la izquierda el grupo de valores correspondiente a $m = 10$ y en la fila superior el valor de $p = 0'25$. En el cruce de esta columna con el grupo de valores de la fila $m = 10$ se encuentran las probabilidades que queremos calcular:

m	k	p=0'25
10	0	$P[X = 0] = 0'0563$
	1	$P[X = 1] = 0'1877$
	2	$P[X = 2] = 0'2816$
	3	$P[X = 3] = 0'2503$
	4	$P[X = 4] = 0'1460$
	5	$P[X = 5] = 0'0584$
	6	$P[X = 6] = 0'0162$
	7	$P[X = 7] = 0'0031$
	8	$P[X = 8] = 0'0004$
	9	$P[X = 9] = 0'0000$
	10	$P[X = 10] = 0'0000$

□

Nota. En la tabla se omiten los valores de p mayores de $0'50$, pues las probabilidades correspondientes se pueden obtener considerando la variable $Y = m - X$, cuya distribución es Binomial(m,1-p).

Así, por ejemplo, si X $\sim$ Binomial(10, $0'70$), para calcular $\mathbf{P}(X = x_0)$, tendremos en cuenta que $Y = m - X \sim$ Binomial(10, $0'30$) y que $\mathbf{P}(X = x_0) = \mathbf{P}(Y = m - x_0)$, pues $\mathbf{P}(Y = m - x_0)$ ya se encuentra en la tabla. En concreto, para $x_0 = 4$, $\mathbf{P}(X = 4) = \mathbf{P}(Y = 6) = 0'0368$.

3 Variable aleatoria continua

En este tema estudiaremos las variables aleatorias continuas y la distribución normal, que es el modelo más común para este tipo de variables. La idea de continuidad ya fue estudiada en el tema de Estadística descriptiva. Una variable es continua si sus valores posibles se encuentra en un intervalo, y resulta improbable la repetición de resultados idénticos en individuos diferentes, salvo por errores de redondeo. Sirvan como ejemplos magnitudes físicas que se vienen modelizando como variables continuas: peso, longitud, tiempo, etcétera. En este tema pensaremos en este tipo de mediciones antes de realizar el experimento, esto es, como variables aleatorias.

3.1 Concepto de variable aleatoria continua, función de densidad y función de distribución

Por *variable aleatoria* entendemos, al igual que en la sección anterior, una aplicación que a cada resultado de un experimento aleatorio le asigna un número real, obtenido por la medición de cierta característica.

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longrightarrow & X(\omega) \end{array}$$

Ahora nos centramos en las variables aleatorias *continuas*, cuyos resultados posibles se encuentran en un intervalo (acotado o no).

Si una variable aleatoria X tiene asociada una función f no negativa, tal que la integral de f sobre cada intervalo da la probabilidad de que X esté en ese intervalo, entonces llamamos a f la **función de densidad de probabilidad** de la variable aleatoria X, y decimos que X tiene una **distribución continua**.

Tenemos, entonces, que:

1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}$

2. $\mathbf{P}(a \leq X \leq b) = \displaystyle\int_a^b f(x)\,dx$
   (Desde un punto de vista geométrico, es el área bajo la función de densidad sobre dicho intervalo).

Y, por lo tanto:

3. $\mathbf{P}\,(X \leq x_0) = \displaystyle\int_{-\infty}^{x_0} f(x)\,dx \qquad \forall x_0 \in \mathbb{R}$

4. $\displaystyle\int_{\mathbb{R}} f(x)\,dx = 1$

En la figura siguiente se representa un ejemplo de función de densidad bimodal, con modas en los valores 3 y 6, que son los lugares donde la densidad es más alta. Se escoge como ejemplo de x_0 el valor 5 y se representa el área (sombreada) bajo la curva hasta $x_0 = 5$. Dicha área representa la probabilidad $\mathbf{P}(X \leq x_0) = \mathbf{P}(X \leq 5) = 0'663$. Los valores de la variable se sitúan en el eje de abscisas (eje horizontal), y la idea es que las áreas bajo la curva son las probabilidades de las zonas en el eje de abscisas, que soportan esas áreas. Si la densidad es más alta se consigue más probabilidad con similar longitud en la base.

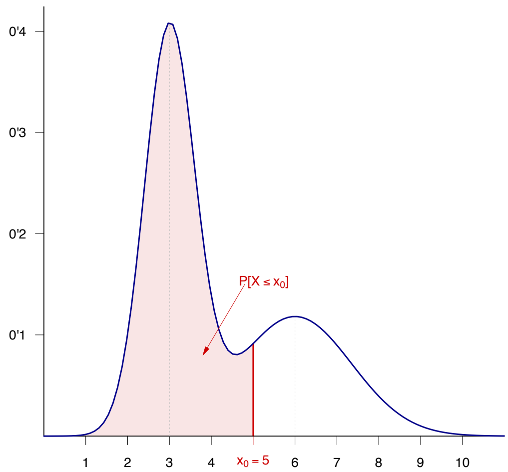


Figura 5. Función de densidad de una variable continua bimodal, con modas en los valores 3 y 6.

La **función de distribución** fue definida en la sección anterior para cualquier variable aleatoria, por tanto también para las continuas. De la definición de función de densidad se deduce la siguiente relación con la función de distribución:

$$F\,(x_0) = \mathbf{P}[X \leq x_0] = \int_{-\infty}^{x_0} f(x)\,dx$$

Por aplicación del teorema fundamental del cálculo integral obtenemos:

1. F es continua 2. $f\,(x_0) = F'\,(x_0)$

En la figura siguiente se representa la función de distribución asociada a la densidad de la figura anterior. Se indica el valor de la función de distribución para $x_0 = 5$, que coincide con el área sombreada en la gráfica de la función de densidad, y resulta ser $F(x_0) = \mathbf{P}(X \leq x_0) = \mathbf{P}(X \leq 5) = 0'663$.

En general observamos que la función de distribución es creciente partiendo de cero y llegando a uno, que es continua (pues la variable aleatoria es continua) y que crece más deprisa en las zonas donde la densidad es más alta. Así, en este ejemplo la función de distribución crece más rápido en el entorno del 3 y en el entorno del 6.

