

## Parte I

# Estadística descriptiva

## Capítulo 2

# El tratamiento de los datos. Estadística descriptiva

Es un error capital el teorizar antes de poseer datos. Insensiblemente uno comienza a alterar los hechos para encajarlos en las teorías, en lugar encajar las teorías en los hechos

Sherlock Holmes (A. C. Doyle), en *Un escándalo en Bohemia*

**Resumen.** En este capítulo aprenderemos métodos para resumir y describir conjuntos de datos a través de distintos tipos de tablas, gráficos y medidas estadísticas.

**Palabras clave:** datos cuantitativos, datos cualitativos, datos discretos, datos continuos, distribución de frecuencias, diagrama de barras, diagrama de sectores, histograma, media, mediana, moda, cuantiles, varianza, desviación típica, asimetría, datos atípicos.

### 2.1. Introducción

Obtenidos a través de encuestas, experimentos o cualquier otro conjunto de medidas, los datos estadísticos suelen ser tan numerosos que resultan prácticamente inútiles si no son resumidos de forma adecuada. Para ello la Estadística utiliza tanto técnicas gráficas como numéricas, algunas de las cuales describimos en este capítulo.

Podemos decir que existe una clasificación, un tanto artificial, de los datos, según se refieran a una población tangible, en cuyo caso se conocerán todos los casos, o a una población conceptual, en cuyo caso sólo se conocerá una muestra (aleatoria simple). Sin embargo, esta clasificación no tiene ningún efecto en lo relativo a lo que vamos a estudiar en este capítulo.

### 2.2. Tipos de datos

Los datos (o variables) pueden ser de dos tipos: **cuantitativos** y **cualitativos**.

Los datos **cuantitativos** son los que representan una cantidad reflejada en una escala numérica. A su vez, pueden clasificarse como datos **cuantitativos discretos** si se refieren al conteo de alguna característica, o datos **cuantitativos continuos** si se refieren a una medida.

Los datos **cualitativos o categóricos** se refieren a características de la población que no pueden asociarse a cantidades con significado numérico, sino a características que sólo pueden clasificarse.

**Ejemplo.** Veamos algunos ejemplos de cada uno de estos tipos de variables:

- En el ejemplo del óxido de silicio, la variable *espesor* es cuantitativa continua.
- En el ejemplo de los cojinetes, el *diámetro de los cojinetes* es una variable cuantitativa continua.
- En el ejemplo de los niveles de plomo, se está analizando si una muestra contiene niveles detectables o no. Se trata, por tanto, de una variable cualitativa con dos categorías: *sí contiene niveles detectables* o *no contiene niveles detectables*.
- En el ejemplo de los accidentes laborales, la variable *número de accidentes laborales* es cuantitativa discreta, mientras que las franjas horarias constituyen una variable cualitativa.

### 2.3. Métodos gráficos y numéricos para describir datos cualitativos

La forma más sencilla de describir de forma numérica una variable cualitativa es determinar su distribución de frecuencias. Por su parte, esta distribución de frecuencias determina a su vez las representaciones gráficas más usuales.

Supongamos que tenemos una variable cualitativa, que toma una serie de posibles valores (categorías). El número de veces que se da cada valor es la **distribución de frecuencias** de la variable. Si en vez de dar el número de veces nos fijamos en la proporción de veces, tenemos la **distribución de frecuencias relativas**.

Las representaciones gráficas más usuales son los diagramas de barras y los diagramas de sectores.

Los **diagramas de barras** son una representación de cada una de las categorías de la variable mediante una barra colocada sobre el eje X y cuya altura sea la frecuencia o la frecuencia relativa de dichas categorías.

Los **diagramas de sectores** son círculos divididos en tantos sectores como categorías, sectores cuyo ángulo debe ser proporcional a la frecuencia de cada categoría.

Categoría	Frecuencia	Frecuencia relativa
País	Número de reactores nucleares	Proporción
Bélgica	4	0.041
Francia	22	0.225
Finlandia	2	0.020
Alemania	7	0.071
Holanda	1	0.010
Japón	11	0.112
Suecia	3	0.031
Suiza	1	0.010
Estados Unidos	47	0.480
TOTAL	98	1.000

Cuadro 2.1: Tabla de frecuencias.

**Ejemplo.** Tomamos como población los 98 reactores nucleares más grandes en todo el mundo. Nos fijamos en la variable o dato referente al país donde están localizados.

Los datos serían

[illegible]

Las distribuciones de frecuencias y de frecuencias relativas podemos resumirlas en una **tabla de frecuencias** como la que aparece en el Cuadro 2.1.

Por su parte, las representaciones mediante diagramas de barras y sectores de estos datos aparecen en la Figura 2.1 y la Figura 2.2 respectivamente.

## 2.4. Métodos gráficos para describir datos cuantitativos

Si tenemos una variable cuantitativa discreta y ésta toma pocos valores, podemos tratarla como si fuera una variable cualitativa, calcular su distribución de frecuencias y dibujar un diagrama de barras.

**Ejemplo.** En una empresa con cadena de montaje donde se empaquetan piezas en cajas se realiza un estudio sobre la calidad de producción. Los datos siguientes informan sobre el número de piezas defectuosas encontradas en una muestra de cajas examinadas:

0000001111111122222222223333333344444444555566666777889

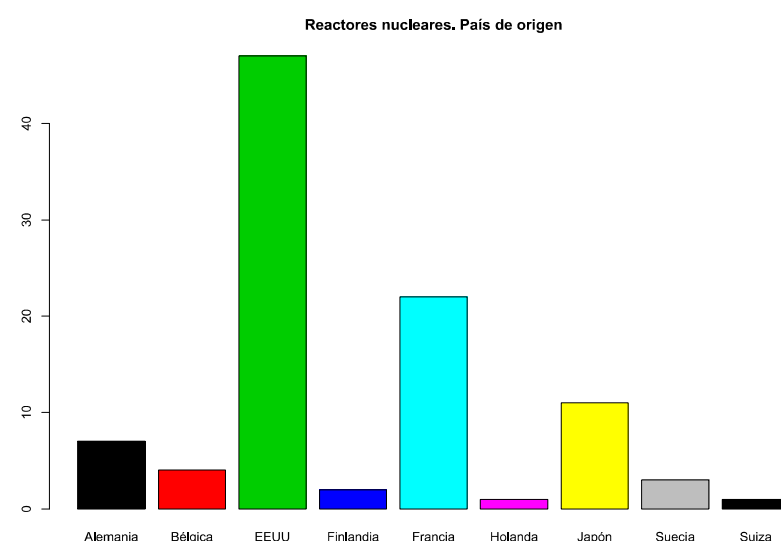


Figura 2.1: Diagrama de barras.

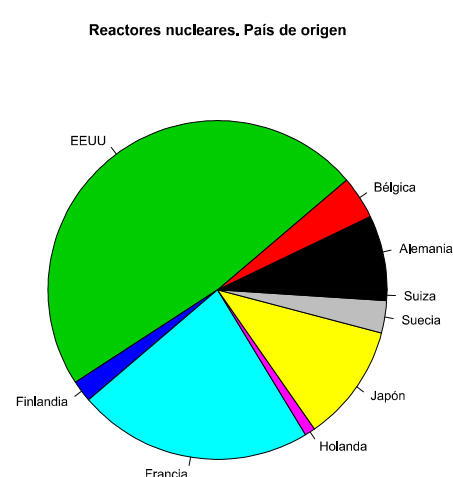


Figura 2.2: Diagrama de sectores.

El diagrama de barras asociado aparecen en la Figura 2.3.

Sin embargo, la mayoría de variables cuantitativas son de tipo continuo, de manera que toman demasiados valores como para que la representación de su distribución de frecuencias sea útil<sup>1</sup>. Por ello el método gráfico más común y tradicional para datos cuantitativos es el histograma.

El **histograma** es una variante del diagrama de barras donde se agrupan los valores de la variable en intervalos para que estos intervalos tengan frecuencias mayores que uno.

Para obtener un histograma de forma manual deben seguirse los siguientes pasos:

1. Calculamos el número,  $N$ , de intervalos que vamos a utilizar. Se recomienda que sea aproximadamente igual a la raíz cuadrada del número de datos. Sin embargo, los programas estadísticos suelen utilizar otro método, llamado *Método de Sturges*, en el que  $N = \lceil \log_2 n + 1 \rceil$ , donde  $n$  es el número de datos y  $\lceil \cdot \rceil$  es la función parte entera.

<sup>1</sup>Si toma muchos valores, muy probablemente la mayor parte de ellos sólo aparezca una vez, por lo que la distribución de frecuencias será casi siempre constante e igual a 1.

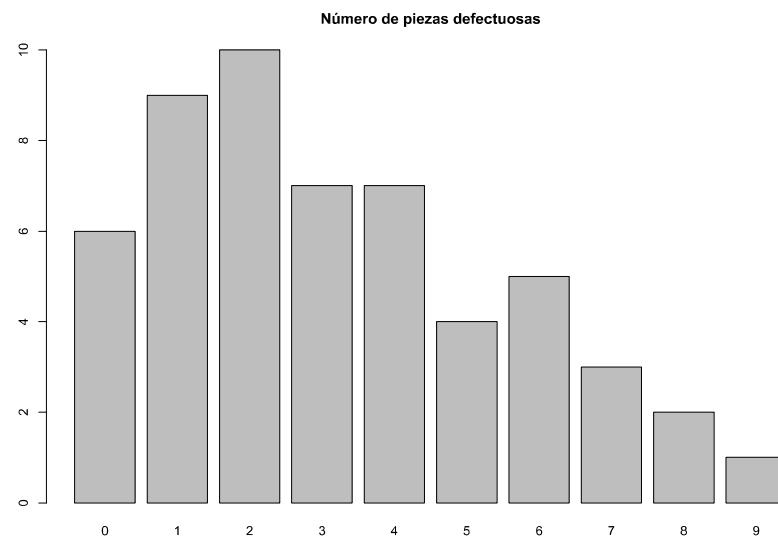


Figura 2.3: Diagrama de barras.

2. Calculamos el rango,  $R$ , del histograma, que será ligeramente más amplio que el rango de los datos. El histograma debe comenzar en un número ( $x_m$ ) ligeramente por debajo del mínimo de los datos y terminar en un número ( $x_M$ ) ligeramente por encima del máximo. El rango del histograma será, por tanto,  $R = x_M - x_m$ .
3. Calculamos la longitud,  $L$ , de los intervalos, como el cociente entre el rango del histograma y el número de intervalos, es decir,  $L = \frac{R}{N}$ .
4. Se construyen los  $N$  intervalos:

$$I_1 = [x_m, x_m + L)$$

$$I_2 = [x_m + L, x_m + 2L)$$

$$I_3 = [x_m + 2L, x_m + 3L)$$

...

$$I_N = [x_m + N \times L, x_M).$$

5. Para cada intervalo, contamos el número de datos que hay en él, es decir, la frecuencia del intervalo.
6. El histograma es un diagrama de barras donde en el eje X se colocan los intervalos y sobre ellos se construyen barras cuya altura sea la frecuencia o la frecuencia relativa del intervalo. En este caso, las barras deben dibujarse sin espacio entre ellas. En ocasiones, en vez de tomar la frecuencia relativa como altura de las barras, se toma dicha frecuencia relativa como área de las barras: en ese caso, se habla de un histograma en escala de densidad.

**Nota.** Por cuestiones que detallaremos más adelante es importante destacar que el porcentaje de datos que cae dentro de un intervalo es proporcional al área de la barra que se construye sobre ese intervalo. Por ejemplo, si el área de una barra es el 30 % del área total del intervalo, entonces el 30 % de los datos están en dicho intervalo.

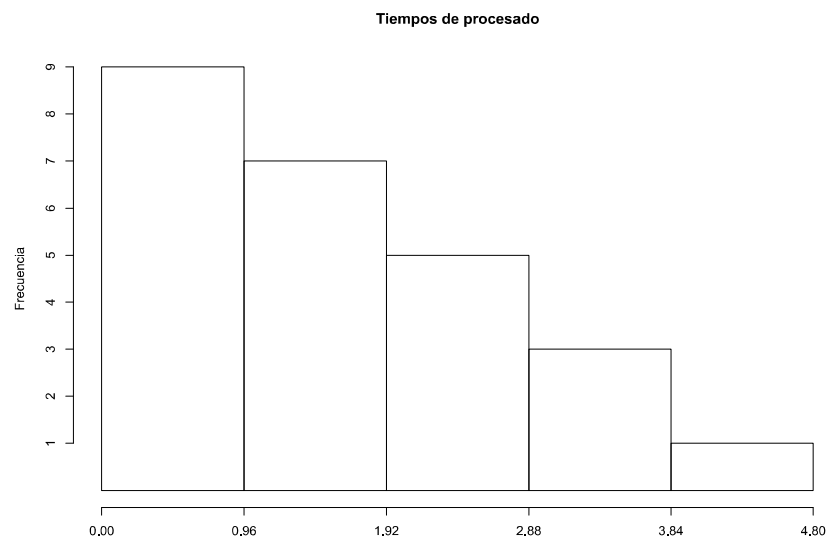


Figura 2.4: Histograma.

Por otra parte, ¿qué pasaría si tomamos un número muy grande de datos? El número de intervalos del histograma sería también muy grande, y las barras serían muy estrechas, de manera que en vez de parecer un diagrama de barras, parecería la gráfica de una función real de variable real. Hablaremos de esta función y del área debajo de ella en breve. Por cierto, ¿cómo se calcula el área bajo esta función?

