Medidas descriptivas

Introducción

Los fenómenos que se observan sometidos al azar no suelen ser constantes, por lo que será necesario que junto a una medida que indique el valor alrededor del cual se agrupan los datos, se disponga de una medida que haga referencia a la variabilidad que refleje dicha fluctuación. En este sentido pueden examinarse varias características, siendo las más comunes: la tendencia central de los datos, la dispersión o variación con respecto a este centro, los datos que ocupan ciertas posiciones, la simetría de los datos y la forma en la que los datos se agrupan.

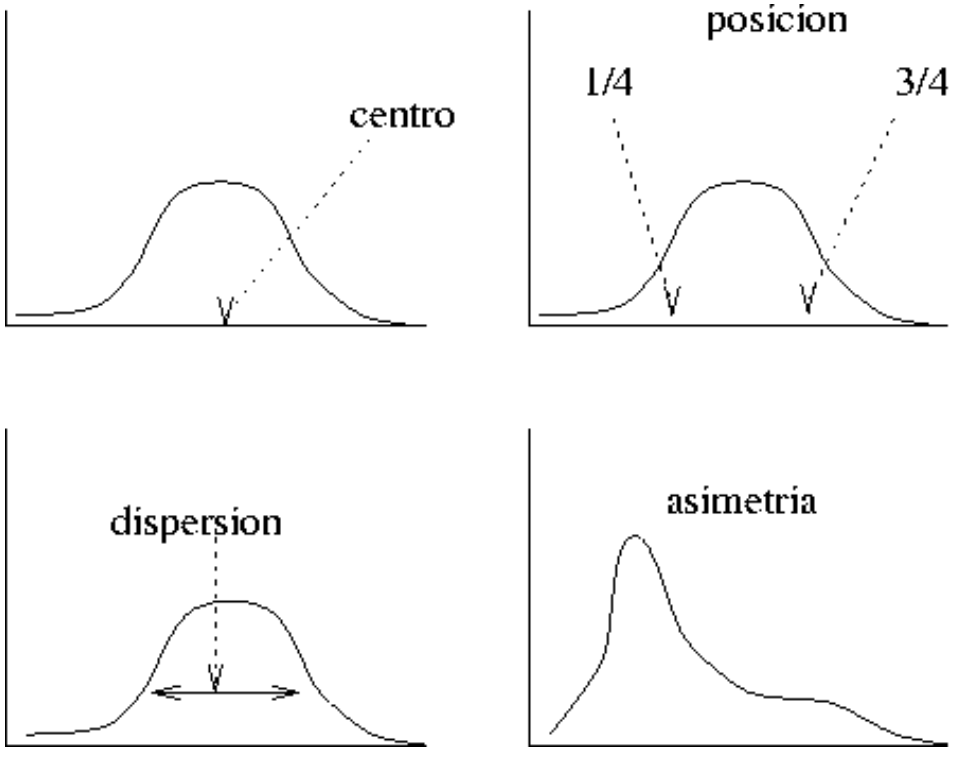


Figura 1: Medidas representativas de un conjunto de datos estadísticos

A lo largo de este tema, y siguiendo este orden, iremos estudiando los estadísticos que nos van a orientar sobre cada uno de estos niveles de información: valores alrededor de los cuales se agrupa la muestra, la mayor o menor fluctuación alrededor de esos valores, nos interesaremos en ciertos valores que marcan posiciones características de una distribución de frecuencias así como su simetría y su forma.

Medidas de centralización

Dependiendo del tipo de variable se pueden considerar diferentes medidas.

Media Aritmética

La media aritmética de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos. Es decir, si la tabla de valores de una variable X es

X	n_i	f_i
x_1	n_1	f_1
...	...	...
x_k	n_k	f_k

la media es el valor que podemos escribir de las siguientes formas equivalentes:

$$\overline{x} = x_1 f_1 + \cdots + x_k f_k = \frac{1}{n}\left(x_1 n_1 + \cdots + x_k n_k\right) = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

Si los datos no están ordenados en una tabla, entonces

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

Hemos supuesto implícitamente en la definición de media que tratábamos con una variable X discreta. Si la variable es continua y se utilizan tablas con intervalos se tienen que cambiar los valores de x_i por las marcas de clase correspondientes c_i. En general, la media aritmética obtenida a partir de las marcas de clase c_i, diferirá de la media obtenida con los valores reales, x_i. Es decir, habrá una perdida de precisión que será tanto mayor cuanto mayor sea la diferencia entre los valores reales y las marcas de clase, o sea, cuanto mayores sean las longitudes a_i, de los intervalos.