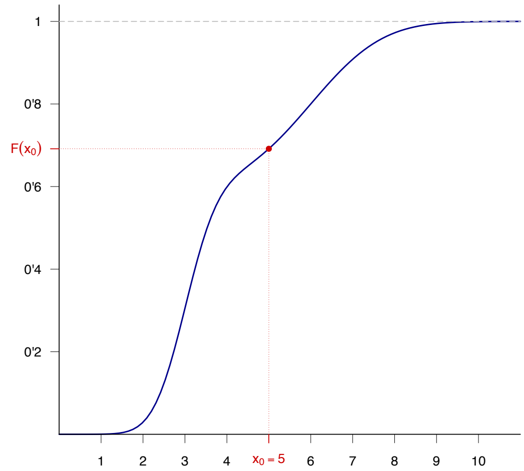


Figura 6. Función de distribución asociada a la densidad de la figura 5.

Ejemplo 8

La extremidad de un animal mide aproximadamente 6 centímetros. Sobre esta extremidad se pueden ubicar ciertos parásitos. Cada parásito se sitúa sobre algún punto de la extremidad de manera aleatoria. A este punto lo denotamos por X, y lo medimos desde 0 cm en el punto donde la extremidad se une al tronco y 6 cm en el punto más lejano del tronco. Suponemos que la variable aleatoria X se distribuye de acuerdo con la siguiente función de densidad:

$$f(x) = \begin{cases} \dfrac{x^2}{72} & \text{si } \; x \in [0,6] \\[2ex] 0 & \text{si } \; x \notin [0,6] \end{cases}$$

Vamos a calcular la función de distribución de X y obtener la probabilidad $\mathbf{P}(3 < X \leq 5)$.

Solución. Lo primero que queremos destacar es que la variable X sólo toma valores en el intervalo $[0,6]$. Entonces diremos que el intervalo $[0,6]$ es el **soporte** de la distribución de X. Este concepto lo aplicaremos a cualquier otra variable aleatoria, para la cual su soporte será el menor conjunto donde la variable aleatoria toma valores (con probabilidad uno).

Para calcular la función de distribución, observamos que antes de 0 vale cero y después de 6 vale uno, pues $[0,6]$ es el soporte. Ahora, para un valor $x \in [0,6)$, se tiene que

$$F(x) = \mathbf{P}(X \leq x) = \int_0^x \frac{u^2}{72}\, du = \left[\frac{u^3}{216}\right]_0^x = \frac{x^3}{216}$$

Por tanto, podemos escribir la función de distribución de la siguiente manera:

$$F(x) = \begin{cases} 0 & \text{si } \; x \in (-\infty, 0) \\[2ex] \dfrac{x^3}{216} & \text{si } \; x \in [0,6) \\[2ex] 1 & \text{si } \; x \in [6, +\infty) \end{cases}$$

Respecto de la probabilidad solicitada, tenemos que

$\mathbf{P}(3 < X \leq 5) = F(5) - F(3) = 0'5787037 - 0'125 = 0'4537037 \approx 0'454$

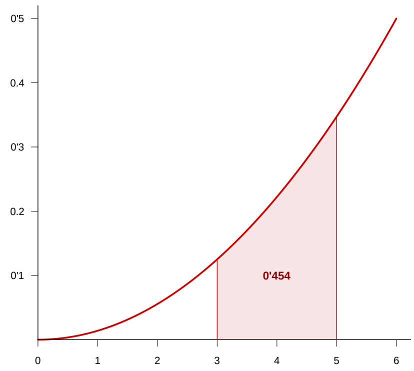


Figura 7. La función de densidad

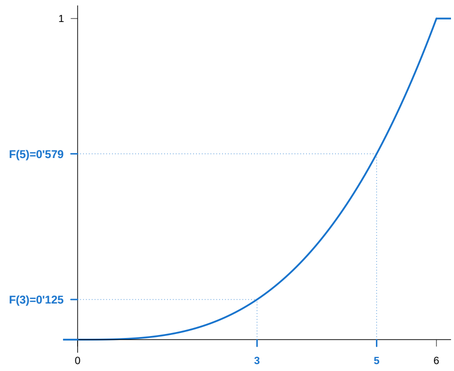


Figura 8. La función de distribución

□

3.2 Medidas de posición y dispersión calculadas sobre una variable aleatoria continua

En la sección anterior hemos calculado las medidas de posición y dispersión sobre variables aleatorias discretas. Estas medidas también se pueden calcular sobre variables aleatorias continuas. La diferencia radica en que para las discretas aparece un sumatorio ponderado por las probabilidades, mientras que para las continuas aparece una integral ponderada por la función de densidad: en lugar de usar los valores de la variable aleatoria discreta y sus probabilidades, ahora emplearemos la función de densidad sobre cierto soporte, donde toma valores la variable aleatoria continua.
Las interpretaciones y propiedades de cada medida se conservan, sólo cambia la manera de calcularlas.

Media o Esperanza. Si X es una variable aleatoria continua con función de densidad f, su esperanza se calcula así:

$$E(X) = \mu = \int x f(x)\, dx$$

Se siguen cumpliendo las propiedades en relación con cambios de localización y escala, esto es, $E(a + bX) = a + bE(X)$ siendo a y b constantes.

Mediana. En esta situación, es el valor "m" de la variable aleatoria que deja probabilidad 0'5 tanto a su izquierda como a su derecha. Se puede calcular a través de la función de distribución así:

$$m = F^{-1}(0'5)$$

Si X es una variable aleatoria continua, no habrá saltos, pero sí puede haber zonas constantes donde la inversa sea todo un intervalo. En ese caso tomaremos como mediana el menor valor de ese intervalo.

Moda. La moda se suele definir como el valor más probable. Para las variables continuas los puntos tienen probabilidad cero, por lo que no tendría sentido definir la moda de esa manera. Sin embargo, se puede definir la moda como el máximo de la función de densidad (nótese que aquí no es necesario agrupar en intervalos, como se hacía en Descriptiva). En general se suelen aceptar como modas todos los máximos relativos de la función de densidad, y no sólo el máximo absoluto.