**Ejemplo.** Los datos siguientes corresponden al tiempo necesario para procesar 25 trabajos en una CPU.

1.17	1.61	1.16	1.38	3.53	1.23	3.76	1.94	0.96	4.75
0.15	2.41	0.71	0.02	1.59	0.19	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.4					

Vamos a calcular un histograma para esos datos.

1. Dado que  $\sqrt{25} = 5$ , utilizaremos 5 intervalos.
2. El mínimo de los datos es 0.02 y el máximo 4.75, de manera que podemos considerar como rango del histograma el intervalo  $[0, 4.8]$ , cuya longitud (rango del histograma) es 4.8.
3. La longitud de los intervalos es, en ese caso,  $\frac{4.8}{5} = 0.96$ .
4. Construimos los intervalos:

$$I_1 = [0, 0.96)$$

$$I_2 = [0.96, 1.92)$$

$$I_3 = [1.92, 2.88)$$

$$I_4 = [2.88, 3.84)$$

$$I_5 = [3.84, 4.8)$$

5. Calculamos la distribución de frecuencia asociada a esos intervalos:

Tiempo de procesado	Frecuencia
[0, 0.96)	8
[0.96, 1.92)	8
[1.92, 2.88)	5
[2.88, 3.84)	3
[3.84, 4.8)	1

6. Finalmente, representamos el diagrama de barras (Figura 2.4).

## 2.5. Métodos numéricos para describir datos cuantitativos

Es cierto que un diagrama de barras o un histograma nos ayudan a tener una imagen de cómo son los datos, pero normalmente es necesario complementar esa imagen mediante medidas que, de forma objetiva, describan las características generales del conjunto de datos.

Vamos a ver en este apartado tres tipos de medidas, que básicamente responden a tres preguntas: *por dónde están los datos* (medidas de posición), *cómo de agrupados están los datos* (medidas de dispersión) y *qué forma tienen los datos* (medidas de forma).

### 2.5.1. Medidas de tendencia central

Las **medidas de tendencia central** son medidas de posición que tratan de establecer un valor que pueda considerarse *el centro* de los datos en algún sentido.

#### 2.5.1.1. Media

Sea un conjunto de datos de una variable cuantitativa,  $x_1, \dots, x_n$ . La **media** de los datos es

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

Esta medida es la más común dentro de las de tendencia central y corresponde al *centro de gravedad* de los datos.

Es inmediato comprobar que si se realiza un cambio de origen y escala sobre los datos, del tipo  $y = ax + b$ , la media sufre el mismo cambio, es decir,  $\bar{y} = a\bar{x} + b$ .

De igual forma, si tenemos datos de la suma de dos o más variables, la media de la suma es la suma de las medias de cada variable.



### 2.5.1.2. Mediana

Sea un conjunto de datos de una variable cuantitativa,  $x_1, \dots, x_n$ . Ordenemos la muestra de menor a mayor,  $x_{(1)}, \dots, x_{(n)}$ .

La **mediana** es el valor de la variable que deja el mismo número de datos antes y después que él, una vez ordenados estos.

El cálculo de la mediana dependerá de si el número de datos,  $n$ , es par o impar:

- Si  $n$  es impar, la mediana es el valor que ocupa la posición  $\frac{n+1}{2}$  una vez que los datos han sido ordenados (en orden creciente o decreciente), porque éste es el valor central. Es decir:  $M_e = x_{(\frac{n+1}{2})}$ .
- Si  $n$  es par, la mediana es la media aritmética de las dos observaciones centrales. Cuando  $n$  es par, los dos datos que están en el centro de la muestra ocupan las posiciones  $\frac{n}{2}$  y  $\frac{n}{2} + 1$ . Es decir:  $M_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2} + 1)}}{2}$ .

La mediana corresponde exactamente con la idea de valor central de los datos. De hecho, puede ser un valor más representativo de éstos que la media, ya que es más *robusta* que la media. Veámos qué significa esto en un ejemplo.

**Ejemplo.** Consideremos los datos siguientes:

0 0 1 2 3 4 5

Su media es  $\frac{0+0+1+2+3+4+5}{7} = 2.1429$ , y su mediana 2.

Pero imaginemos que por error o por casualidad obtenemos un nuevo dato enormemente grande en relación al resto de datos, 80. En ese caso, la media sería

$$\frac{0 + 0 + 1 + 2 + 3 + 4 + 5 + 80}{8} = 11.875$$

y la mediana 2.5. Es decir, un solo dato puede desplazar enormemente la media, hasta convertirla en una medida poco representativa, pero sólo desplazará ligeramente la mediana. Ese es el motivo por el que se dice que la mediana es una medida **robusta**.

### 2.5.1.3. Moda o intervalo modal

En principio la **moda** se define como el valor más frecuente de los datos. Lo que ocurre es que si éstos son datos de una variable continua o discreta con muchos valores, puede que los datos apenas se repitan. En ese caso, en el que, como vimos en las representaciones gráficas, se debe agrupar por intervalos, no debe darse un valor como moda, sino un **intervalo modal**, aquél con mayor frecuencia asociada.

### 2.5.2. Cuantiles

Los **cuantiles** son medidas de posición pero no necesariamente ligados al *centro* de los datos. La idea a la que responden es muy sencilla y muy práctica. Se trata de valorar de forma relativa cómo es un dato respecto del conjunto global de todos los datos.

Si, por ejemplo, un niño de 4 años pesa 13 kilos, ¿está desnutrido? ¿está sano? La respuesta debe ser que *depende*. ¿Dónde vive el niño? Es importante porque, por ejemplo, en Estados Unidos los niños son en general más grandes que, por ejemplo, en Japón. Quizá más que el peso nos interese saber qué posición relativa tiene el peso del niño dentro de la población de la que forma parte. Por ejemplo, si nos dicen que el niño está entre el 1 % de los niños que menos pesan, probablemente tiene un problema de crecimiento.

El **cuantil**  $p$  ( $Q_p$ ) de unos datos ( $0 \leq p \leq 1$ ), sería un valor de la variable situado de modo que el  $100p$  % de los valores sean menores o iguales que él y el resto ( $100(1 - p)$  %) mayores.

No obstante, en la práctica vamos a encontrar un problema para encontrar cuantiles, sobre todo con pocos datos: lo más habitual es que no exista el valor exacto que deje a la izquierda el  $100p$  % de los valores y el resto a la derecha. Por ese motivo, los programas estadísticos utilizan unas fórmulas de interpolación para obtener el valor del cuantil entre los dos valores de los datos que lo contienen. En nuestro caso, a la hora de obtener cuantiles, la aplicación de esas fórmulas de interpolación *a mano* harían muy lentos y pesados los cálculos, por lo que vamos a aplicar un convenio mucho más sencillo: aproximaremos el valor del cuantil correspondiente de la siguiente forma:

1. Si el  $100p$  % de  $n$ , donde  $n$  es el número de datos, es un entero,  $k$ , entonces  $Q_p = \frac{x_{(k)} + x_{(k+1)}}{2}$ .
2. Si el  $100p$  % de  $n$  no es un entero, lo redondeamos al entero siguiente,  $k$ , y entonces  $Q_p = x_{(k)}$ .

No olvidemos, sin embargo, que los programas estadísticos van a utilizar las fórmulas de interpolación para calcular el valor de los cuantiles, de manera que no debe extrañar si se observan pequeñas diferencias al comparar nuestros resultados *a mano* con los de estos programas.

Existen diversos nombres para referirse a algunos tipos de cuantiles. Entre ellos:

- Los **percentiles** son los cuantiles que dividen la muestra en 100 partes, es decir, son los cuantiles 0.01 (percentil 1), 0.02 (percentil 2), ..., 0.99 (percentil 99). Si notamos por  $P_\alpha$  al percentil  $\alpha$ , con  $\alpha = 1, 2, 3, \dots, 99$ , se tiene que  $P_\alpha = Q_{\alpha/100}$ . En Estadística Descriptiva es más frecuente hablar de percentiles que de cuantiles porque se refieren a cantidades entre 0 y 100, en tanto por ciento, que son más habituales de valorar por todo el mundo.
- Los **cuartiles** dividen a la población en cuatro partes iguales, es decir, corresponden a los cuantiles 0.25, 0.5 (mediana) y 0.75.

**Ejemplo.** Consideremos de nuevo los datos correspondientes al tiempo de procesado de 25 tareas en una CPU. Ahora los hemos ordenado de menor a mayor (en 5 filas):

0.02	0.75	1.17	1.61	2.59
0.15	0.82	1.23	1.94	3.07
0.19	0.92	1.38	2.01	3.53
0.47	0.96	1.40	2.16	3.76
0.71	1.16	1.59	2.41	4.75

Vamos a calcular distintas medidas de posición y a comentarlas.

En primer lugar, la media es 1.63. La mediana ocupa el lugar 13 en la muestra ordenada, y su valor es 1.38. Obsérvese que la media es algo mayor que la mediana: esto es debido a la presencia de algunos valores significativamente más altos que el resto, como pudimos ver en el histograma.

Por su parte, el  $P_{25}$  o cuantil 0.25 ocupa la posición 7, ya que el 25 % de 25 es 6.25. Por tanto,  $P_{25} = 0.82$ . De igual forma,  $P_{75} = Q_{0.75} = 2.16$ , el valor que ocupa la posición 19. Podemos ver, por tanto, que los valores más bajos están muy agrupados al principio, y se van dispersando más conforme se hacen más altos.

### 2.5.3. Medidas de variación o dispersión

Las **medidas de variación o dispersión** están relacionadas con las medidas de tendencia central, ya que lo que pretenden es cuantificar cómo de concentrados o dispersos están los datos respecto a estas medidas. Nosotros nos vamos a limitar a dar medidas de dispersión asociadas a la media.

La idea de estas medidas es valorar en qué medida los datos están agrupados en torno a la media. Esta cuestión tan simple es uno de los motivos más absurdos de la mala prensa que tiene la Estadística en la sociedad en general. La gente no se fía de lo que ellos llaman *la Estadística* entre otros motivos, porque parece que todo el mundo cree que una media tiene que ser un valor válido para todos, y eso es materialmente imposible.

**Ejemplo.** Pensemos en la media del salario de los españoles. En 2005 fue de 18.750 euros al año. Ahora bien, esa media incluye tanto a las regiones más desarrolladas como a las más desfavorecidas y, evidentemente, la cifra generará mucho malestar en gran parte de la población (con toda seguridad, más del 50 %), cuyo salario está por debajo.

**Ejemplo.** Existe una frase muy conocida que dice que “*la Estadística es el arte por el cuál si un español se come un pollo y otro no se come ninguno, se ha comido medio pollo cada uno*”. Esa frase se usa en muchas ocasiones para ridiculizar a la Estadística, cuando en realidad debería servir para desacreditar a quien la dice, por su ignorancia.

Hay que decir que la Estadística no tiene la culpa de que la gente espere de una media más de lo que es capaz de dar, ni de que muy poca gente conozca medidas de dispersión asociadas a la media.

#### 2.5.3.1. Varianza muestral

Dados unos datos de una variable cuantitativa,  $x_1, \dots, x_n$ , la **varianza muestral**<sup>2</sup> de esos datos es

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

**Nota.** Para calcular *a mano* la varianza resulta más cómodo desarrollar un poco su fórmula, como vamos a ver:

$$\begin{aligned} s_{n-1}^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2}{n-1} \\ &= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}. \end{aligned}$$

Cuanto mayor sea la varianza de unos datos, más dispersos, heterogéneos o variables son esos datos. Cuanto más pequeña sea una varianza de unos datos, más agrupados u homogéneos son dichos datos.

**Ejemplo.** Una muestra aleatoria simple de la altura de 5 personas arroja los siguientes resultados:

1.76   1.72   1.80   1.73   1.79

Calculemos su media y su varianza muestral.

Lo único que necesitamos es  $\sum_{i=1}^5 x_i = 8.8$  y  $\sum_{i=1}^5 x_i^2 = 15.493$ . A partir de estos datos,

$$\bar{x} = \frac{8.8}{5} = 1.76$$

y

$$s_{n-1}^2 = \frac{15.493 - 5 \times 1.76^2}{4} = 0.00125$$

En lo que respecta al comportamiento de la varianza muestral frente a cambios de origen y escala, sólo le afectan los segundos. Es decir, si tenemos que  $y = ax + b$ , se verifica que  $s_{y;n-1}^2 = a^2 s_{x;n-1}^2$ .

Finalmente, si bien habíamos comentado que en el caso de la media, si tenemos la suma de varias variables, la media total es la suma de las medias de cada variable, no ocurre así con la varianza en general.

#### 2.5.3.2. Desviación típica o estandar muestral

El principal problema de la varianza es su unidad de medida. Por cómo se define si, por ejemplo, la variable se expresa en kilos, la media también se expresa en kilos, pero la varianza se expresa en kilos<sup>2</sup>, lo que hace que sea difícil valorar si una varianza es muy elevada o muy pequeña.

Es por ello que se define la **desviación típica o estandar muestral** de los datos como  $s_{n-1} = \sqrt{s_{n-1}^2}$ , cuya unidad de medida es la misma que la de la media.

**Nota.** La Regla Empírica

Si el histograma asociado a unos datos tiene la forma de una campana o de una joroba, el conjunto de datos tendrá las siguientes características, lo que en algunos libros se conoce como **Regla Empírica**:

1. Aproximadamente el 68 % de los datos estará en el intervalo  $(\bar{x} - s_{n-1}, \bar{x} + s_{n-1})$ .
2. Aproximadamente el 95 % de los datos estará en el intervalo  $(\bar{x} - 2s_{n-1}, \bar{x} + 2s_{n-1})$ .
3. Casi todos los datos estarán en el intervalo  $(\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1})$ .

