# Cómo se distribuyen los datos

Todos los valores de las variables cuantitativas, obtenidos de una muestra determinada, representan una gran cantidad de datos. Es cierto que estos datos no brindan gran información, pero una vez agrupados en determinados índices de tendencia central, fundamentalmente la media, y analizando en qué medida se dispersan de este valor, principalmente con la desviación estándar, ofrecen ya una visión clara de lo que ocurre en la muestra de la que se han obtenido estos valores.

A partir de los índices descritos y en función de las frecuencias con que los valores de las variables se repiten, se puede estimar un modelo que describa el comportamiento de estas variables, lo que permite disponer de mucha más información acerca de los resultados que se han obtenido en el estudio.

## Cómo se pueden distribuir los datos

El conjunto de valores de una variable se puede caracterizar por unos parámetros sencillos. No obstante, no todos los conjuntos lo hacen de la misma forma. Hay, entonces, varias maneras de distribución de datos; a continuación se indica cómo se analizan las más comunes.

### Distribución de Bernouilli

Imaginemos que se efectúa un programa de cribado para detección de individuos con presión arterial elevada, se realiza el cribado en 675 personas y se identifica a 108 personas que presentan un valor superior a 140/90 mmHg. El resultado de la variable identificación de pacientes con presión arterial elevada sólo puede tomar dos valores: PA ± 140/90 mmHg y PA < 140/90 mmHg; el primer valor es el que se tomará como medida de la efectividad de la intervención. De este modo, la variable analizada que presenta resultados dicotómicos (mayor o menor de una cifra) se puede identificar mediante su probabilidad. En este caso, la probabilidad de identificar a estos individuos es p1: 108/675 = 0,16 o 16%, mientras que la de no identificarles es p2: (675-108)/675 = 0,84 u 84%.

### Distribución de Poisson

Cuando se analizan sucesos raros, donde el número de individuos de la muestra es grande, > 30, y la probabilidad del suceso analizado es pequeña, < 0,1, la distribución de valores adopta una distribución de Poisson, que se caracteriza por un único parámetro, su media y su varianza, simultáneamente. Esto ocurre, por ejemplo, al analizar la incidencia de espina bífida en partos, que en nuestro país supone sólo 2,87 casos por cada 10.000 nacimientos vivos (n:10.000 p: 2,87/10.000 = 0,000287). Como en España hubo en 2004 458.752 nacimientos, esta distribución se puede caracterizar por medio de su media, que es $\mu$: 458.752 * (2,87/10.000) = 132, que a la vez será su varianza.

### Otras distribuciones

Hay otros tipos de distribuciones cuya utilización es menos habitual y más compleja, pero no por ello menos importante. Un tipo de distribución que se verá con cierta frecuencia cuando se implementa una intervención para producir o prevenir un determinado acontecimiento es la exponencial negativa, que se estudia mediante análisis de supervivencia. En él se estudian los datos de la variable tiempo hasta que ocurre un episodio y cómo depende este tiempo de otras variables. Así, por ejemplo, el tiempo ocurrido hasta la observación de una complicación cardiovascular mayor a consecuencia de una presión elevada sigue una distribución exponencial negativa.

### Distribución normal

Suele ser la forma más habitual de distribuir los valores de una variable fisiológica. Para predecir que el conjunto de datos de una variable determinada presenta una distribución normal es necesario que ésta dependa de diversos factores y que, además, cada uno de ellos ejerza sobre ella una influencia pequeña y que actúe de forma independiente de otros factores.

En nuestro caso, el valor de la presión arterial sistólica (PAS) depende de muchos otros factores, como edad, sexo, raza, antecedentes familiares, cumplimiento terapéutico, idoneidad del tratamiento, peso corporal, etc. Además, cada uno de estos factores ejerce una influencia que es independiente de la ejercida por otros factores. Por último, cada factor ejerce una repercusión relativamente pequeña frente al conjunto de todos los factores. Pues bien, al cumplirse estas tres condiciones, podemos suponer que los valores de la variable PAS siguen una distribución normal.

## Análisis de una distribución normal

En un análisis preliminar de los datos de la distribución normal se estiman dos índices que caracterizan este tipo de distribución, su media (μ) y su desviación típica o estándar (σ).