Cuantiles. Sea $p \in (0, 1)$. Recordamos que el cuantil p de una variable aleatoria es el valor "Q_p" que deja una probabilidad p a su izquierda. En consecuencia,

$$Q_p = F^{-1}(p)$$

De nuevo, en caso de que haya una zona constante en p, tendríamos que tomar como cuantil Q_p el menor valor del intervalo correspondiente.

Varianza. Si X es una variable aleatoria continua con función de densidad f, su esperanza se calcula así:

$$\text{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)\,dx$$

Se sigue cumpliendo que $\text{Var}[a+bX] = b^2\text{Var}(X)$.

Desviación típica. Sigue siendo la raíz cuadrada de la varianza y, por tanto, se denota por $\sigma = \text{Dt(X)} = \sqrt{\text{Var(X)}}$.

Ejemplo 1 (continuación)

Vamos a calcular la media, varianza y desviación típica de X. También calcularemos la mediana de X.

Solución.

La media de X sería:

$$E(X) = \int x f(x)\,dx = \int_0^6 x\left(\frac{x^2}{72}\right)\,dx = \left[\frac{x^4}{288}\right]_0^6 = \frac{6^4}{288} = 4'5\,cm$$

Para calcular la varianza:

$$\begin{aligned}
Var(X) = & \int [x-\mu]^2 f(x)\,dx = \int_0^6 \left(x-\frac{9}{2}\right)^2\left(\frac{x^2}{72}\right)\,dx = \\
= & \int_0^6 \left[x^2 - 2x\left(\frac{9}{2}\right) + \left(\frac{9}{2}\right)^2\right]\left(\frac{x^2}{72}\right)\,dx = \\
= & \int_0^6 \frac{x^4}{72}\,dx + \int_0^6 \frac{9x^3}{72}\,dx + \int_0^6 \frac{81}{4}\frac{x^2}{72}\,dx = \\
= & \frac{1}{72}\left[\frac{6^5}{5} - \frac{9\cdot 6^4}{4} + \frac{81}{4}\frac{6^3}{3}\right] = \\
= & 1'35\ cm^2
\end{aligned}$$

La desviación típica será entonces

$$Dt(X) = \sqrt{Var(X)} = \sqrt{1'35} = 1'16\quad cm$$

Para la mediana, tendremos que invertir la función de distribución en $0'5$, esto es, $Mediana(X) = F^{-1}(0'5)$. Entonces, si m es la mediana, se tiene la ecuación $F(m) = m^3/216 = 0'5$, cuya solución es

$$Mediana(X) = m = (0'5\cdot 216)^{1/3} = 4'76\quad cm$$

□

3.3 Variable aleatoria continua multidimensional

Como vimos en la sección anterior, una variable aleatoria multidimensional o vector aleatorio es una colección de variables aleatorias $(X_1,\ldots,X_k)$ medidas simultáneamente sobre el mismo individuo o sobre el mismo resultado de un experimento aleatorio. Por tanto, se puede representar como una aplicación:

$$X = (X_1,\ldots,X_k) : \Omega \longrightarrow \mathbb{R}^k$$

Definición 9

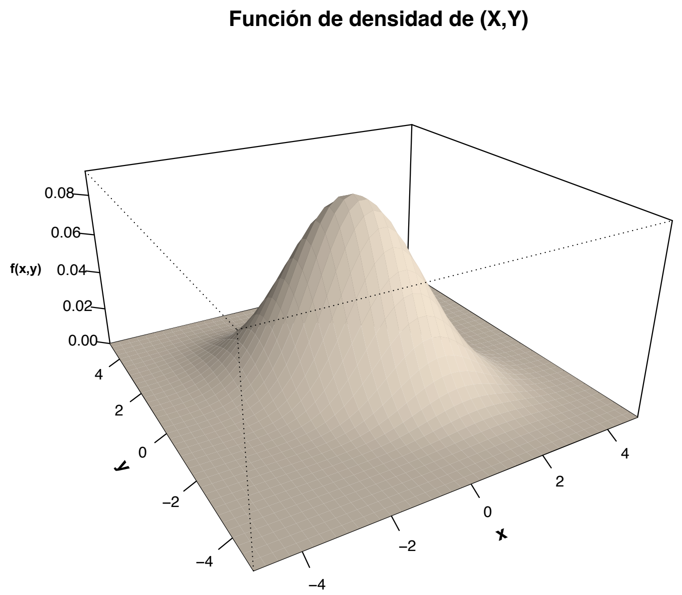
Consideramos que el vector aleatorio $(X_1,\ldots,X_k)$ es ***continuo*** *si existe una* ***función de densidad conjunta*** *$f:\mathbb{R}^k\to\mathbb{R}$ no negativa y tal que*

$$\mathsf{P}(X\le x_0) = \int_{-\infty}^{x_0} f(x)\,dx \qquad \forall x_0\in\mathbb{R}^k$$

La notación $X \leq x$ en el caso de vectores se interpreta como $X_j \leq x_j$ para todo $j \in \{1, 2, \ldots, l\}$ siendo X_j y x_j las j-ésimas componentes de X y x, respectivamente. Además, debemos tener presente que la integral se está calculando en un recinto del espacio de dimensión k. En general, la probabilidad de cualquier conjunto en ese espacio se puede calcular integrando la función de densidad sobre ese conjunto.

En particular, si consideramos un vector aleatorio bidimensional, $k = 2$, lo denotaremos por (X, Y) y tendremos que una función de densidad conjunta $f : \mathbb{R}^2 \to \mathbb{R}$ tal que

$$\mathbf{P}(X \leq x_0, Y \leq y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f(x, y)\ dy\, dx \qquad \text{para todo } (x_0, y_0) \in \mathbb{R}^2$$



A partir de la densidad conjunta, se obtiene la marginal de X, integrando en y:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\ dy \qquad \forall x \in \mathbb{R}$$

De manera similar se define la densidad marginal de Y.
La función de densidad condicionada de Y a X se definiría así:

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

y de modo similar con la densidad condicionada de X a Y.
Las variables X e Y serán independientes si

$$f(x, y) = f_X(x) f_Y(y)\ \forall x, y.$$

Las medias y las varianzas se calculan mediante las integrales correspondientes.
A continuación veremos uno de los modelos de distribución de probabilidad más importantes en variables aleatorias continuas.