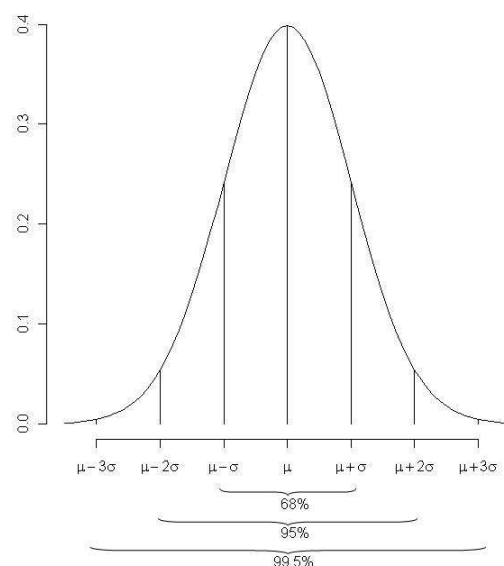


Figura 2.5: Representación gráfica de la regla empírica.

### 2.5.3.3. Coeficiente de variación

Como acabamos de decir, debemos proporcionar cada media junto con alguna medida de dispersión, preferentemente la desviación típica. Una forma de valorar en términos relativos cómo es de dispersa una variable es precisamente proporcionar el cociente entre la desviación típica y la media (en valor absoluto), lo que se conoce como **coeficiente de variación**.

Dado un conjunto de datos de media  $\bar{x}$  y desviación típica  $s_{n-1}$ , se define su **coeficiente de variación** como

$$CV = \frac{s_{n-1}}{|\bar{x}|}.$$

La principal ventaja del coeficiente de variación es que no tiene unidades de medida, lo que hace más fácil su interpretación.

**Ejemplo.** Para los datos de tiempo de procesado en una CPU de 25 tareas, la varianza es 1.42, luego su desviación estandar es 1.19, y el coeficiente de variación  $\frac{1.19}{1.63} = 0.73$ . Por tanto, la desviación estándar es algo más del 70 % de la media. Esto indica que los datos no están muy concentrados en torno a la media, probablemente debido a la presencia de los valores altos que hemos comentado antes.

**Nota.** El coeficiente de variación, tal y como está definido, sólo tiene sentido para conjuntos de datos con el mismo signo, es decir, todos positivos o todos negativos. Si hubiera datos de distinto signo, la media podría estar próxima a cero o ser cero, imposibilitando que aparezca en el denominador.

**Nota.** Suele ser frecuente el error de pensar que el coeficiente de variación no puede ser mayor que 1, lo cual es rigurosamente falso. Si lo expresamos en porcentaje, el coeficiente de variación puede ser superior al 100 % sin más que la desviación típica sea mayor que la media, cosa bastante frecuente, por cierto.

**Nota.** A la hora de interpretar el coeficiente de variación inmediatamente surge la pregunta de *¿cuándo podemos decir que es alto y cuándo que es bajo?* Realmente, no existe una respuesta precisa, sino que depende del contexto de los datos que estemos analizando. Si, por ejemplo, estamos analizando unos datos que por su naturaleza deben ser muy homogéneos, un coeficiente de variación del 10 % sería enorme, pero si por el contrario estamos analizando datos que por su naturaleza son muy variables, un coeficiente de variación del 10 % sería muy pequeño.

Por todo ello, lo recomendable es analizar el coeficiente de variación entendiendo su significado numérico, es decir, entendiendo que se refiere a la comparación de la desviación típica con la media, e interpretando su valor en relación al contexto en el que estemos trabajando.

#### 2.5.4. Medidas de forma. Coeficiente de asimetría

Las **medidas de forma** comparan la forma que tiene la representación gráfica, bien sea el histograma o el diagrama de barras de la distribución, con una situación *ideal* en la que los datos se reparten en igual medida a la derecha y a la izquierda de la media.

Esa situación en la que los datos están repartidos de igual forma a uno y otro lado de la media se conoce como **simetría**, y se dice en ese caso que la distribución de los datos es simétrica. En ese caso, además, su mediana, su moda y su media coinciden.

Por contra, se dice que una distribución es **asimétrica a la derecha** si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda. Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es **asimétrica a la izquierda**.

Para valorar la simetría de unos datos se suele utilizar el **coeficiente de asimetría de Fisher**:

$$As = \frac{\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n-1}}{s_{n-1}^3}.$$

Obsérvese que para evitar el problema de la unidad y hacer que la medida sea escalar y por lo tanto relativa, dividimos por el cubo de su desviación típica. De esta forma podemos valorar si unos datos son más o menos simétricos que otros, aunque no estén medidos en la misma unidad de medida. La interpretación de este coeficiente de asimetría es la siguiente:

- Tanto mayor sea el coeficiente en valor absoluto, más asimétricos serán los datos.
- El signo del coeficiente nos indica el sentido de la asimetría:
  - Si es positivo indica que la asimetría es a la derecha.
  - Si es negativo, indica que la asimetría es a la izquierda.

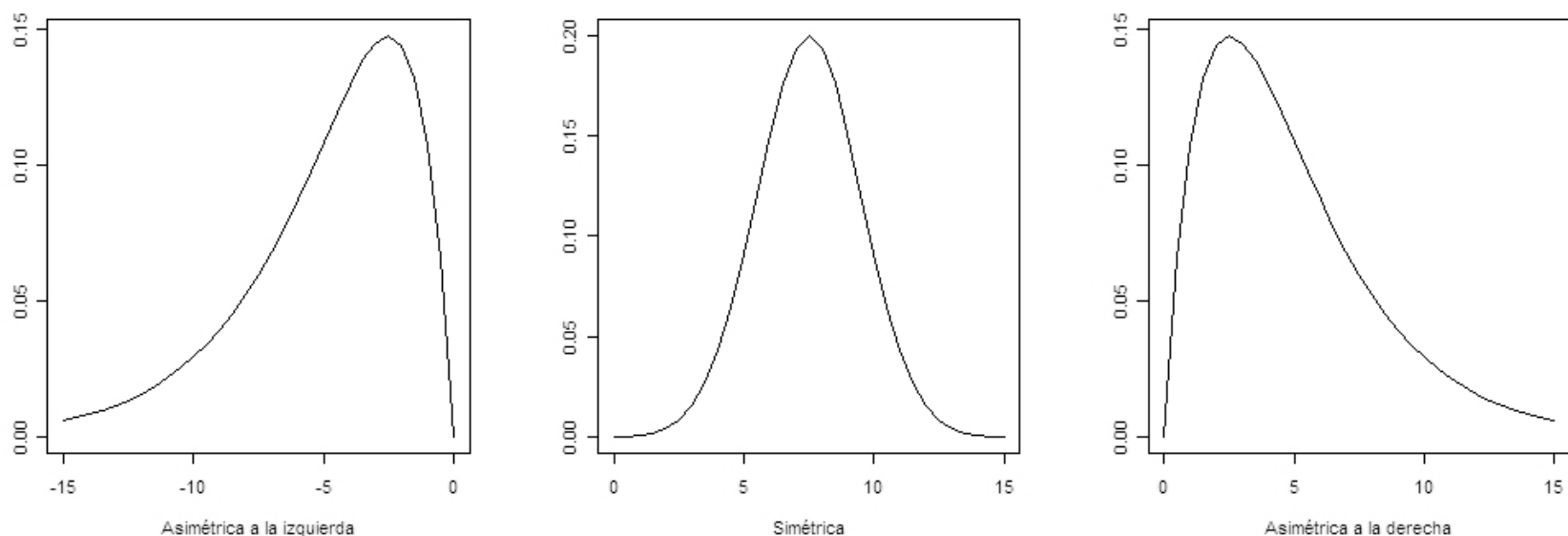


Figura 2.6: Formas típicas de distribuciones de datos.

**Ejemplo.** Para los datos de tiempo de procesado en una CPU de 25 tareas, el coeficiente de asimetría de Fisher es 0.91, lo que, como habíamos visto y comentado con anterioridad, pone de manifiesto que la distribución es asimétrica a la derecha, debido a la presencia de tiempos de procesado bastante altos en relación al resto.

### 2.5.5. Parámetros muestrales y parámetros poblacionales

Cuando se trabaja con una muestra de una población, ya sea ésta tangible o conceptual, las distintas medidas de posición, dispersión y forma, se denominan **parámetros muestrales**. Hay que tener en cuenta que prácticamente siempre se trabaja con muestras, ya que o bien trabajamos con poblaciones conceptuales o con poblaciones tangibles (finitas, por tanto), pero con muchísimos elementos.

Frente a estos parámetros muestrales se encuentran los parámetros análogos referidos a toda la población. Estos parámetros, llamados **parámetros poblacionales**, son, en general, imposibles de conocer<sup>3</sup>. Por ejemplo, la media poblacional se calcularía igual que la media muestral de unos datos, pero aplicada la fórmula a todos los elementos de la población. Como eso es prácticamente imposible de poner en la práctica, veremos

<sup>3</sup>Salvo en el caso de poblaciones finitas con pocos elementos.

en capítulos posteriores que los parámetros muestrales se utilizan en la práctica para aproximar o estimar los parámetros poblacionales.

## 2.6. Métodos para detectar datos cuantitativos atípicos o fuera de rango

Hay ocasiones en que un conjunto de datos contiene una o más observaciones *inconsistentes* en algún sentido. Por ejemplo, en los datos de tiempo de procesado en una CPU de 25 tareas, supongamos que tenemos una observación más, igual a 85, debido a que la CPU se bloqueó y hubo que reiniciarla. Este dato, que probablemente no deseemos incluir, es un ejemplo de caso de dato atípico o valor fuera de rango.

En general, una observación que es inusualmente grande o pequeña en relación con los demás valores de un conjunto de datos se denomina **dato atípico o fuera de rango**.

Estos valores son atribuibles, por lo general, a una de las siguientes causas:

1. El valor ha sido introducido en la base de datos incorrectamente.
2. El valor proviene de una población distinta a la que estamos estudiando.
3. El valor es correcto pero representa un suceso muy poco común.

A continuación vamos a proponer dos maneras de determinar si un dato es un valor fuera de rango.

### 2.6.1. Mediante la regla empírica

Este método es adecuado si el histograma de los datos tiene forma de campana, en cuyo caso podemos aplicar la regla empírica para detectar qué datos están fuera de los rangos *lógicos* según esta regla.

Según ella, el 99.5 % de los datos están en el intervalo  $[\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1}]$ , luego *se considerarán datos atípicos los  $x_i$  que no pertenezcan al intervalo  $[\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1}]$ .*

### 2.6.2. Mediante los percentiles

Supongamos que tenemos un conjunto de datos  $x_1, \dots, x_n$ . El procedimiento es el siguiente:

1. Se calculan los cuartiles primero y tercero, es decir, los percentiles 25 y 75,  $P_{25}$  y  $P_{75}$ . Se calcula el llamado *rango intercuartílico* ( $IR$  o  $RI$ ),  $IR = P_{75} - P_{25}$ .
2. Se consideran **datos atípicos** aquellos inferiores a  $P_{25} - 1.5IR$  o superiores a  $P_{75} + 1.5IR$ .



	Medias	Desv. Típica	CV	Coef. Asimetría
Serie 1	92.01	3.62	25.40	-1.79
Serie 2	92.74	3.73	24.86	1.71

Cuadro 2.2: Resumen descriptivo de los datos de las placas de silicio

**Ejemplo.** Vamos a ver si hay algún dato atípico entre los datos de tiempo de procesado en una CPU de 25 tareas.

Dado que el histograma no tenía forma de campana, el método de la regla empírica no es el método más adecuado para la detección de valores atípicos.

Por su parte,  $P_{50} = 1.38$ ,  $P_{25} = 0.82$  y  $P_{75} = 2.16$ . Por tanto,  $IR = 2.16 - 0.82 = 1.34$ , y el intervalo fuera del cual consideramos valores fuera de rango es  $[0.82 - 1.5 \times 1.34, 2.16 + 1.5 \times 1.34] = [-1.19, 4.17]$ . De esta forma, el valor 4.75 es un valor fuera de rango.

Hay una versión gráfica de este método para detectar valores atípicos mediante los percentiles: se llama **diagrama de caja o diagrama de cajas y bigotes** o (en inglés) **boxplot**. Este diagrama incluye en un gráfico:

1. El valor de la mediana (o segundo cuartil,  $Q_2$ ): ese es el centro de la caja.
2. El valor de los percentiles 25 y 75, cuartiles primero y tercero respectivamente ( $Q_1$  y  $Q_3$ ): son los lados inferior y superior de la caja.
3. El diagrama no representa los límites  $P_{25} - 1.5 \times IR$  y  $P_{75} + 1.5 \times IR$ . En su lugar, señala los últimos puntos no atípicos por debajo ( $L_i$ ) y por encima ( $L_s$ ), es decir, señala el último dato por encima de  $P_{25} - 1.5 \times IR$  y el último dato por debajo de  $P_{75} + 1.5 \times IR$ , y los representa como *bigotes* que salen de la caja.
4. Normalmente representa con círculos los datos atípicos.

## 2.7. Sobre el ejemplo de las capas de dióxido de silicio

Ya estamos en condiciones de responder en parte a las cuestiones que quedaron latentes en el tema de introducción sobre el ejemplo de las placas de silicio.

Vamos a comenzar realizando un resumen descriptivo de los datos, separando por series, proporcionando media, desviación típica, coeficiente de variación y coeficiente de asimetría. Todos estos resultados aparecen en la Tabla 2.2.