Proposición:

La suma de las diferencias de la variable con respecto a la media es nula, es decir,

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

Demostración:

Basta desarrollar el sumatorio para obtener

$$\sum_{i=1}^{n} (x_i - \overline{x}) = (x_1 - \overline{x}) + \cdots + (x_n - \overline{x}) = (x_1 + \cdots + x_n) - n\overline{x} = n\overline{x} - n\overline{x} = 0$$

Ejemplo

Obtener las desviaciones con respecto a la media en la siguiente distribución y comprobar que su suma es cero.

$l_{i-1} - l_i$	n_i
0 - 10	1
10 - 20	2
20 - 30	4
30 - 40	3

Solución:

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})\, n_i$
0 - 10	1	5	5	-19	-19
10 - 20	2	15	30	-9	-18
20 - 30	4	25	100	+1	+4
30 - 40	3	35	105	+11	+33
	$n = 10$		$\sum_{i=1}^{k} x_i n_i = 240$		$\sum_{i=1}^{k} (x_i - \overline{x})\, n_i = 0$

La media aritmética es:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i = \frac{240}{10} = 24$$

Como se puede comprobar sumando los elementos de la última columna la suma es 0.

Linealidad de la media

Si $Y = a + bX$ entonces la correspondiente media de Y es

$$\overline{y} = a + b\overline{x},$$

es decir, el *operador media* es una función *lineal*.

Observaciones

No tiene sentido su cálculo en variables de tipo cualitativo o nominal (media de *sí* y *no*...).

Inconvenientes de la media:

- Es muy sensible a los valores extremos de la variable: todas las observaciones intervienen en el cálculo de la media, así, la aparición de una observación extrema hará que la media se desplace en esa dirección.
- No es recomendable usar la media como medida central en las distribuciones muy asimétricas.
- Depende de la división en intervalos en el caso de utilizar tablas estadísticas.
- Si se considera una variable discreta, por ejemplo el número fallos de un sistema, el valor de la media puede no pertenecer al conjunto de posibles valores que pueda tomar la variable.

Cálculo abreviado

Se puede utilizar la propiedad de la linealidad de la media para simplificar las operaciones necesarias para su cálculo mediante un *cambio de origen* y de *unidad de medida*, en el caso de tener datos con muchos dígitos.

Otras tipos de medias

En función del tipo de problema se pueden considerar varias posibles generalizaciones de la media aritmética. He aquí algunas de ellas aplicadas a un conjunto de posibles observaciones $x_1, \ldots, x_n$.

Media geométrica

$$\overline{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n},$$

alternativamente se puede tomar el antilogaritmo de

$$\ln(\overline{x}_G) = \frac{\ln(x_1) + \cdots + \ln(x_n)}{n}.$$

Se recomienda su uso cuando se tienen porcentajes, tasas o números índice; es decir, cuando una variable presenta variaciones acumulativas.

Media armónica

Se define como la *inversa de la media de las inversas* de las observaciones:

$$\begin{aligned} \overline{x}_A &= \left(\frac{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}{n} \right)^{-1} = \\ &= \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}. \end{aligned}$$

Se suele usar cuando se promedian variables como productividades, velocidades o rendimientos.

Media cuadrática

Es la raíz cuadrada de la media aritmética de los cuadrados de las observaciones:

$$\overline{x}_C = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}}.$$

Mediana

Se considera una variable X cuyas observaciones han sido **ordenadas** de menor a mayor. Llamaremos mediana, Me, al primer valor de la variable que deja por debajo de sí al 50 % de las observaciones. Por tanto, si n es el número de observaciones, la mediana corresponderá a la observación que ocupa la posición $[n/2] + 1$ (donde representamos por $[\cdot]$ la *parte entera* de un número), si el número de datos es impar, y la semisuma de los valores que ocupan las posiciones $n/2$ y $n/2 + 1$, si el número de datos es par.

Entre las propiedades de la mediana, se pueden destacar las siguientes:

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.

- Es de cálculo rápido y de interpretación sencilla, pero no tiene sentido su cálculo en variables de tipo cualitativo o nominal, al igual que la media.

- A diferencia de la media, la mediana de una variable es siempre un valor de la variable que se estudia (ej. La mediana de una variable número de hijos toma siempre valores enteros).

- El mayor defecto de la mediana es que tiene unas propiedades matemáticas complicadas, lo que hace que sea muy difícil de utilizar en *Inferencia Estadística*.