3.4 La distribución normal

La distribución normal es la más importante y de mayor uso de todas las distribuciones continuas de probabilidad. Por múltiples razones se viene considerando la más idónea para modelizar una gran diversidad de mediciones de la Física, Química o Biología. Entre estas razones estudiaremos el teorema central del límite, que justifica la utilización de la normal como aproximación para las distribuciones de variables aleatorias bajo ciertas condiciones. En particular, lo usaremos para aproximar la distribución binomial.

Empezamos definiendo la distribución normal estándar, que es la que tiene media cero y varianza uno, para después estudiar la distribución normal (general), esto es, con cualquier media y varianza.

3.4.1 La distribución normal estándar, $N(0, 1)$

Definición 10

*Una variable aleatoria continua Z se dice que tiene **distribución normal estándar**, y lo denotamos $Z \sim N(0, 1)$, si su función de densidad viene dada por:*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad si \quad z \in \mathbb{R}$$

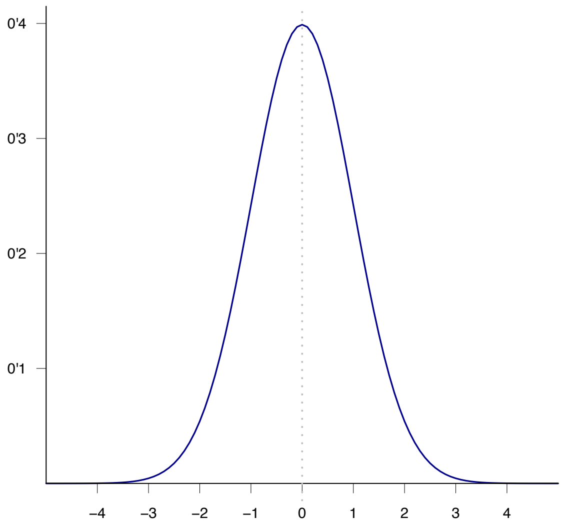


Figura 9. Función de densidad normal estándar, $f(z)$.

Propiedades

1. $Z \sim N(0, 1)$ toma valores en toda la recta real. ($f(z) > 0 \quad \forall z \in \mathbb{R}$)
2. f es simétrica en torno a cero.
3. f tiene dos puntos de inflexión en -1 y +1.
4. Si $Z \sim N(0, 1)$ entonces $E(Z) = 0$ y $Var(Z) = 1$.

Ejemplo 9

Supongamos que $Z \sim N(0, 1)$. Calcularemos $\mathsf{P}(Z \leq 1'3)$.

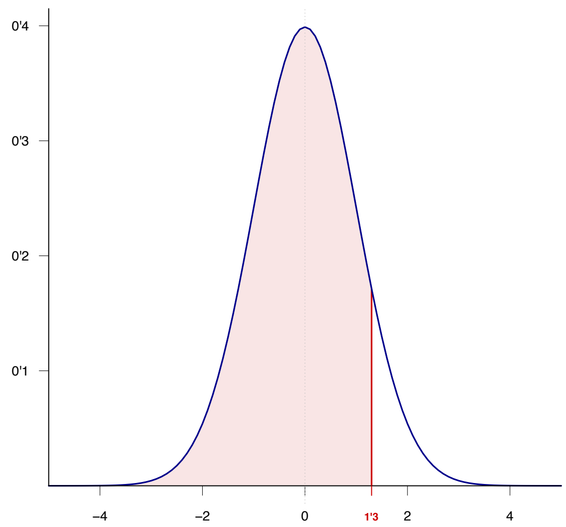


Figura 10. El área sombreada representa $\mathsf{P}(Z \leq 1'3)$*, para* $Z \sim N(0,1)$.

No existe una expresión explícita para el área, sino que es preciso emplear algoritmos de resolución numérica. Los programas de cálculo más comunes, como las hojas de cálculo y software matemático o estadístico como el lenguaje R, ya tienen implementados estos algoritmos y ofrecen funciones para obtener las probabilidades bajo la distribución normal.

En ausencia de ordenador vamos a emplear una tabla que proporciona los valores de la función de distribución, $F(z_0) = \mathsf{P}(Z \leq z_0)$ para una rejilla de valores positivos de z_0, desde 0 hasta 4'09, con incremento de 0'01.

Presentamos a continuación un fragmento de las tablas que utilizaremos:

z_0	0'00	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
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	0'90147
1'3	0'90320	0'90490	0'90658	0'90824	0'90988	0'91149	0'91308	0'91466	0'91621	0'91774
1'4	0'91924	0'92073	0'92220	0'92364	0'92507	0'92647	0'92785	0'92922	0'93056	0'93189

En el ejemplo, para encontrar $\mathsf{P}(Z \leq 1'3)$ tenemos que buscar en el cruce de la fila del 1'3 y la columna del 0'00, pues allí se encuentra $\mathsf{P}(Z \leq 1'3) = 0'90320$.

Para los valores negativos de z_0 aplicaremos argumentos de simetría, para las probabilidades a la derecha emplearemos el complementario, pues $\mathsf{P}(Z > z_0) = 1 - \mathsf{P}(Z \leq z_0)$, y para un intervalo usaremos la diferencia, ya que $\mathsf{P}(Z \in (a,b]) = \mathsf{P}(Z \leq b) - \mathsf{P}(Z \leq a)$. Además, será lo mismo $\mathsf{P}(Z \leq z_0)$ que $\mathsf{P}(Z < z_0)$, pues $\mathsf{P}(Z = z_0) = 0$, de modo que es indiferente incluir o no los extremos de los intervalos. Con estas ideas se puede calcular la probabilidad de cualquier intervalo, utilizando las tablas o un programa que proporcione la función de distribución. Lo vemos en el ejemplo siguiente.