En primer lugar, es cierto que, como apuntábamos en el tema de introducción, los valores están en torno a 90 (la media es 92 más o menos). Además, vemos que sí que hay una variabilidad moderada de los datos, con un CV en torno al 25 %, lo que indica que, al parecer, las distintas condiciones en que cada medición se realizó, afectaron en alguna medida el resultado: todo esto es muy preliminar porque no tenemos la información completa de en qué condiciones se realizaron cada una de las mediciones. Por el contrario, podemos observar algo muy llamativo. Los datos de la primera serie son claramente asimétricos a la izquierda (coeficiente de

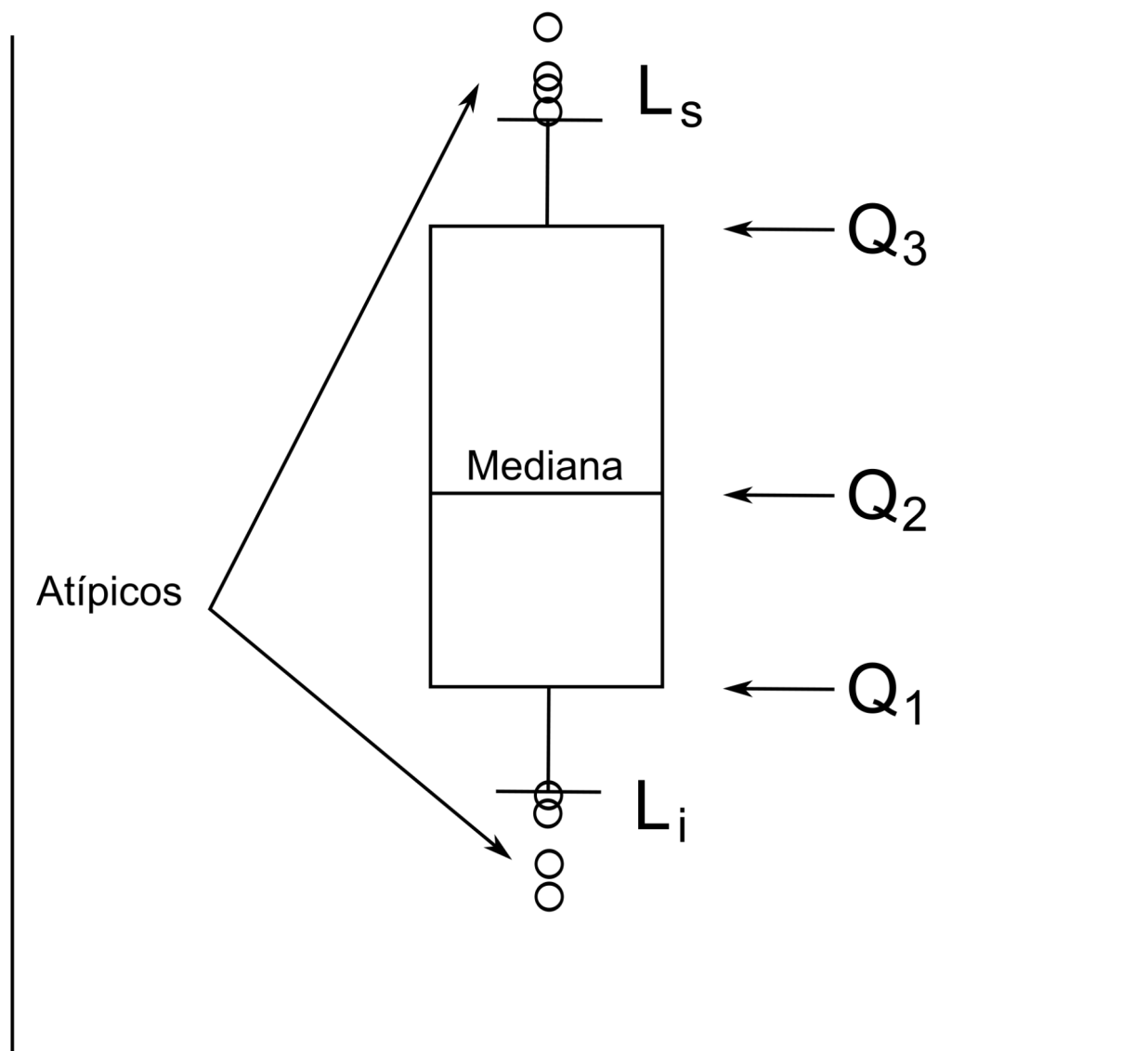


Figura 2.7: Descripción de un diagrama de caja. Fuente: [http://es.wikipedia.org/wiki/Diagrama\\_de\\_caja](http://es.wikipedia.org/wiki/Diagrama_de_caja)

asimetría de -1.79), mientras que los de la segunda serie son claramente asimétricos a la derecha (coeficiente de asimetría de 1.71). Dado que no era esperable que surgieran diferencias entre las dos series, debemos preguntarnos qué pasó.

Para tratar de analizar más profundamente los datos, vamos a proporcionar también los dos diagramas de caja de ambas series. Aparecen en la Figura 2.8. Con ellas, vamos a resumir ahora las decisiones que los autores tomaron en vista de los resultados y las conclusiones a las que llegaron.

Obsérvese que las diferencias entre las series no afectan sorprendentemente al conjunto de las muestras, sino sólo a los valores atípicos que se ven en ambos diagramas de caja. Eso *probaría* que, en efecto, no hay ninguna diferencia sistemática entre las series.

La siguiente tarea es la de inspeccionar los datos atípicos. Si miramos con atención los datos, vemos que las 8 mediciones más grandes de la segunda serie ocurrieron en la placa 10. Al ver este hecho, los autores del trabajo inspeccionaron esta placa y descubrieron que se había contaminado con un residuo de la película, lo que ocasionó esas mediciones tan grandes del espesor. De hecho, los ingenieros eliminaron esa placa y toda la serie entera por razones técnicas. En la primera serie, encontraron también que las tres mediciones más bajas se habían debido a un calibrador mal configurado, por lo que las eliminaron. No se pudo determinar causa alguna a la existencia de los dos datos atípicos restantes, por lo que permanecieron en el análisis. Por último, nótese que después de este proceso de depuración de los datos que el análisis mediante Estadística Descriptiva ha motivado, la distribución de los datos tiene una evidente forma de campana.

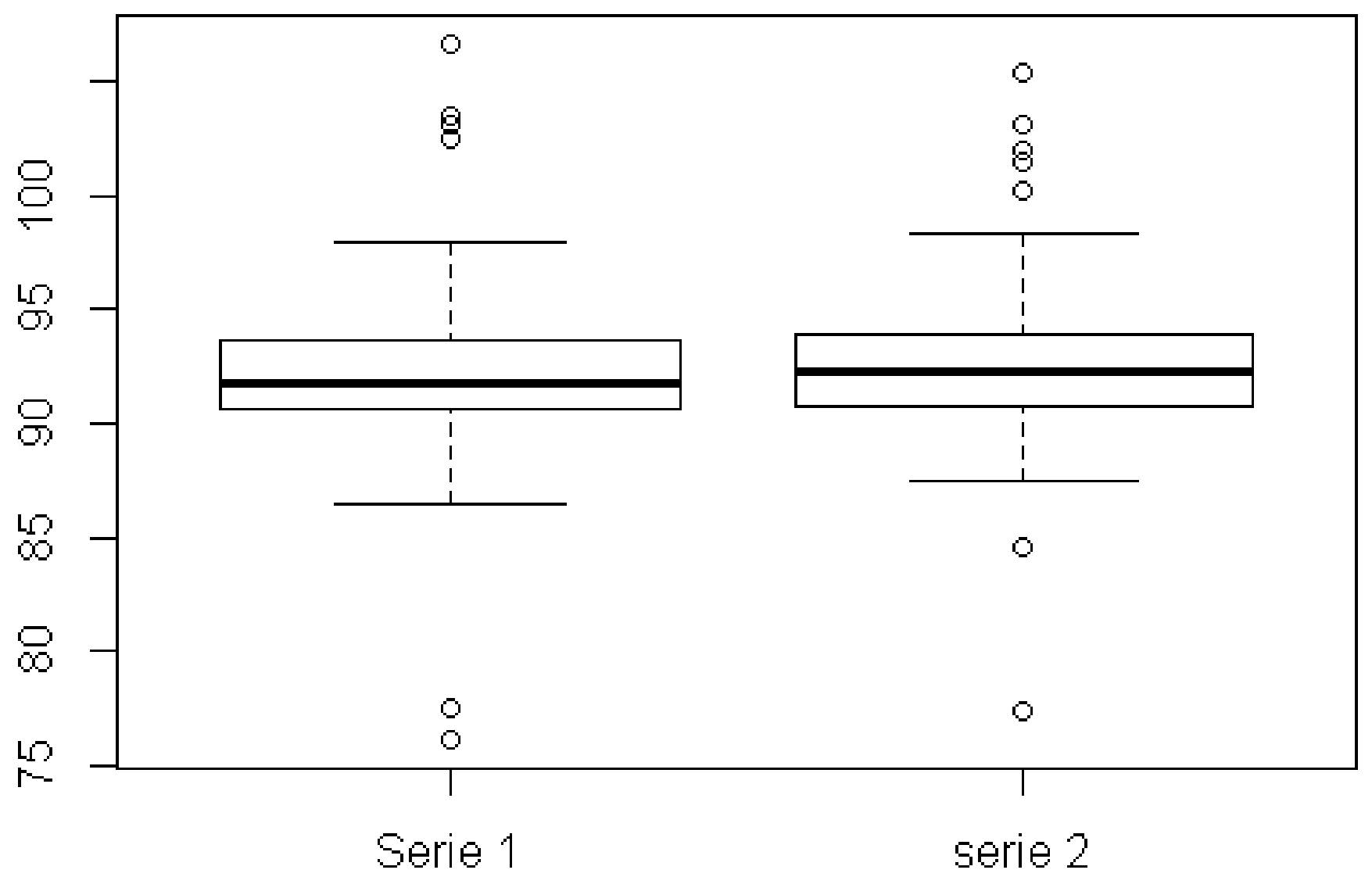


Figura 2.8: Diagramas de caja de los datos del espesor de las capas de dióxido de silicio

## Parte II

# Cálculo de Probabilidades

## Capítulo 3

# Probabilidad

Vemos que la teoría de la probabilidad en el fondo sólo es sentido común reducido a cálculo; nos hace apreciar con exactitud lo que las mentes razonables toman por un tipo de instinto, incluso sin ser capaces de darse cuenta[...] Es sorprendente que esta ciencia, que surgió del análisis de los juegos de azar, llegara a ser el objeto más importante del conocimiento humano[...] Las principales cuestiones de la vida son, en gran medida, meros problemas de probabilidad.

Pierre Simon, Marqués de Laplace

**Resumen.** El capítulo proporciona un tratamiento de los experimentos cuyos resultados no se pueden predecir con certeza a través del concepto de probabilidad. Se analizan las propiedades de la probabilidad y se introduce también el concepto de probabilidad condicionada, que surge cuando un suceso modifica la asignación de probabilidades previa.

**Palabras clave:** experimento aleatorio, experimento determinístico, espacio muestral, suceso, probabilidad, probabilidad condicionada, independencia de sucesos.

### 3.1. Introducción

En nuestra vida cotidiana asociamos usualmente el concepto de **Probabilidad** a su calificativo **probable**, considerando **probables** aquellos eventos en los que tenemos un alto grado de creencia en su ocurrencia. En esta línea, **Probabilidad** es un concepto asociado a la medida del **azar**. También pensamos en el azar vinculado, fundamentalmente, con los juegos de azar, pero desde esa óptica tan reducida se nos escapan otros muchísimos ejemplos de fenómenos de la vida cotidiana o asociados a disciplinas de distintas ciencias donde el azar juega un papel fundamental. Por citar algunos:

- ¿Qué número de unidades de producción salen cada día de una cadena de montaje? No existe un número fijo que pueda ser conocido a priori, sino un conjunto de posibles valores que podrían darse, cada uno de ellos con un cierto grado de certeza.
- ¿Cuál es el tamaño de un paquete de información que se transmite a través de HTTP? No existe en realidad un número fijo, sino que éste es desconocido a priori.

- ¿Cuál es la posición de un objeto detectado mediante GPS? Dicho sistema obtiene, realmente, una estimación de dicha posición, pero existen márgenes de error que determinan una región del plano donde el objeto se encuentra con alta probabilidad.
- ¿Qué ruido se adhiere a una señal que se envía desde un emisor a un receptor? Dependiendo de las características del canal, dicho ruido será más o menos relevante, pero su presencia no podrá ser conocida a priori, y deberá ser diferenciada de la señal primitiva, sin que se conozca ésta, teniendo en cuenta que se trata de un ruido *aleatorio*.

En todos estos ejemplos el azar es un factor insoslayable para conocer el comportamiento del fenómeno en estudio.

## 3.2. Experimentos aleatorios y experimentos determinísticos

En general, un experimento del que se conocen todos sus posibles resultados y que, repetido en las mismas condiciones, no siempre proporciona los mismos resultados se conoce como **experimento aleatorio**.

En contraposición, un **experimento determinístico** es aquel donde las mismas condiciones aseguran que se obtengan los mismos resultados.

Lo que el Cálculo de Probabilidades busca es encontrar una medida de la incertidumbre o de la certidumbre que se tiene de todos los posibles resultados, ya que jamás (o muy difícilmente) se podrá conocer a priori el resultado de cualquier experimento donde el azar esté presente: a esta medida de la incertidumbre la denominaremos *probabilidad*<sup>1</sup>.