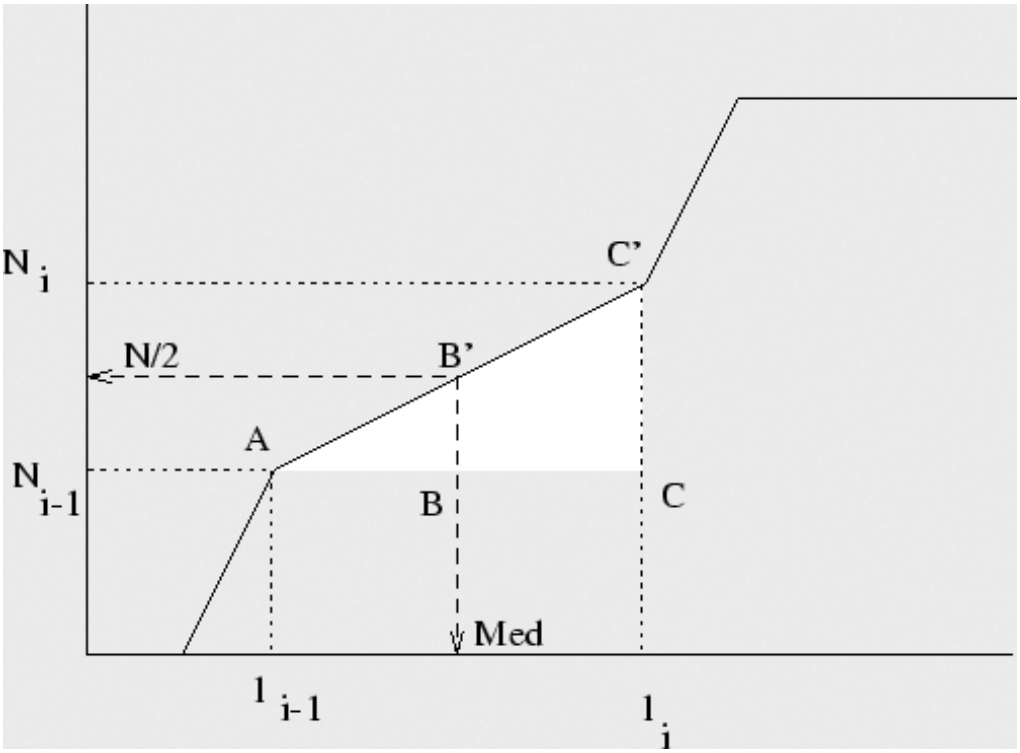
Observación

En el caso de las variables continuas agrupadas en intervalos, el cálculo de la mediana es algo más complicado. Se supone que la mediana se encuentra en un intervalo dado $(l_{i-1}, l_i]$ y hay que determinar el punto que deja exactamente la mitad de observaciones a un lado y al otro. Mediante un argumento geométrico se deduce que la mediana es el valor tal que

$$Me = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i,$$

donde l_{i-1} es el extremo inferior del intervalo donde se encuentra el valor de la mediana, n es el tamaño total de la muestra, n_i es la frecuencia absoluta que aparece en el intervalo donde se encuentra el valor de la mediana y a_i es la amplitud de dicho intervalo.

NOTA: Si se utiliza interpolación lineal, aplicando el teorema de Thales:



$$\begin{aligned} \frac{CC'}{AC} &= \frac{BB'}{AB} \Longrightarrow \\ \frac{n_i}{a_i} &= \frac{\frac{n}{2} - N_{i-1}}{Me - l_{i-1}} \Longrightarrow \\ Me &= l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i \end{aligned}$$

Ejemplo

Sea X una variable discreta que ha presentado sobre una muestra las siguientes modalidades:

$$X \sim 2, 5, 7, 9, 12$$

$\overline{x} = \frac{x_1+\cdots+x_n}{n} = \frac{2+5+7+9+12}{5} = 7$

$Med = 7$, ya que es el valor que deja por detrás dos observaciones y por delante otras dos (está en medio de todas las observaciones).

Si cambiamos la última observación por otra anormalmente grande, esto no afecta a la mediana, pero sí a la media:

$$X \sim 2, 5, 7, 9, 125$$

$\overline{x} = \frac{x_1+\cdots+x_n}{n} = \frac{2+5+7+9+125}{5} = 29{,}6$ y $Me = 7$.

En este caso la media no es un posible valor de la variable (discreta), y se ha visto muy afectada por la observación extrema. Este no ha sido el caso para la mediana.

Moda

Llamaremos moda a cualquier máximo relativo de la distribución de frecuencias, es decir, cualquier valor de la variable que posea una frecuencia mayor que su anterior y su posterior.

De la moda se puede destacar las siguientes propiedades:

- Es muy fácil de calcular y se puede considerar para variables tanto cuantitativas como cualitativas.
- Puede no ser única.

Estadísticos de posición

Para una variable discreta, se define el percentil de orden k, como la observación, P_k, que deja por debajo de sí el $k\,\%$ de la muestra. Esta definición es semejante a la de la mediana, pues como consecuencia de la definición, es evidente que $Me = P_{50}$.

Por su propia naturaleza, el percentil puede estar situado en cualquier lugar de la distribución, por lo que no se puede considerar como una medida de tendencia central, sino más bien de *posición*.