$\square$

Ejemplo 10
Sea $Z \sim N(0,1)$. Calcularemos:

(a) $\mathsf{P}(Z \leq 1'52)$

(b) $\mathsf{P}(Z > 2)$

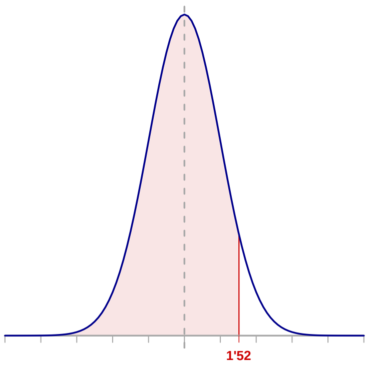
(c) $\mathsf{P}(Z > -2'51)$

(d) $\mathsf{P}(Z < -0'76)$

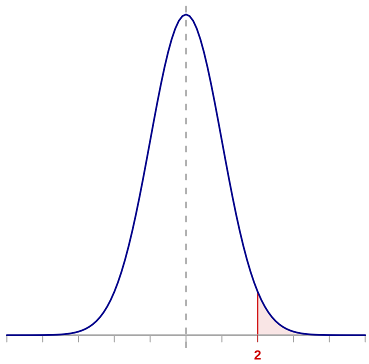
(e) $\mathsf{P}(\mid Z \mid \leq 1)$

(f) $\mathsf{P}(-1 \leq Z \leq 2)$

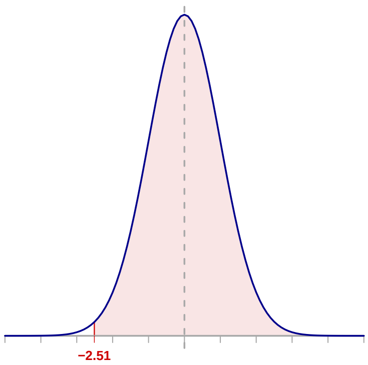
(g) ¿Cuánto vale aproximadamente $\mathsf{P}(Z > 4'2)$?



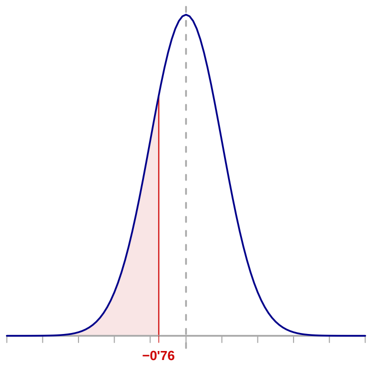
$\mathsf{P}(Z \leq 1'52) = 0'93574$



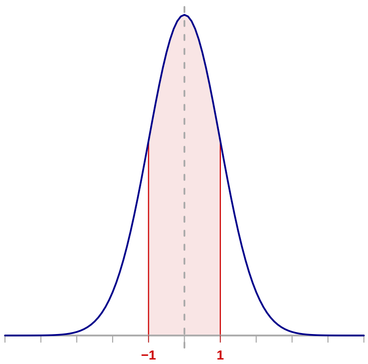
$\mathsf{P}(Z > 2) = 1 - \mathsf{P}(Z \leq 2) = 1 - 0'97725 = 0'02275$



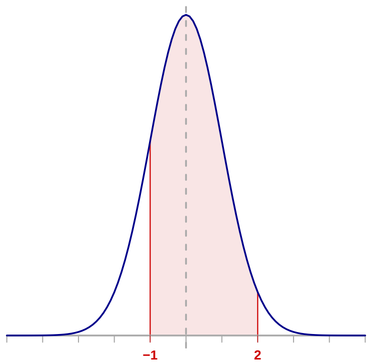
$\mathsf{P}(Z > -2'51) = \mathsf{P}(Z < 2'51) = 0'993963$



$\mathsf{P}(Z < -0'76) = 1 - \mathsf{P}(Z < 0'76) = 1 - 0'7764 = 0'2236$



$\mathsf{P}(\mid Z \mid \leq 1) = 1 - 2 \cdot \mathsf{P}(Z > 1) = 1 - 2 \cdot (1 - 0'8413) = 0'6826$



$\mathsf{P}(-1 \leq Z \leq 2) = \mathsf{P}(Z \leq 2) - \mathsf{P}(Z < -1) = 0'81855$

□

También se pueden leer las tablas al revés, esto es, conociendo la probabilidad, buscar el número que deje esa probabilidad a la izquierda, o a la derecha, o a ambos lados. Esto será muy utilizado en los temas siguientes. Realmente estaríamos calculando *cuantiles* de la distribución normal. No sólo se podrá hacer mediante las tablas, sino que los programas más comunes, como las hojas de cálculo y desde luego el lenguaje R, también contienen *la función cuantil* de la distribución normal.

Veremos cómo se pueden obtener cuantiles, expresados en diversas formas, a través del ejemplo siguiente.

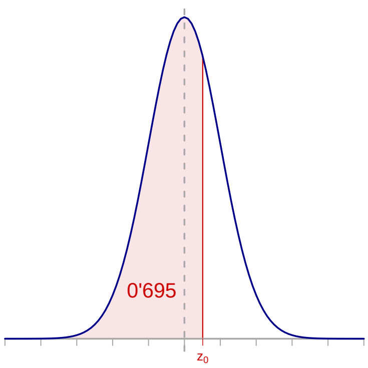
Ejemplo 11
Sea $Z \sim N(0,1)$. Halla los valores z_0 tales que

(a) $\mathrm{P}\left(Z \leq z_0\right) = 0'695$

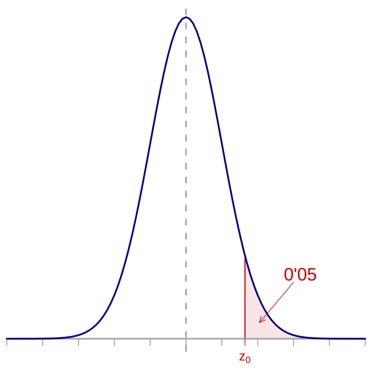
(b) $\mathrm{P}\left(Z > z_0\right) = 0'05$

(c) $\mathrm{P}\left(Z > z_0\right) = 0'975$

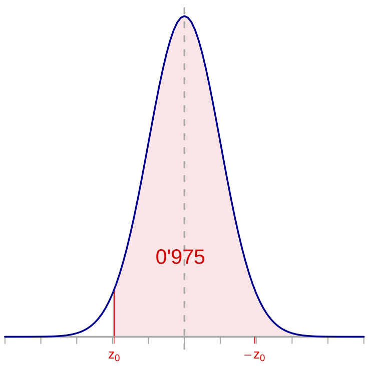
(d) $\mathrm{P}\left(\mid Z \mid > z_0\right) = 0'01$