## 3.3. Definición de probabilidad

Tenemos, por tanto, que probabilidad es la asignación que hacemos del grado de creencia que tenemos sobre la ocurrencia de algo. Esta asignación, sin embargo, debe ser *coherente*. Esta necesidad de que asignemos probabilidades adecuadamente se va a plasmar en esta sección en tres reglas, conocidas como *axiomas*, que debe cumplir cualquier reparto de probabilidades.

### 3.3.1. Álgebra de conjuntos

Si consideramos un experimento aleatorio, podemos caracterizar los posibles resultados de dicho experimento como conjuntos. Es de interés, por tanto, repasar los conceptos y propiedades básicas del álgebra de conjuntos. En todo este apartado no debemos olvidar que los conjuntos representan en nuestro caso los posibles resultados de un experimento aleatorio.

Un **conjunto** es una colección de elementos.

Se dice que  $B$  es un **subconjunto de**  $A$  si todos sus elementos lo son también de  $A$ , y se notará  $B \subset A$ .

<sup>1</sup>Es mejor que aceptemos desde el principio que la Estadística no es la ciencia de la adivinación: tan sólo se ocupa de cuantificar cómo de incierto es un evento y, ocasionalmente, de proponer estrategias de predicción basadas en dicha medida de la incertidumbre.

### *Apuntes de Estadística para Ingenieros*

Para cada  $A$  se verifica  $\emptyset \subset A \subset A \subset \Omega$ .

Si  $C \subset B$  y  $B \subset A$ , entonces,  $C \subset A$ . Esto se conoce como propiedad transitiva.

La **unión** de  $B$  y  $A$  es un conjunto cuyos elementos son los elementos de  $A$  y  $B$ , y se nota  $A \cup B$ . Esta operación verifica la propiedad conmutativa y asociativa.

Si  $A \subset B$ , entonces  $A \cup B = B$ .

La **intersección** de  $A$  y  $B$  es el conjunto formado por los elementos comunes de  $A$  y  $B$ , y se nota  $AB$  o  $A \cap B$ . Esta operación verifica la propiedad conmutativa, asociativa y distributiva respecto de la unión. Dos conjuntos,  $A$  y  $B$ , se dicen **mutuamente excluyentes, disjuntos o incompatibles** si su intersección es vacía, es decir,  $A \cap B = \emptyset$ .

Si dos conjuntos  $A$  y  $B$  son disjuntos, su unión suele notarse  $A + B$ .

Los conjuntos  $A_1, \dots, A_N$  se dicen **mutuamente excluyentes** si  $A_i \cap A_j = \emptyset$  para todo  $i \neq j$ .

Una **partición** es una colección de conjuntos,  $A_1, \dots, A_N$  tal que:

a)  $A_1 \cup \dots \cup A_N = \Omega$

b)  $A_i \cap A_j = \emptyset$  para todo  $i \neq j$ .

El **conjunto complementario** de un conjunto  $A$ ,  $\bar{A}$  ó  $A^c$ , está formado por todos los elementos de  $\Omega$  que no pertenecen a  $A$ .

Se sigue por tanto,

$$A \cup \bar{A} = \Omega$$

$$A \cap \bar{A} = \emptyset$$

$$(A^c)^c = A$$

$$\bar{\Omega} = \emptyset$$

$$\text{Si } B \subset A \rightarrow \bar{A} \subset \bar{B}$$

$$\text{Si } A = B \rightarrow \bar{A} = \bar{B}.$$

Finalmente, mencionemos las llamadas Leyes de Morgan:

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}.$$

### **3.3.2. Espacio muestral**

Consideremos un experimento aleatorio.

El conjunto formado por todos los posibles resultados del experimento aleatorio recibe el nombre de **espacio muestral**, y lo notaremos habitualmente como  $\Omega$ .

Cualquier subconjunto de un espacio muestral recibe el nombre de **suceso** o **evento**.

Hablaremos de **ensayo** o **realización** de un experimento aleatorio refiriéndonos a una ejecución de dicho experimento. Así, diremos que en un ensayo **ocurre un suceso**  $A$  si se observa en dicho ensayo cualquier resultado incluido en el suceso  $A$ .

Una observación importante es que el espacio muestral no tiene por qué ser único, sino que dependerá de lo que deseemos observar del experimento aleatorio. Vamos a poner este hecho de manifiesto en los siguientes ejemplos.

**Ejemplo.** Si consideramos el lanzamiento de un dado, un espacio muestral sería  $\Omega = \{1, 2, 3, 4, 5, 6\}$ .

Los sucesos más elementales posibles son  $\{1\}$ ,  $\{2\}$ ,  $\{3\}$ ,  $\{4\}$ ,  $\{5\}$  y  $\{6\}$ . Otros sucesos no elementales pueden ser  $\{1, 2\}$ ,  $\{\text{mayor que } 2\}$ ,  $\{\text{par}\}$ , ...

Sin embargo, supongamos que estamos lanzando un dado porque no tenemos ninguna moneda a mano, y sólo deseamos ver si el resultado es par o impar. En ese caso, el espacio muestral sería  $\Omega = \{\text{par}, \text{impar}\}$ .

**Ejemplo.** Un experimento habitual en Biología consiste en extraer, por ejemplo, peces de un río, hasta dar con un pez de una especie que se desea estudiar. El número de peces que habría que extraer hasta conseguir el ejemplar deseado de la especie en estudio formaría el espacio muestral,  $\Omega = \{1, 2, 3, \dots\}$ , si es que el investigador desea observar exactamente el número de peces hasta extraer ese ejemplar deseado. Obsérvese que se trata de un conjunto no acotado, pero numerable.

Como ejemplos de posibles sucesos de interés podríamos poner los eventos  $\{1, 2, 3, 4, 5\}$ ,  $\{\text{mayor o igual a } 5\}$ , ...

Supongamos ahora que el investigador sólo está interesado en comprobar si hacen falta más de 5 extracciones para obtener un ejemplar de la especie en estudio. En ese caso, el espacio muestral sería  $\Omega = \{> 5, \leq 5\}$ .

**Ejemplo.** Si consideramos el experimento aleatorio consistente en elegir un número absolutamente al azar entre 0 y 1, un espacio muestral sería  $\Omega = [0, 1]$ . A diferencia de los anteriores ejemplos, este espacio muestral no es finito, ni siquiera numerable.

Como ejemplo de sucesos posibles en este espacio muestral podemos destacar, entre otros,  $\{\text{menor que } 0.5\}$ ,  $\{\text{mayor que } 0.25\}$ ,  $\{\text{menor que } 0.75\}$ , ...

Otro espacio muestral podría ser observar el valor decimal mayor más cercano. Por ejemplo, si sale 0.25, me interesa 0.3. En ese caso el espacio muestral sería  $\Omega = 0.1, 0.2, \dots, 1$ . Este espacio muestral serviría, por ejemplo, para sortear números entre 1 y 10, sin más que multiplicar el resultado obtenido por 10.



En estos últimos ejemplos podemos ver que hay dos grandes tipos de espacios muestrales según el número de sucesos elementales.

Un espacio muestral se dice **discreto** si está formado por un conjunto finito o infinito numerable de sucesos elementales.

Por el contrario, un espacio muestral se dice **continuo** si está formado por un conjunto no numerable de sucesos elementales.

### 3.3.3. Función de probabilidad

Dado un espacio muestral  $\Omega$  correspondiente a un experimento aleatorio, una **función de probabilidad** para ese espacio muestral es cualquier función que asigne a cada suceso un número en el intervalo  $[0, 1]$  y que verifique

$P[A] \geq 0$ , para cualquier evento  $A$ .

$P[\Omega] = 1$ .

Dada una colección de sucesos  $A_1, A_2, \dots, A_n$  mutuamente excluyentes, es decir, tales que  $A_i \cap A_j = \emptyset$  para todo  $i \neq j$ ,

$$P[\cup_{i=1}^n A_i] = \sum_{i=1}^n P[A_i].$$

**Nota.** Hay que notar que se puede dar más de una función de probabilidad asociada al mismo espacio muestral. Por ejemplo, asociado al espacio muestral  $\Omega = \{cara, cruz\}$ , del lanzamiento de una moneda, pueden darse un número infinito no numerable de medidas de la probabilidad; concretamente, asociadas a cada elección

$$\begin{aligned} P[cara] &= p \\ P[cruz] &= 1 - p, \end{aligned}$$

para cada  $p \in [0, 1]$ . Aunque si la moneda no está cargada, como sucede habitualmente, se considera el caso en que  $p = \frac{1}{2}$ .

**Ejemplo.** Volviendo sobre el lanzamiento del dado, si éste no está cargado, podemos definir la siguiente función de probabilidad:

$$P[\{i\}] = \frac{1}{6}, \quad i = 1, 2, \dots, 6.$$

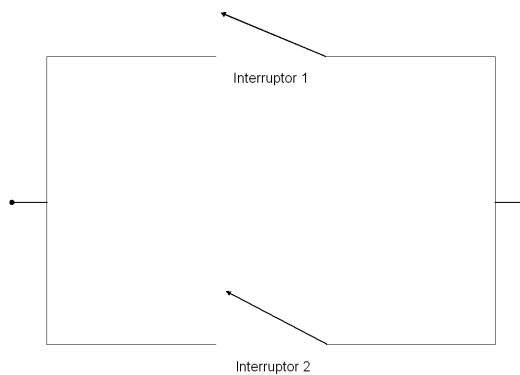


Figura 3.1: Circuito

En ese caso, podemos, a su vez, calcular algunas probabilidades. Por ejemplo,

$$\begin{aligned} P(\{par\}) &= P[\{2, 4, 6\}] \\ &= P[\{2\}] + P[\{4\}] + P[\{6\}] \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5. \end{aligned}$$

En este cálculo se ha tenido en cuenta la tercera condición de la definición axiomática.

Como consecuencia de la definición se verifican, entre otras, las siguientes propiedades, que además facilitan bastante los cálculos:

- $P[\emptyset] = 0$ .
- Sea  $A$  un suceso cualquiera. Entonces,  $P[\bar{A}] = 1 - P[A]$ .
- Sean  $A$  y  $B$  dos sucesos cualesquiera. Entonces,  $P[A \cap \bar{B}] = P[A] - P[A \cap B]$ .
- Sean  $A$  y  $B$  dos sucesos cualesquiera. Entonces,  $P[A \cup B] = P[A] + P[B] - P[A \cap B]$ .

**Ejemplo.** El circuito que aparece en la Figura 3.1 está constituido por dos interruptores (*switches*) en paralelo. La probabilidad de que cualquiera de ellos esté cerrado es de  $\frac{1}{2}$ .

Para que pase corriente a través del circuito basta con que pase corriente por alguno de los dos interruptores, esto es, que al menos uno de ellos esté cerrado. Por tanto, si notamos por  $E$  al suceso *que pase corriente a través del circuito* y  $E_i$  al suceso *que el interruptor  $i$  esté cerrado*, entonces,

$$\begin{aligned} P[E] &= P[E_1 \cup E_2] = P[E_1] + P[E_2] - P[E_1 \cap E_2] \\ &= \frac{1}{2} + \frac{1}{2} - P[E_1 \cap E_2] \leq 1. \end{aligned}$$

Para conocer esta probabilidad de forma exacta necesitamos saber cómo actúan de forma conjunta ambos circuitos.

Nº de lanzamientos	10	100	250	500	750	1000
Nº de caras	4	46	124	244	379	501
$\frac{N. \text{ de caras}}{N. \text{ de lanzamientos}}$	0.4	0.46	0.496	0.488	0.5053	0.501

Cuadro 3.1: Aproximación frecuentista a la probabilidad de cara en el lanzamiento de una moneda.

### 3.4. Interpretación frecuentista de la probabilidad

La interpretación más común al concepto de probabilidad tiene que ver con los promedios de ocurrencia de los sucesos del experimento en cuestión.

Pensemos en el lanzamiento de una moneda: si decimos que la probabilidad de cara es 0.5, entendemos que si lanzamos la moneda un gran número de veces y anotamos el número de caras, éstas serán más o menos la mitad.

Generalizando este proceso, podríamos decir que la probabilidad de un evento  $A$ ,  $P[A]$ , es

$$P[A] = \lim_{n \rightarrow \infty} \frac{n_A}{n},$$

donde  $n_A$  es el número de ocurrencias de  $A$  en  $n$  ensayos del experimento.

Esta interpretación se conoce como *definición frecuentista de la probabilidad*. Se trata de una interpretación de carácter eminentemente práctico porque permite una aproximación física al concepto de probabilidad, pero se ve limitada por las complicaciones que supone la definición en términos de un límite que, como tal, sólo se alcanza *en el infinito*. Además, desde un punto de vista realista, ¿en qué ocasiones podremos repetir el experimento un gran número de veces?

**Ejemplo.** Se han realizado 1000 lanzamientos de una moneda. En el Cuadro 3.1 aparece un resumen de ese proceso. Puede observarse como cuanto mayor es el número de lanzamientos, más se aproxima la frecuencia relativa al valor  $\frac{1}{2}$ , de manera que podríamos pensar que la probabilidad de cara es igual que la probabilidad de cruz e iguales ambas a  $\frac{1}{2}$ , aunque esto sólo es una suposición, o una aproximación, ya que para aplicar estrictamente la definición frecuentista deberíamos continuar hasta el infinito, lo que resulta imposible.

Esta interpretación frecuentista de la probabilidad permite inferir lo que podemos llamar *frecuencias esperadas*. Si un evento  $A$  tiene asignada una probabilidad  $P[A]$ , entonces, si repetimos el experimento aleatorio  $n$  veces, *lo más esperable* es que el número de veces que se de el evento  $A$  será  $n \times P[A]$ . Más adelante podremos matizar con más rigor a qué nos referimos con *lo más esperable*.