Imaginemos los valores de PAS obtenidos de una muestra de 50 individuos (tabla 1).

Si se representan las frecuencias relativas o las probabilidades de obtener una determinada PAS en un paciente, frente a los diversos valores de PAS se observa una distribución con una forma acampanada, relativamente simétrica, que constituye la forma típica de la distribución normal (fig. 1). Para ello se procede en Excel, mediante la función HISTOGRAMA, en HERRAMIENTAS/ANÁLISIS DE DATOS.

Teóricamente, la PAS puede tomar cualquier valor. Sin embargo, en la práctica se observará que habrá muchos individuos que tienen una PAS próxima al valor estimado de la media aritmética (μ: 147,66 mmHg) y que habrá valores más alejados, tanto por exceso como por defecto, tanto más cuanto mayor sea la desviación estándar calculada (σ: 18,77). Ello se debe a que la mayor parte de los datos, aproximadamente el 95%, se encuentra entre la media más dos desviaciones típicas y la media menos dos desviaciones estándar.

**Fig. 1. Representación gráfica de la distribución de valores respecto de su frecuencia relativa y su probabilidad.**

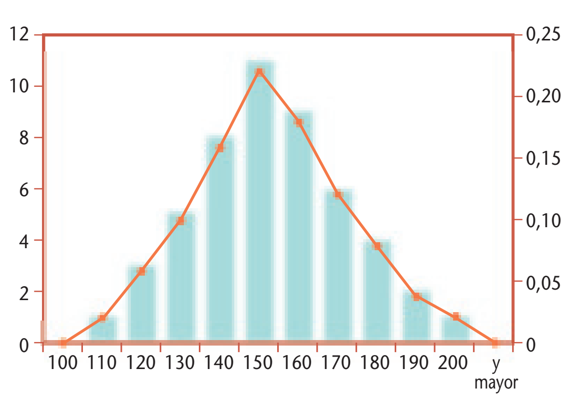


**Tabla 1.** Valores de PAS tomados de una hipotética muestra de 50 individuos

| | | | | |
|---|---|---|---|---|
| 118 | 156 | 148 | 136 | 171 |
| 138 | 164 | 152 | 163 | 152 |
| 162 | 123 | 148 | 134 | 153 |
| 151 | 192 | 172 | 184 | 162 |
| 171 | 144 | 149 | 116 | 138 |
| 134 | 147 | 149 | 162 | 159 |
| 159 | 152 | 133 | 124 | 146 |
| 106 | 143 | 136 | 181 | 145 |
| 132 | 129 | 145 | 125 | 122 |
| 153 | 168 | 171 | 116 | 149 |

PAS: presión arterial sistólica.

Un posterior análisis evaluará si realmente los datos que tenemos de cada variable siguen o no una distribución normal, pues ello es de vital importancia para aplicar posteriormente sobre ellos un tipo u otro de análisis de datos.

Para proceder, primero se observará la forma del gráfico de la curva (fig. 1) y se comparará con la de una representación simétrica perfecta. A continuación se estiman en Excel los coeficientes de curtosis y de asimetría, que deberán ser lo más próximo posible a 0. La curtosis representa el achatamiento o elevación de la distribución. En el ejemplo que nos ocupa, la curtosis se determina en Excel, mediante la función CURTOSIS, simplemente indicando la matriz de datos obtenidos en la variable, con lo que se obtiene un valor de –0,208 que, al ser negativo, se corresponde con una distribución más plana respecto a una distribución normal perfecta; si hubiera sido positiva indicaría una distribución más elevada.

El coeficiente de asimetría caracteriza el grado de asimetría de la distribución de valores con respecto a su media, puesto que en una distribución normal perfecta la curva es totalmente simétrica respecto de la media. En el ejemplo de referencia, el coeficiente, estimado en Excel mediante la función COEFICIENTE.ASIMETRÍA, ofrece un valor de 0,037, lo que indica una distribución ligeramente sesgada hacia los mayores valores; si hubiera sido negativo, indicaría una desviación hacia los menores o negativos.

Finalmente se indica que hay métodos matemáticos para evaluar la normalidad de una distribución de datos, como el test de Kolmogorov-Smirnov, que cuantifican más exactamente las desviaciones de nuestra distribución. ■

PDF
Consultora científica