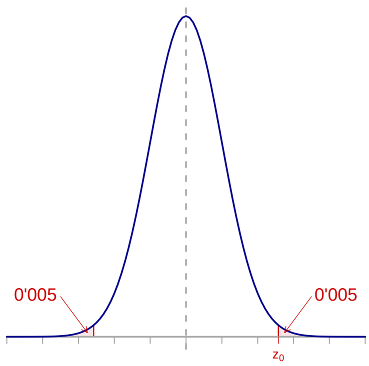
$\mathrm{P}(Z \leq z_0) = 0'695 \Rightarrow z_0 = 0'51$



$\mathrm{P}(Z > z_0) = 0'05 \Rightarrow \mathrm{P}(Z \leq z_0) = 0'95 \Rightarrow z_0 = 1'645$



$\mathrm{P}(Z > z_0) = 0'975 \Rightarrow \mathrm{P}(Z < -z_0) = 0'975 \Rightarrow z_0 = -1'96$



$\mathrm{P}(\mid Z \mid > z_0) = 0'01 \Rightarrow \mathrm{P}(Z \leq z_0) = 0'995 \Rightarrow z_0 = 2'575$

□

3.4.2 La distribución normal, $N(\mu, \sigma)$

Si efectuamos un cambio de localización y escala sobre la normal estándar, podemos obtener una distribución con la misma forma pero con la media y desviación típica que queramos.

Definición 11
Si $Z \sim N(0,1)$ entonces

$$X = \mu + \sigma\, X \sim N\left(\mu, \sigma\right),$$

*y diremos que X tiene **distribución normal de media** μ **y desviación típica** σ.*

Así, la función de densidad de X tendrá la misma forma de campana, será simétrica en torno a la media μ y sus puntos de inflexión serán $(\mu - \sigma)$ y $(\mu + \sigma)$. Basándonos en el cambio de variable, se puede obtener la expresión de la función de densidad de X:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{si} \quad x \in \mathbb{R}$$

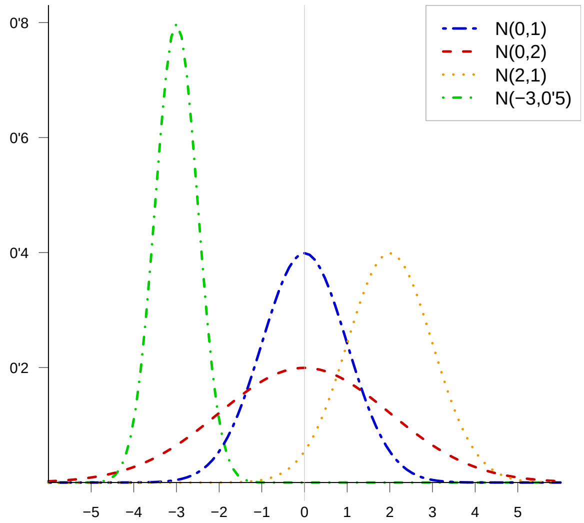


Figura 11. Funciones de densidad de variables normales con distintas medias y desviaciones típicas.

Si a una variable aleatoria normal le restamos su media y dividimos por su desviación típica, obtenemos una variable aleatoria normal estándar (con media cero y desviación típica uno). A este proceso le llamamos **estandarización** de una variable aleatoria.

$$\text{Si } X \sim N(\mu,\sigma) \text{ entonces } Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

Debemos observar que la estandarización se puede aplicar a cualquier variable aleatoria, tenga o no distribución normal. Al estandarizar una variable aleatoria, obtendremos otra (variable estandarizada) con media cero y desviación típica uno.

En R disponemos de funciones para obtener los valores de la función de densidad, de distribución y cuantiles de la normal con cualquier media y desviación típica.

Sin embargo, sólo disponemos de tablas de la normal estándar, por lo que con las tablas siempre se va a necesitar estandarizar la variable.

Ejemplo 12

Supongamos que $X \sim N(5,2)$. Calcularemos $\mathbf{P}(X \leq 1)$ y el valor x_0 tal que $\mathbf{P}(X > x_0) = 0'3$.

Solución.

$$\mathbf{P}(X \leq 1) = \mathbf{P}\left(\frac{X-5}{2} \leq \frac{1-5}{2}\right) = \mathbf{P}(Z \leq -2) = 1 - 0'97725 = 0'02275$$

donde $Z = (X-5)/2 \sim N(0,1)$.

Para obtener x_0 tal que $\mathbf{P}(X > x_0) = 0'3$, empleamos la estandarización, de modo que

$0'3 = \mathbf{P}(X > x_0) = \mathbf{P}\left(\frac{X-5}{2} > \frac{x_0-5}{2}\right) = \mathbf{P}(Z > z_0)$ siendo $z_0 = (x_0-5)/2$.

Para poder acudir a la tabla tenemos que argumentar así

$\mathbf{P}(Z \leq z_0) = 1 - \mathbf{P}(Z > z_0) = 0'7$.

Ahora, el valor z_0 de la tabla tal que $\mathbf{P}(Z \leq z_0) = 0'7$ está entre $0'52$ y $0'53$ por lo que tomamos $z_0 = 0'525$. Despejando se obtiene $x_0 = 0'525 * 2 + 5 = 6'05$.

□

4 Aproximación de otras distribuciones por la distribución normal. Teorema central del límite

Empezamos aproximando la distribución binomial por la normal. Si mantenemos fija la probabilidad de éxito p e incrementamos el número de intentos m hasta valores muy grandes (se escribiría $m \to +\infty$), al representar la distribución de probabilidad Binomial(m, p) observamos que adopta una forma de campana muy parecida a la función de densidad normal. Al mismo tiempo, para m grande no disponemos de tablas para las probabilidades binomiales y la fórmula se hace impracticable. Por todo esto, resulta tentador calcular esa probabilidad "como si la variable fuera normal". Sin embargo, necesitamos una justificación de que esto es válido, de que el error cometido es pequeño o se hace pequeño en ciertas condiciones. Esto lo aporta el Teorema de De Moivre-Laplace, precursor del teorema central del límite.