**Ejemplo.** Siguiendo con el ejemplo de la moneda, si la lanzamos 348 veces, lo esperable es que salgan alrededor de  $348 \times 0.5 = 174$  caras.

### 3.5. Interpretación subjetiva de la probabilidad

Si nos dicen que la probabilidad de que llueva mañana es del 35 %, ¿cómo podemos interpretar eso en términos frecuentistas? No tiene sentido pensar en que podemos repetir el experimento *día de mañana* muchas veces y contar cuántas veces llueve. ¿Podríamos pensar *si hubiera muchos días como el de mañana, aproximadamente llovería en el 35 % de ellos*? Pero eso no tiene sentido porque el día de mañana es único.

La interpretación subjetiva de la probabilidad tiene que ver con la vinculación de este concepto con el grado de incertidumbre que tenemos sobre las cosas. Si tenemos un experimento aleatorio, el resultado de dicho experimento es incierto. La probabilidad de un resultado del experimento es el grado de creencia que yo tengo en la ocurrencia de dicho resultado. Ese grado de creencia es personal, luego es subjetivo, pero lógicamente, deberá estar acorde con la información que tenemos sobre el experimento.

### 3.6. Espacio muestral con resultados equiprobables. Fórmula de Laplace

Otro punto de vista que permite abordar el proceso de asignación de probabilidad a sucesos es el siguiente: continuando con el ejemplo de la moneda, en este experimento son dos los resultados posibles, y no hay razones para pensar que uno de ellos es *más probable* que otro, así que tiene sentido considerar que la probabilidad de cara y la probabilidad de cruz son ambas del 50 %.

En general, si el espacio muestral está formado por  $N$  resultados posibles y todos ellos tienen la misma probabilidad (equiprobables), podríamos decir que la probabilidad de un evento  $A$ ,  $P[A]$ , es

$$P[A] = \frac{N_A}{N},$$

donde  $N_A$  es el número de resultados favorables a la ocurrencia de  $A$ .

Esta fórmula, conocida como *fórmula de Laplace* también es fundamentalmente práctica. Por ejemplo, nos permite deducir que

$$P[\text{cara}] = \frac{1}{2}$$

en el lanzamiento de una moneda sin tener que lanzar la moneda un gran número de veces.

Sin embargo, la definición tiene dos grandes inconvenientes: el conjunto de resultados posibles,  $N$ , tiene que ser finito y, además, todos los resultados posibles deben tener la misma probabilidad (con lo cual, lo definido queda implícitamente inmerso en la definición).

### 3.7. Probabilidad condicionada. Independencia de sucesos

Para introducir de manera intuitiva el concepto de probabilidad condicionada debemos pensar en la probabilidad como medida de la creencia en la ocurrencia de los sucesos.

Pensemos en un experimento aleatorio y en un suceso de dicho experimento,  $A$ , en el que, en principio, tenemos un grado de creencia  $P[A]$ ; pero supongamos que conocemos algo del resultado de dicho experimento; concretamente, sabemos que ha ocurrido un suceso  $B$ . Parece lógico pensar que esa información conocida sobre el resultado del ensayo modificará nuestro grado de creencia en  $A$ : llamemos a este nuevo grado de creencia  $P[A | B]$ , **probabilidad de  $A$  conocida  $B$**  o **probabilidad de  $A$  condicionada a  $B$** .

**Ejemplo.** Consideremos el suceso  $A$  : el día de hoy va a llover y el suceso  $B$  : el día de hoy está nublado. Obviamente, la probabilidad  $P[A]$  será menor que la probabilidad  $P[A | B]$ , ya que el hecho de que esté nublado refuerza nuestra creencia en que llueva.

**Ejemplo.** Consideremos el experimento aleatorio de extraer una carta de una baraja española. Sea el suceso  $A$  : obtener una sota, el suceso  $B_1$  : obtener una figura y el suceso  $B_2$  : obtener una carta de copas. Las distintas probabilidades, condicionadas o no, bajo la definición clásica, son las siguientes:

$$\begin{aligned} P[A] &= \frac{4 \text{ sotas}}{40 \text{ cartas}} = \frac{1}{10} \\ P[A | B_1] &= \frac{4 \text{ sotas}}{12 \text{ figuras}} = \frac{1}{3} \\ P[A | B_2] &= \frac{1 \text{ sota de copas}}{10 \text{ copas}} = \frac{1}{10}. \end{aligned}$$

Como puede verse,  $B_1$  modifica la probabilidad a priori, pero no así  $B_2$ . Puede decirse que  $B_2$  no ofrece información acerca de  $A$ , o que  $A$  y  $B_2$  son **independientes**.

Vamos a dar a continuación una definición de **probabilidad condicionada** que responde a esta idea de recalculer la probabilidad en función de la información existente.

La **probabilidad condicionada de un suceso  $A$ , conocido otro suceso  $B$** , denotada por  $P[A | B]$ , se define como el cociente

$$P[A | B] = \frac{P[A \cap B]}{P[B]},$$

siempre que  $P[B] \neq 0$ .

Una función de probabilidad condicionada  $P[\cdot/B]$  es una función de probabilidad en toda regla: por tanto, cumple las mismas propiedades que cualquier función de probabilidad “sin condicionar”.

Como hemos comentado, la idea de la probabilidad condicionada es utilizar la información que nos da un suceso conocido sobre la ocurrencia de otro suceso. Pero, como ya hemos puesto de manifiesto en un ejemplo, no siempre un suceso da información sobre otro. En este caso se dice que ambos sucesos son **independientes**. Por tanto:

Dos sucesos  $A$  y  $B$  se dicen independientes si  $P[A | B] = P[A]$ , o equivalentemente si  $P[B | A] = P[B]$ , o equivalentemente si  $P[A \cap B] = P[A] \times P[B]$ .

**Ejemplo.** Continuando con el Ejemplo 3.3.3, lo más lógico es pensar que los dos interruptores actúan de forma independiente, en cuyo caso  $P[E_1 \cap E_2] = P[E_1] P[E_2]$  y tenemos que,

$$\begin{aligned} P[E] &= \frac{1}{2} + \frac{1}{2} - P[E_1 \cap E_2] \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2} \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

**Nota.** Es muy importante no confundir la probabilidad condicionada de un suceso a otro con la probabilidad de la intersección de ambos sucesos. En la Figura 3.2 puede verse la diferencia entre las probabilidades condicionadas entre dos sucesos y la probabilidad de su intersección. En términos coloquiales, podemos

analizar estas probabilidades como el cociente entre *una parte* y *un todo*. Cuando la probabilidad es condicionada ese *todo* es el suceso que condiciona. Cuando la probabilidad no es condicionada, ese *todo* es todo el espacio muestral. En ambos casos esa *parte* es la intersección.

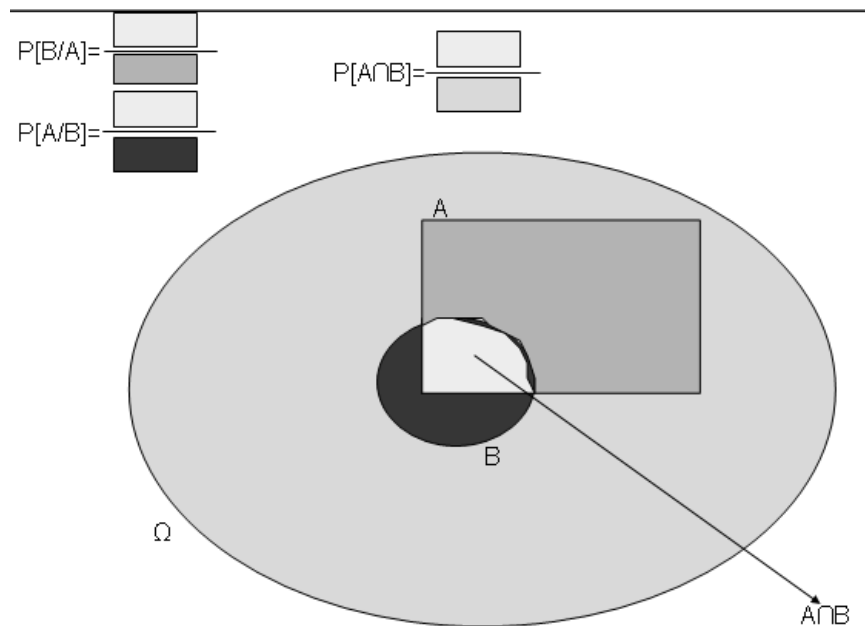


Figura 3.2: Esquema acerca de la definición de probabilidad condicionada.

**Nota.** También suele ser bastante común la confusión entre sucesos independientes y sucesos incompatibles o mutuamente excluyentes.

En este sentido, recordemos que dos sucesos  $A$  y  $B$  son incompatibles o mutuamente excluyentes si  $A \cap B = \emptyset$ , en cuyo caso  $P[A \cap B] = 0$ .

Por su parte,  $A$  y  $B$  serán independientes si  $P[A \cap B] = P[A] P[B]$ .

Las diferencias entre ambos conceptos son obvias.

**Ejemplo.** La probabilidad de que el producto no sea elaborado a tiempo es 0.05. Se solicitan tres pedidos del producto con la suficiente separación en el tiempo como para considerarlos eventos independientes.

1. ¿Cuál es la probabilidad de que todos los pedidos se envíen a tiempo?

En primer lugar, notemos  $E_i$  al suceso *enviar a tiempo el pedido  $i$ -ésimo*. En ese caso, sabemos que  $P[E_i] = 0.95$ .

Por su parte, nos piden

$$P[E_1 \cap E_2 \cap E_3] = P[E_1] P[E_2] P[E_3] = 0.95^3,$$

debido a que los pedidos son independientes.

2. ¿Cuál es la probabilidad de que exactamente un pedido no se envíe a tiempo?

En este caso el suceso que nos piden es más complejo:

$$\begin{aligned} & P [\bar{E}_1 \cap E_2 \cap E_3 \cup E_1 \cap \bar{E}_2 \cap E_3 \cup E_1 \cap E_2 \cap \bar{E}_3] \\ &= P [\bar{E}_1 \cap E_2 \cap E_3] + P [E_1 \cap \bar{E}_2 \cap E_3] + P [E_1 \cap E_2 \cap \bar{E}_3] \\ &= 0.05 \times 0.95^2 + 0.05 \times 0.95^2 + 0.05 \times 0.95^2 = 0.135, \end{aligned}$$

donde se ha utilizado que los sucesos  $\bar{E}_1 \cap E_2 \cap E_3$ ,  $E_1 \cap \bar{E}_2 \cap E_3$  y  $E_1 \cap E_2 \cap \bar{E}_3$  son incompatibles.

3. ¿Cuál es la probabilidad de que dos o más pedidos no se envíen a tiempo?

Tengamos en cuenta que ya hemos calculado la probabilidad de que todos se envíen a tiempo y de que todos menos uno se envíen a tiempo. Entonces,

$$\begin{aligned} & P [\text{dos o más pedidos no se envíen a tiempo}] \\ &= 1 - P [\text{todos se envíen a tiempo} \cup \text{un pedido no se envíe a tiempo}] \\ &= 1 - (0.95^3 + 0.135). \end{aligned}$$

**Ejemplo.** Consideremos un proceso industrial como el que se esquematiza en la Figura 3.3. En dicho esquema se pone de manifiesto que una unidad será producida con éxito si pasa en primer lugar un chequeo previo (A); después puede ser montada directamente (B), redimensionada (C) y después montada (D) o adaptada (E) y después montada (F); posteriormente debe ser pintada (G) y finalmente embalada (H). Consideremos que las probabilidades de pasar exitosamente cada subproceso son todas ellas iguales a 0.95, y que los subprocesos tienen lugar de forma independiente unos de otros. Vamos a calcular en esas condiciones la probabilidad de que una unidad sea exitosamente producida.

Si nos damos cuenta, A, G y H son ineludibles, mientras que una unidad puede ser producida si pasa por B, por C y D o por E y F. En notación de conjuntos, la unidad será producida si se da

$$A \cap (B \cup C \cap D \cup E \cap F) \cap G \cap H.$$

Como los procesos son independientes unos de otros, no tenemos problemas con las probabilidades de las intersecciones, pero tenemos que calcular la probabilidad de una unión de tres conjuntos,  $B \cup C \cap D \cup E \cap F$ . En general,

$$\begin{aligned} P[A_1 \cup A_2 \cup A_3] &= P[(A_1 \cup A_2) \cup A_3] = P[A_1 \cup A_2] + P[A_3] - P[(A_1 \cup A_2) \cap A_3] \\ &= P[A_1] + P[A_2] - P[A_1 \cap A_2] + P[A_3] - P[A_1 \cap A_3 \cup A_2 \cap A_3] \end{aligned}$$

$$= P[A_1] + P[A_2] - P[A_1 \cap A_2] + P[A_3] \\ - (P[A_1 \cap A_3] + P[A_2 \cap A_3] - P[A_1 \cap A_2 \cap A_3])$$

$$= P[A_1] + P[A_2] + P[A_3] \\ - P[A_1 \cap A_2] - P[A_1 \cap A_3] - P[A_2 \cap A_3] \\ + P[A_1 \cap A_2 \cap A_3]$$

En nuestro caso,

$$P[B \cup C \cap D \cup E \cap F] = P[B] + P[C \cap D] + P[E \cap F] \\ - P[B \cap C \cap D] - P[B \cap E \cap F] - P[C \cap D \cap E \cap F] \\ + P[B \cap C \cap D \cap E \cap F] \\ = 0.95 + 2 \times 0.95^2 - 2 \times 0.95^3 - 0.95^4 + 0.95^5 \\ = 0.9995247$$

Ya estamos en condiciones de obtener la probabilidad que se nos pide:

$$P[A \cap (B \cup C \cap D \cup E \cap F) \cap G \cap H] = P[A] P[B \cup C \cap D \cup E \cap F] P[G] P[H] \\ = 0.95 \times (0.9995247) \times 0.95 \times 0.95 \\ = 0.8569675.$$

En estos ejemplos, el cálculo de la probabilidad de las intersecciones ha resultado trivial porque los sucesos son independientes. Son embargo, esto no siempre ocurre. ¿Cómo podemos, en general, obtener la probabilidad de la intersección de dos o más sucesos no necesariamente independientes?