4.1 Teorema de De Moivre-Laplace

Tomemos una probabilidad de éxito fija p.
Consideremos una sucesión de variables aleatorias $X_m \sim$ Binomial(m, p), $m \in \{1, 2, 3, \ldots\}$, y sea $Z \sim N(0, 1)$.
Entonces:

$$\lim_{m\to\infty} \mathbf{P}\left(\frac{X_m - mp}{\sqrt{mp(1-p)}} \leq z\right) = \mathbf{P}(Z \leq z) \qquad \forall\, z \in \mathbb{R}$$

Denotaremos esto más brevemente así:

$$\frac{X_m - mp}{\sqrt{mp(1-p)}} \xrightarrow{d} Z$$

Nótese que hemos estandarizado la variable Binomial. Utilizaremos el siguiente criterio:

- Si $m \geq 30$, $mp \geq 10$ y $m(1-p) \geq 10$, entonces la Binomial(m, p) se puede aproximar por una normal de media $\mu = mp$ y desviación típica $\sigma = \sqrt{mp(1-p)}$.

4.2 Elemento de corrección por continuidad

Al pasar de la distribución binomial a la normal estamos aproximando una distribución discreta por otra continua. Esto produce trastornos al calcular la probabilidad de un punto (no nula para la discreta y nula para la continua) o a la hora de discernir entre "$<$" y "$\leq$" en las expresiones. Para solventar esto, aproximamos la probabilidad de un valor de la distribución binomial $k \in \{0, 1, 2, \ldots, n\}$ por la del intervalo centrado en k de longitud unidad $(k - 0'5, k + 0'5]$, excepto para el 0 y el n, valores menor y mayor de la distribución discreta, respectivamente, cuyas probabilidades se aproximan por las de los intervalos $(-\infty, 0'5]$ y $(n - 0'5, +\infty)$.

Ejemplo 13
En un campo de 100 árboles, cada árbol tiene una probabilidad del 20% de ser infectados por una enfermedad de la raíz, independientemente de otros árboles. ¿Cuál es la probabilidad de que estén infectados entre 10 y 15 árboles, ambos extremos incluidos?

Solución. Denotemos X ="Número de árboles infectados".

Sabemos que $X \sim$ Binomial$(m = 100, p = 0'2)$.

Como $m \geq 30$, $mp \geq 10$ y $m(1-p) \geq 10$, podemos aproximar la binomial por una normal de media $\mu = mp = 20$ y desviación típica $\sigma = \sqrt{mp(1-p)} = 4$.
Denotaremos $X_N \sim N(20, 4)$ a una variable aleatoria normal en esas condiciones, que nos sirve de aproximación de X, y $Z \sim N(0, 1)$. Entonces

$$\begin{aligned}
\mathbf{P}(10 \leq X \leq 15) \quad &\approx \quad \mathbf{P}(9'5 \leq X_N \leq 15'5) = \mathbf{P}\left(\frac{9'5 - 20}{4} \leq \frac{X_N - 20}{4} \leq \frac{15'5 - 20}{4}\right) = \\
&= \quad \mathbf{P}(-2'625 \leq Z \leq -1'125) = \mathbf{P}(1'125 \leq Z \leq 2'625) \\
&= \quad \mathbf{P}(Z \leq 2'625) - \mathbf{P}(Z \leq 1'125) = 0'995667 - 0'8697 = 0'125967
\end{aligned}$$

Observemos que $\mathbf{P}(Z \leq 2'625)$ es un valor comprendido entre $\mathbf{P}(Z \leq 2'62) = 0'995603$ y $\mathbf{P}(Z \leq 2'63) = 0'995731$, por lo que tomaremos dicha probabilidad como su punto medio ($0'995667$).
Análogamente, la $\mathbf{P}(Z \leq 1'125)$ es un valor comprendido entre $\mathbf{P}(Z \leq 1'12) = 0'8686$ y $\mathbf{P}(Z \leq 1'13) = 0'8708$, por lo que adoptaremos su punto medio ($0'8697$) como el valor de dicha probabilidad.

□

4.3 Teorema central del límite

No sólo se puede aproximar la distribución binomial por la normal, sino que hay un teorema general, que es el Teorema Central del Límite, que permite aproximar por la normal cualquier variable que resulte de la suma de variables independientes. Lo enunciamos a continuación.

Sea $X_1, X_2, \ldots, X_m, \ldots$ una sucesión de variables aleatorias independientes y con la misma distribución. Denotamos $S_m = X_1 + X_2 + \ldots + X_m$ a la suma, y sea $Z \sim N(0,1)$. Entonces

$$\frac{S_m - E(S_m)}{\sqrt{Var(S_m)}} \xrightarrow{d} Z$$

Como la binomial se puede interpretar como la suma de m variables independientes con distribución de Bernoulli, el Teorema de De Moivre-Laplace, que nos proporciona la aproximación de la binomial por la normal, se puede obtener como consecuencia del teorema central del límite.

5 Propiedades de aditividad para la Binomial y la Normal

Por último presentamos una serie de propiedades muy útiles sobre la suma de variables independientes.

- Si $X_1 \sim \text{Binomial}(m_1, p)$, $X_2 \sim \text{Binomial}(m_2, p)$ y son independientes, entonces
$$X_1 + X_2 \sim \text{Binomial}(m_1 + m_2, p).$$
- Si $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2)$ y son independientes, y $a, b_1, b_2 \in \mathbb{R}$, entonces
$$a + b_1X_1 + b_2X_2 \sim N\left(a + b_1\mu_1 + b_2\mu_2, \sqrt{b_1^2\sigma_1^2 + b_2^2\sigma_2^2}\right).$$

Ejemplo 14
Un aparato electrónico consta de dos piezas A y B. El peso de la pieza A tiene distribución normal de media 50 g y desviación típica 4 g, y el peso de la pieza B tiene distribución normal de media 20 g y desviación típica 3 g. Suponiendo que los pesos de las dos piezas son independientes, calcular la probabilidad de que el peso total del aparato supere los 80 g.

Solución. Denotemos $X_A =$"Peso de la pieza A" y $X_B =$"Peso de la pieza B".