En el caso de sólo dos sucesos,  $A$  y  $B$ , podemos deducir que

$$P[A \cap B] = P[A|B] \times P[B]$$

directamente de la definición de probabilidad condicionada. A partir de esta fórmula, por inducción, se puede obtener la llamada fórmula producto, que se enuncia de la siguiente forma: si  $A_1, A_2, \dots, A_n$  son sucesos de un espacio muestral no necesariamente independientes, se verifica

$$P[A_1 \cap A_2 \cap \dots \cap A_n] = P[A_1]P[A_2|A_1] \dots P[A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}]$$



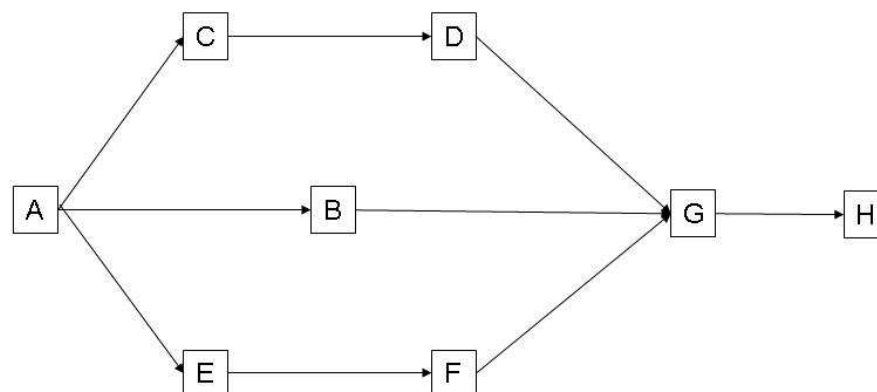


Figura 3.3: Esquema del proceso industrial del ejemplo

**Ejemplo.** Un lote de 50 arandelas contiene 30 arandelas cuyo grosor excede las especificaciones de diseño. Suponga que se seleccionan 3 arandelas al azar y sin reemplazo del lote.

1. ¿Cuál es la probabilidad de que las tres arandelas seleccionadas sean más gruesas que las especificaciones de diseño?

Comenzamos notando los sucesos  $A_i$ : la  $i$ -ésima arandela extraída es más gruesa que las especificaciones de diseño,  $i = 1, 2, 3$ .

Entonces, nos piden

$$\begin{aligned}
 P[A_1 \cap A_2 \cap A_3] &= P[A_1] P[A_2/A_1] P[A_3/A_1 \cap A_2] \\
 &= \frac{30}{50} \frac{29}{49} \frac{28}{48}.
 \end{aligned}$$

2. ¿Cuál es la probabilidad de que la tercera arandela seleccionada sea más gruesa que las especificaciones de diseño si las dos primeras fueron más delgadas que la especificación?

$$P[A_3/\bar{A}_1 \cap \bar{A}_2] = \frac{30}{48}.$$

### 3.8. Teorema de la probabilidad total y Teorema de Bayes

Los siguientes dos resultados se conocen como **Teorema de la probabilidad total** y **Teorema de Bayes** respectivamente, y juegan un importante papel a la hora de calcular probabilidades. Los dos utilizan como

principal herramienta el concepto de probabilidad condicionada.

**Teorema de la Probabilidad Total.** Sea  $P$  una función de probabilidad en un espacio muestral. Sea  $\{A_1, \dots, A_N\} \subset F$  una partición del espacio muestral  $\Omega$  y sea  $B$  un suceso cualquiera. Entonces,

$$P[B] = P[B | A_1] P[A_1] + \dots + P[B | A_N] P[A_N].$$

**Teorema de Bayes.** En esas mismas condiciones, si  $P[B] \neq 0$ ,

$$P[A_i | B] = \frac{P[B | A_i] P[A_i]}{P[B | A_1] P[A_1] + \dots + P[B | A_N] P[A_N]}.$$

**Ejemplo.** Supongamos que tenemos 4 cajas con componentes electrónicas dentro. La caja 1 contiene 2000 componentes, con un 5% de defectuosas; la caja 2 contiene 500 componentes, con un 40% de defectuosas; las cajas 3 y 4 contienen 1000 componentes, con un 10% de defectuosas.

1. ¿Cuál es la probabilidad de escoger al azar una componente defectuosa?

Notemos  $D$  : componente defectuosa y  $C_i$  : componente de la caja  $i$ -ésima. Entonces, se tiene que

$$\begin{aligned} P[C_1] &= \frac{2000}{2000 + 500 + 1000 + 1000} = \frac{4}{9} \\ P[C_2] &= \frac{500}{2000 + 500 + 1000 + 1000} = \frac{1}{9} \\ P[C_3] &= \frac{1000}{2000 + 500 + 1000 + 1000} = \frac{2}{9} \\ P[C_4] &= \frac{1000}{2000 + 500 + 1000 + 1000} = \frac{2}{9} \end{aligned}$$

Además,  $P[D | C_1] = 0.05$ ,  $P[D | C_2] = 0.4$ ,  $P[D | C_3] = 0.1$  y  $P[D | C_4] = 0.1$ .

Utilizando el Teorema de la probabilidad total,

$$\begin{aligned} P[D] &= P[D | C_1] P[C_1] + P[D | C_2] P[C_2] + P[D | C_3] P[C_3] \\ &\quad + P[D | C_4] P[C_4] \\ &= 0.05 \frac{4}{9} + 0.4 \frac{1}{9} + 0.1 \frac{2}{9} + 0.1 \frac{2}{9} = 0.11111 \end{aligned}$$

2. Si se escoge una componente al azar y resulta ser defectuosa, ¿cuál es la probabilidad de que pertenezca a la caja 1?

$$P[C_1 | D] = \frac{P[D | C_1] P[C_1]}{P[D]} = \frac{0.05 \frac{4}{9}}{0.11111} = 0.2$$

	Número en cada caja			
$\mu F$	1	2	3	Total
0.01	20	95	25	140
0.1	55	35	75	165
1.0	70	80	145	295
Total	145	210	245	600

Cuadro 3.2: Acumuladores.

**Ejemplo.** Se disponen tres cajas donde se almacenan acumuladores según aparece en el Cuadro 3.2.

Se escoge al azar una caja y de ella, a su vez, un acumulador.

1. ¿Cuál es la probabilidad de que se haya seleccionado un acumulador de  $0.01\mu F$ ?

Notemos  $0.01\mu F$ ,  $0.1\mu F$  y  $1.0\mu F$  a los sucesos *extraer un acumulador de*  $0.01\mu F$ ,  $0.1\mu F$  y  $1.0\mu F$  respectivamente. De igual forma, notemos  $c1$ ,  $c2$  y  $c3$  a los sucesos *elegir la caja 1, la caja 2 y la caja 3*, respectivamente. Utilizando el teorema de la probabilidad total,

$$\begin{aligned} P[0.01\mu F] &= P[0.01\mu F / c1] P[c1] + P[0.01\mu F / c2] P[c2] + P[0.01\mu F / c3] P[c3] \\ &= \frac{20}{145} \frac{1}{3} + \frac{95}{210} \frac{1}{3} + \frac{25}{245} \frac{1}{3} = \frac{5903}{25578} = 0.23078. \end{aligned}$$

2. Si ha sido seleccionado un acumulador de  $1.0\mu F$ , ¿cuál es la probabilidad de que proceda de la caja 1? Utilizando el teorema de Bayes,

$$P[c1 / 1.0\mu F] = \frac{P[1.0\mu F / c1] P[c1]}{P[1.0\mu F]}.$$

Por su parte,

$$\begin{aligned} P[1.0\mu F] &= P[1.0\mu F / c1] P[c1] + P[1.0\mu F / c2] P[c2] + P[1.0\mu F / c3] P[c3] \\ &= \frac{70}{145} \frac{1}{3} + \frac{80}{210} \frac{1}{3} + \frac{145}{245} \frac{1}{3} = \frac{6205}{12789} = 0.48518, \end{aligned}$$

luego

$$P[c1 / 1.0\mu F] = \frac{\frac{70}{145} \frac{1}{3}}{\frac{6205}{12789}} = \frac{2058}{6205} = 0.33167.$$

**Ejemplo.** Siguiendo con el ejemplo de las arandelas con grosor fuera de las especificaciones de diseño, ¿cuál es la probabilidad de que la tercera arandela seleccionada sea más gruesa que las especificaciones de diseño?

$$\begin{aligned} P[A_3] &= P[A_3 | A_1 \cap A_2] P[A_1 \cap A_2] + P[A_3 | \bar{A}_1 \cap A_2] P[\bar{A}_1 \cap A_2] \\ &\quad + P[A_3 | A_1 \cap \bar{A}_2] P[A_1 \cap \bar{A}_2] + P[A_3 | \bar{A}_1 \cap \bar{A}_2] P[\bar{A}_1 \cap \bar{A}_2] \end{aligned}$$

$$\begin{aligned}
 &= P[A_3|_{A_1 \cap A_2}]P[A_1]P[A_2|_{A_1}] + P[A_3|_{\bar{A}_1 \cap A_2}]P[\bar{A}_1]P[A_2|_{\bar{A}_1}] \\
 &+ P[A_3|_{A_1 \cap \bar{A}_2}]P[A_1]P[\bar{A}_2|_{A_1}] + P[A_3|_{\bar{A}_1 \cap \bar{A}_2}]P[\bar{A}_1]P[\bar{A}_2|_{\bar{A}_1}] \\
 &= \frac{28}{48} \frac{30}{50} \frac{29}{49} + \frac{29}{48} \frac{20}{50} \frac{30}{49} \\
 &+ \frac{29}{48} \frac{30}{50} \frac{20}{49} + \frac{30}{48} \frac{20}{50} \frac{19}{49}.
 \end{aligned}$$

**Ejemplo.** En el canal de comunicaciones ternario que se describe en la Figura 3.4, se ha observado que el dígito 3 es enviado tres veces más frecuentemente que 1, y 2 dos veces más frecuentemente que 1. Calculemos la probabilidad de que un dígito cualquiera enviado a través del canal sea recibido correctamente.

En primer lugar, si notamos  $P[X = 1] = p$ , entonces  $P[X = 2] = 2p$  y  $P[X = 3] = 3p$ . Por otra parte, como

$$1 = P[X = 1] + P[X = 2] + P[X = 3] = 6p,$$

se tiene que

$$P[X = 1] = \frac{1}{6}, P[X = 2] = \frac{1}{3} \text{ y } P[X = 3] = \frac{1}{2}.$$

Ahora, utilizando el teorema de la probabilidad total,

$$\begin{aligned}
 P[\text{dígito OK}] &= P[\text{dígito OK} / X = 1] P[X = 1] \\
 &+ P[\text{dígito OK} / X = 2] P[X = 2] \\
 &+ P[\text{dígito OK} / X = 3] P[X = 3] \\
 &= P[Y = 1 / X = 1] P[X = 1] \\
 &+ P[Y = 2 / X = 2] P[X = 2] \\
 &+ P[Y = 3 / X = 3] P[X = 3] \\
 &= (1 - \alpha) \frac{1}{6} + (1 - \beta) \frac{1}{3} + (1 - \gamma) \frac{1}{2} = P.
 \end{aligned}$$

**Ejemplo.** Continuando con el anterior, si se recibe un 1, ¿cuál es la probabilidad de que se hubiera enviado un 1?

Utilizando el teorema de Bayes,

$$P[X = 1 / Y = 1] = \frac{P[Y = 1 / X = 1] P[X = 1]}{P[Y = 1]}.$$

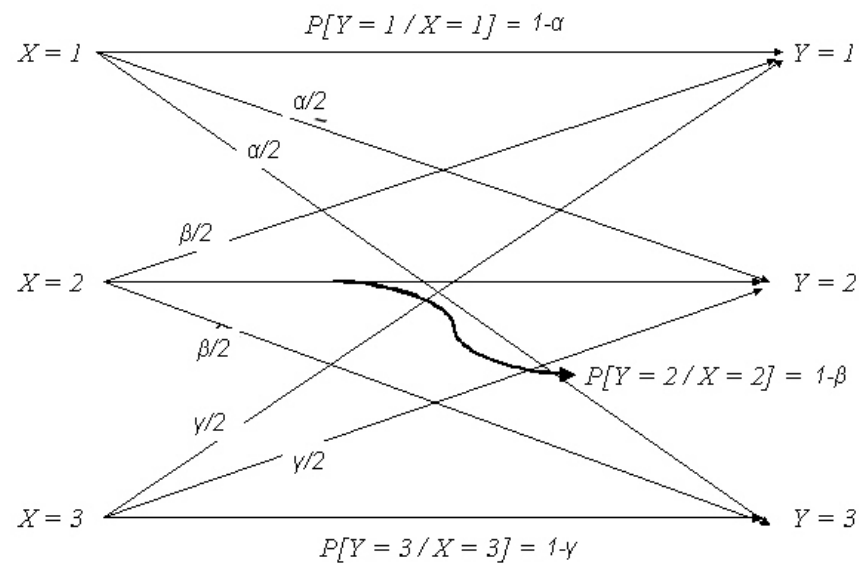


Figura 3.4: Canal ternario de comunicaciones con probabilidad de cruce

Por su parte,

$$\begin{aligned} P[Y = 1] &= P[Y = 1 / X = 1] P[X = 1] \\ &\quad + P[Y = 1 / X = 2] P[X = 2] \\ &\quad + P[Y = 1 / X = 3] P[X = 3] \\ &= \frac{1-\alpha}{6} + \frac{\beta}{6} + \frac{\gamma}{4}, \end{aligned}$$

luego

$$P[X = 1 / Y = 1] = \frac{\frac{1-\alpha}{6}}{\frac{1-\alpha}{6} + \frac{\beta}{6} + \frac{\gamma}{4}} = 2 \frac{-1 + \alpha}{-2 + 2\alpha - 2\beta - 3\gamma}.$$

### 3.9. Más sobre el Teorema de Bayes

La importancia del Teorema de Bayes en Estadística va mucho más allá de su aplicación como fórmula que facilita probabilidades condicionadas. La filosofía que subyace en él ha dado lugar a toda una forma de entender la Estadística, llamada por ello *Estadística Bayesiana*. Vamos a tratar de explicar los fundamentos de esta manera de entender el teorema.