El peso del aparato electrónico, denotado por X, es $X = X_A + X_B$.

Sabemos que $X \sim N(50,4)$ y $X_B \sim N(20,3)$, por lo tanto, $X \sim N(50+20, \sqrt{4^2+3^2})$, es decir, $X \sim N(70,5)$.

Ahora se puede calcular la probabilidad pedida, estandarizando X, esto es, considerando $Z = (X-70)/5 \sim N(0,1)$:

$$\mathbf{P}(X > 80) = P\left(\frac{X-70}{5} > \frac{80-70}{5}\right) = \mathbf{P}(Z > 2) = 1 - \mathbf{P}(Z \leq 2) = 1 - 0'97725 = 0'02275$$

□

6 Ejercicios

1. Supongamos que el 5% de los nidos de águila en la época de anidación no tienen ninguna cría, el 35% tienen una sola cría, el 50% tienen dos crías y el 10% tienen tres aguiluchos.

   (a) Se observa un nido de águila en la época de anidación, ¿cuál es la probabilidad de que tenga al menos dos aguiluchos?

   (b) Calcula la media o esperanza del número de crías en un nido de águila en época de anidación.

   (c) Calcula la varianza y la desviación típica del número de crías en un nido de águila en época de anidación.

2. Sea (X, Y) un vector aleatorio bidimensional con distribución de probabilidad:

	Y		
X	2	3	
20	0'2	0'1	
30	0'2	0'2	
40	0'1	0'2	

   (a) Calcula las distribuciones marginales de X e Y.

   (b) Calcula la distribución de Y condicionada a que X vale 30.

   (c) ¿Son X e Y independientes?

3. En el estudio de los hábitos migratorios de las cigüeñas, se ha anillado aproximadamente el 5% de la población total de estas aves. En un día determinado se capturan ocho cigüeñas. Sea X la variable aleatoria "número de cigüeñas anilladas".

   (a) Calcula la probabilidad de que al menos una cigüeña esté anillada.

   (b) Calcula la esperanza de X, esto es, el número esperado de cigüeñas anilladas.

4. Se sabe que la temperatura afecta a la probabilidad de germinación de cierto tipo de semillas. En concreto, a 10°C, la probabilidad de que una semilla germine es de $0'3$ y a 20°C la probabilidad de que germine es de $0'9$. Se ponen dos semillas en condiciones de germinar. Sea Y="número de semillas que germinan".

   (a) Calcula la distribución de probabilidad de Y, sabiendo que la temperatura es de 20°C.

   (b) Si la temperatura es una variable aleatoria, que denotamos por X, y con probabilidades respectivas de $0'4$, y $0'6$ para 10°C y 20°C, calcula la distribución conjunta del vector aleatorio (X, Y) y la distribución marginal de Y.

   (c) En las condiciones del apartado (b), calcula la media, la varianza y la desviación típica del número de semillas que germinan.

5. Una compañía de explotación petrolífera va a perforar 10 pozos. Cada uno de ellos tiene una probabilidad $0'1$ de producir petróleo en forma comercial. A la compañía le cuesta 100 millones de euros perforar cada pozo. Un pozo comercial saca petróleo por valor de 5.000 millones de euros.

   (a) Calcula la ganancia que espera obtener la compañía por los 10 pozos.

   (b) Calcula la desviación típica de las ganancias de la compañía.

6. Supongamos que se conoce que la longitud de los pétalos de una población de plantas de una especie determinada está normalmente distribuida con media $\mu = 3'2$ cm y desviación típica $\sigma = 0'8$ cm. ¿Qué proporción de la población podría esperarse que tenga una longitud de pétalos :

   (a) mayor que 4'0 cm?

   (b) mayor que 1'7 cm?

(c) entre 2'6 y 3'6 cm?

7. Supongamos que se corta una secuencia de ADN mediante una enzima de restricción y se obtienen fragmentos cuyo "tamaño" –longitud medida en pares de bases (*bp*)– es una variable aleatoria que se distribuye normalmente con media $\mu = 5000\ bp$ y desviación típica $\sigma = 900\ bp$.

   Calcula:

   (a) La probabilidad de obtener un fragmento mayor de 6000 bp.

   (b) El percentil 97'5. Interpreta su valor.

   (c) El cuantil 0'025. Interpreta su valor.

8. El contenido de calcio en sangre en la población general es una variable aleatoria con distribución normal cuya media vale $9'5$ mg/dl y cuya desviación típica es de $0'5$ mg/dl. Si se considera anómalo cualquier valor fuera del rango $8'5 - 10'2$ mg/dl.

   (a) ¿Cuál es la probabilidad de que un individuo tomado al azar de esta población tenga un contenido de calcio en sangre considerado como anómalo?

   (b) Si se toman diez individuos de la población general, ¿cuál es la probabilidad de que a lo sumo dos de ellos tenga un contenido de calcio anómalo?

9. El crecimiento de una planta depende del tipo de suelo (tipo A, tipo B) en el que se cultiva. En el suelo tipo A, la altura de las plantas se distribuye normalmente con media 15 cm y desviación típica 2 cm. En el suelo tipo B, la altura de las plantas se distribuye normalmente con media 20 cm y desviación típica 1'5 cm.

   (a) Si una planta se ha cultivado en suelo tipo A, ¿cuál es la probabilidad de que su altura sea superior a 21 cm?

   (b) Si una planta se ha cultivado en suelo tipo B, ¿cuál es la probabilidad de que su altura sea superior a 21 cm?

   (c) Supongamos que la proporción de plantas que se han cultivado en suelo tipo A es 0'5. Elegimos una planta al azar y su altura es superior a 21 cm. ¿Cuál es la probabilidad de haya sido cultivada en el suelo tipo A?

   (d) En un grupo de seis plantas cultivadas en suelo tipo B, ¿cuál es la probabilidad de que todas tengan altura superior a 21 cm?

10. Un cierto tipo de semilla tiene una probabilidad de germinación de 0'8. En una caja de 400 semillas denotamos por X el número de semillas que germinan.

    (a) Calcula la probabilidad de que germinen al menos tres cuartas partes de las semillas (300 semillas).

    (b) Calcula la probabilidad de que germinen a lo sumo el 85% de las semillas (340 semillas).