Supongamos que hay un suceso  $A$  sobre el que tenemos un serio desconocimiento acerca de si se da o no se da. Tanto es así que tenemos que determinar la probabilidad de dicho suceso,  $P[A]$ . Es importante entender que nosotros somos conscientes de que  $A$  ha ocurrido o no ha ocurrido: el problema es precisamente que no sabemos qué ha pasado. Decimos que es importante porque  $P[A]$  no representa la *probabilidad de que A ocurra*, sino nuestro grado de creencia en que ha ocurrido.

Es posible que no tengamos, en principio, datos para conocer de forma exacta cuál es la probabilidad de  $A$ . Aún así, podríamos atrevernos, *como expertos en el tema*, a dar una estimación de dicha probabilidad,  $P[A]$ . A esta probabilidad inicial que damos la vamos a llamar **probabilidad a priori**.

Ahora bien, hemos dado una probabilidad a priori  $P[A]$  sin ninguna información sobre  $A$ . Supongamos ahora

que tenemos nueva información que nos dará pistas acerca de si  $A$  ha ocurrido o no, y que dicha información está recogida en un suceso que llamaremos  $B_1$ . En ese caso, podríamos y deberíamos *actualizar la probabilidad* de  $A$  basándonos en esta nueva información, proporcionando una nueva probabilidad de  $A$  que tenga en cuenta  $B_1$ , es decir,  $P[A | B_1]$ , que llamaremos **probabilidad a posteriori**.

En esa *actualización de la probabilidad* es donde entra el Teorema de Bayes, ya que nos dice que

$$P[A | B_1] = \frac{P[B_1 | A] P[A]}{P[B_1 | A] P[A] + P[B_1 | \bar{A}] P[\bar{A}]}.$$

Obsérvese que la probabilidad a posteriori es proporcional a la probabilidad a priori.

Finalmente, es muy importante ver que podemos extender esta forma de trabajar aplicando el teorema de una forma recursiva. Después de conocer  $B_1$ , nuestra nueva probabilidad para  $A$  es  $P[A | B_1]$ . Abusando de la notación, podemos decir que esa es nuestra nueva probabilidad a priori y si, por ejemplo, tenemos más información sobre  $A$ , dada por otro suceso  $B_2$ , **información independiente de  $B_1$** , la nueva probabilidad a posteriori sería

$$\begin{aligned} P[A | B_1 \cap B_2] &= \frac{P[B_2 | A \cap B_1] P[A | B_1]}{P[B_2 | A \cap B_1] P[A | B_1] + P[B_2 | \bar{A} \cap B_1] P[\bar{A} | B_1]} \\ &= \frac{P[B_2 | A] P[A | B_1]}{P[B_2 | A] P[A | B_1] + P[B_2 | \bar{A}] P[\bar{A} | B_1]}. \end{aligned}$$

Es muy importante observar que en este cociente  $P[A | B_1]$  ocupa el lugar que antes ocupaba la probabilidad a priori. Además, esta segunda probabilidad a posteriori podría considerarse como la nueva probabilidad a priori para una nueva aplicación del teorema basada en el conocimiento de nueva información dada por un suceso  $B_3$ . Este proceso de actualización de las probabilidades a priori basada en la información disponible puede realizarse cuantas veces sea necesario.

Vamos a ilustrar esto en un par de ejemplos.

### 3.9.1. Ejemplo del juez

Supongamos que un juez debe decidir si un sospechoso es inocente o culpable. Él sabe que debe ser cuidadoso y garantista con los derechos del acusado, pero también por su experiencia parte de una creencia en que el sospechoso puede ser culpable que, en cualquier caso, estima por debajo de lo que realmente cree para, insisto, ser garantista con los derechos del acusado. Pongamos que estima esta probabilidad en un 10%.

Ahora empieza a examinar las pruebas. La primera de ellas es una prueba de ADN en la que el acusado dio positivo: encontraron material genético en el arma del crimen que, según la prueba, es suyo. Esa prueba de ADN da positivo en el 99.5 % de las veces en que se comparan dos ADN's idénticos, pero también da positivo (erróneamente) en el 0.005 % de las veces en que se aplica a dos ADN's distintos. Teniendo en cuenta esta información, el juez aplica por primera vez el teorema de Bayes con los siguientes datos:

- $P[\text{culpable}] = 0.1$ , que es la probabilidad a priori que el juez considera.
- La probabilidad de que la prueba de ADN de positivo si el acusado es culpable es

$$P[\text{ADN}+ | \text{culpable}] = 0.995.$$

- La probabilidad de que la prueba de ADN de positivo si el acusado es inocente es

$$P[ADN+ | inocente] = 0.00005.$$

Ahora ya puede actualizar su grado de creencia en la culpabilidad del sospechoso:

$$\begin{aligned} P[culpable | ADN+] &= \frac{P[ADN+ | culpable] \times P[culpable]}{P[ADN+ | culpable] \times P[culpable] + P[ADN+ | inocente] \times P[inocente]} \\ &= \frac{0.995 \times 0.1}{0.995 \times 0.1 + 0.00005 \times 0.9} = 0.999548 \end{aligned}$$

Es decir, ahora piensa que el sospechoso es culpable con un 99.9548 % de certeza. Fijémonos en que nuestra probabilidad a priori aparece en los términos 0.1 en el numerador y 0.1 y 0.9 en el denominador. Esa, 0.1, era la probabilidad que teníamos **antes de la prueba** de que fuera culpable (y 0.9 de que fuera inocente); **después de la prueba** esa probabilidad es 0.999548 de que sea culpable (y 0.000452 de que sea inocente).

Sin embargo, el sospechoso insiste en su inocencia, y propone someterse a una prueba de un detector de mentiras. Los expertos saben que un culpable es capaz de engañar a esta máquina en el 10 % de las veces, y que la máquina dirá el 1 % de las veces que un inocente miente. Nuestro sospechoso se somete a la máquina y ésta dice que es inocente. ¿Cuál será ahora la probabilidad que el juez asigna a la culpabilidad del sospechoso? Teniendo en cuenta que:

- $P[maquina- | culpable] = 0.1,$
- $P[maquina+ | inocente] = 0.01,$

debe aplicar de nuevo el Teorema de Bayes, considerando ahora que la probabilidad a priori de que sea culpable es 99.9548 %:

$$\begin{aligned} P[culpable | maquina-] &= \frac{P[maquina- | culpable] \times P[culpable]}{P[maquina- | culpable] \times P[culpable] + P[maquina- | inocente] \times P[inocente]} \\ &= \frac{0.1 \times 0.999548}{0.1 \times 0.999548 + (1 - 0.01) \times (1 - 0.999548)} = 0.9955431. \end{aligned}$$

Es decir, aún con esa prueba negativa, el juez aún tiene un 99.55431 % de certidumbre de que el sospechoso es culpable. De nuevo, podemos resumir este paso diciendo que **antes de la segunda prueba** nuestra probabilidad de que fuera culpable era de 0.999548 (que aparece en la fórmula ocupando la posición de la probabilidad a priori), mientras que **después de la segunda prueba** esa probabilidad es 0.9955431.

El proceso puede verse resumido en el Cuadro 3.3.

### 3.9.2. Ejemplo de la máquina de detección de fallos

En un proceso industrial de producción en serie de capós de coche, existe una máquina encargada de detectar desperfectos que desechen una pieza de capó. Esa máquina está calibrada para detectar una pieza defectuosa con un 90 % de acierto, pero también detecta como defectuosas el 5 % de las piezas no defectuosas. El encargado de calidad estima, por estudios previos, que el porcentaje general de piezas defectuosas es del 5 %. Este encargado, consciente de que la máquina puede dar por buenas piezas que son defectuosas, decide actuar de la siguiente forma: una pieza que sea detectada como no defectuosa pasará otras dos veces por la misma máquina detectora y sólo será declarada no defectuosa cuando en ninguna de esas tres pruebas, de defectuosa.

	$P[Culpable]$	
	Antes de la prueba	Después de la prueba
<b>1ª prueba:</b> $ADN+$	0.1	$\frac{P[ADN+ culpable] \times 0.1}{P[ADN+ culpable] \times 0.1 + P[ADN+ inocente] \times (1-0.1)} = 0.999548$
<b>2ª prueba:</b> $maquina-$	0.999548	$\frac{P[maquina- culpable] \times 0.999548}{P[maquina- culpable] \times 0.999548 + P[maquina- inocente] \times (1-0.999548)} = 0.9955431$

Cuadro 3.3: Esquema del proceso iterativo del teorema de Bayes en el ejemplo del juez. La probabilidad *a priori* (antes de cada prueba) es la que se utiliza en la fórmula para obtener la probabilidad *a posteriori* (después de cada prueba). La probabilidad *a posteriori* (después) de una prueba es la probabilidad *a priori* (antes) de la siguiente prueba.

Supongamos que una pieza pasa las tres veces y da no defectuosa: ¿cuál es la probabilidad de que realmente sea no defectuosa?

Vamos a empezar notando adecuadamente los sucesos. Notaremos  $D$  al suceso ser defectuosa y por  $+$  a dar positivo como defectuosa en la prueba de la máquina. Sabemos que:

- $P[D] = 0.05$ , que es la probabilidad a priori;
- $P[+|_D] = 0.9$  y
- $P[+|_{\bar{D}}] = 0.05$ .

La probabilidad a priori de que una pieza sea no defectuosa es de 0.95, pero si es detectada como defectuosa una primera vez, dicha probabilidad pasa a ser

$$\begin{aligned} P[\bar{D}|_+] &= \frac{P[+|_{\bar{D}}] P[\bar{D}]}{P[+|_{\bar{D}}] P[\bar{D}] + P[+|_D] P[D]} \\ &= \frac{0.95 \times 0.95}{0.95 \times 0.95 + 0.1 \times 0.05} = 0.9944904. \end{aligned}$$

Esa probabilidad pasa a ser la probabilidad a priori para la segunda vez que da no defectuosa. Por tanto, la probabilidad de que sea no defectuosa si da negativo por segunda vez es

$$\begin{aligned} P[\bar{D}|_{++}] &= \frac{P[+|_{\bar{D}}] 0.9944904}{P[+|_{\bar{D}}] 0.9944904 + P[+|_D] (1 - 0.9944904)} \\ &= \frac{0.95 \times 0.9944904}{0.95 \times 0.9944904 + 0.1 \times (1 - 0.9944904)} = 0.9994172. \end{aligned}$$

Finalmente, la probabilidad de que sea no defectuosa si da negativo por tercera vez es

$$\begin{aligned} P[\bar{D}|_{+++}] &= \frac{P[+|_{\bar{D}}] 0.9994172}{P[+|_{\bar{D}}] 0.9994172 + P[+|_D] (1 - 0.9994172)} \\ &= \frac{0.95 \times 0.9994172}{0.95 \times 0.9994172 + 0.1 \times (1 - 0.9994172)} = 0.9999386. \end{aligned}$$

Como podemos ver, si una pieza da no defectuosa tres veces, la probabilidad de que sea realmente no defectuosa es altísima, del orden del 99.99%, así que el método ideado por el responsable de calidad parece consistente.



	$P \mid \bar{D}$	
	Antes de la prueba	Después de la prueba
1ª prueba: $\bar{+}$	0.95	$\frac{P[+ \mid \bar{D}]0.95}{P[+ \mid \bar{D}]0.95 + P[+ \mid D](1-0.95)} = 0.9944904$
2ª prueba: $\bar{+}$	0.9944904	$\frac{P[+ \mid \bar{D}]0.9944904}{P[+ \mid \bar{D}]0.9944904 + P[+ \mid D](1-0.9944904)} = 0.9994172$
3ª prueba: $\bar{+}$	0.9994172	$\frac{P[+ \mid \bar{D}]0.9994172}{P[+ \mid \bar{D}]0.9994172 + P[+ \mid D](1-0.9994172)} = 0.9999386$

Cuadro 3.4: Esquema del proceso iterativo del teorema de Bayes en el ejemplo de la máquina de detección de fallos. La probabilidad *a priori* (antes de cada prueba) es la que se utiliza en la fórmula para obtener la probabilidad *a posteriori* (después de cada prueba). La probabilidad *a posteriori* (después) de una prueba es la probabilidad *a priori* (antes) de la siguiente prueba.