Los cuartiles, Q_l, son un caso particular de los percentiles. Hay 3, y se definen como:

Primer Cuartil: $Q_1 = P_{25}$.

Segundo Cuartil: $Q_2 = P_{50}$. Es equivalente a la Mediana.

Tercer Cuartil: $Q_3 = P_{75}$.

De forma análoga se pueden definir los deciles como los valores de la variable que dividen a las observaciones en 10 grupos de igual tamaño. Más precisamente, definimos $D_1, D_2, \ldots, D_9$ como:

$$D_i = P_{10i},$$

donde $i = 1, \ldots, 9$.

Los percentiles (que incluyen a la mediana, cuartiles y deciles) también son denominados *estadísticos de posición*.

Al igual que en el caso del cálculo de la mediana cuando las variables son continuas y están agrupadas en intervalos, el cálculo de estos estadísticos de posición es más complicado. Se supone que el valor se encuentra en un intervalo dado $(l_{i-1}, l_i]$ y hay que determinar el punto que deja exactamente el porcentaje correspondiente de observaciones a un lado y al otro. Mediante un argumento geométrico se deduce también que el percentil es el valor tal que

$$P_k = l_{i-1} + \frac{n\frac{k}{100} - N_{i-1}}{n_i} \cdot a_i.$$

Ejemplo

Dada la siguiente distribución en el número de hijos de cien familias, calcular sus cuartiles.

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
	$n = 100$	

Solución:

1. **Primer cuartil**: $\frac{n}{4} = 25$; Es el primer valor tal que $N_i > n/4 = 39$; luego $Q_1 = 2$.
2. **Segundo cuartil**: $\frac{2n}{4} = 50$; Es el primer valor tal que $N_i > n/2 = 65$; luego $Q_2 = 3$.
3. **Tercer cuartil**: $\frac{3n}{4} = 75$; Es el primer valor tal que $N_i > 3n/4 = 85$; luego $Q_3 = 4$.

Medidas de variabilidad o dispersión

Los estadísticos de tendencia central o posición indican dónde se sitúa un grupo de puntuaciones. Los de variabilidad o dispersión nos indican si esas puntuaciones o valores están próximos entre sí o si por el contrario están o muy dispersas.

Una medida razonable de la variabilidad podría ser la **amplitud** o **rango**, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto. Es fácil de calcular y sus unidades son las mismas que las de las observaciones originales, aunque posee varios inconvenientes:

- No utiliza todas las observaciones (sólo dos de ellas).
- Se puede ver muy afectada por alguna observación extrema.
- El rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

Existen medidas de dispersión mejores que ésta y se determinan en función de la distancia entre las observaciones y algún estadístico de tendencia central.

Desviación media

Se define la desviación media como la media de las diferencias en valor absoluto de los valores de la variable a la media, es decir, si tenemos un conjunto de n observaciones, $x_1, \ldots, x_n$, entonces

$$D_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}| \, .$$

Como se observa, la desviación media guarda las mismas dimensiones que las observaciones. La suma de valores absolutos es relativamente sencilla de calcular, pero esta simplicidad tiene un inconveniente: desde el punto de vista geométrico, la distancia que induce la desviación media en el espacio de observaciones no es la *natural* (no permite definir ángulos entre dos conjuntos de observaciones). Esto hace que no sea muy conveniente trabajar con ella cuando se considera inferencia estadística.

Nota: Como forma de medir la dispersión de los datos se tiene que descartar $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$, pues esa suma vale 0, ya que las desviaciones con respecto a la media se pueden compensan unas con otras al haber términos en esa suma que son de signos distintos.

Rango Intercuartílico

Se define como la diferencia entre el tercer y el primer cuartil

$$IQR = Q_3 - Q_1$$

Es preferible usarlo cuando aparecen observaciones anómalas (*outliers*).

Varianza y desviación estándar

Si las desviaciones con respecto a la media las consideramos al cuadrado, $(x_i - \overline{x})^2$, de nuevo obtenemos que todos los sumandos tienen el mismo signo (positivo). Esta es además la forma de medir la dispersión de los datos de forma que sus propiedades matemáticas son más fáciles de utilizar. Se pueden definir, entonces, dos estadísticos fundamentales: La varianza y la desviación estándar (o típica).

La varianza, σ^2, se define como la media de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética, es decir

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

Una fórmula equivalente para el cálculo de la varianza está basada en lo siguiente:

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i} + \frac{1}{n}n\overline{x}^2 = \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 + \overline{x}^2 = \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2.
\end{aligned}$$

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en *metros*, la varianza lo hace en $metros^2$) Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la **desviación estándar**, σ, como

$$\sigma = +\sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}.$$

Ejemplo

Calcular la varianza y desviación típica de las siguientes cantidades medidas en metros: 3,3,4,4,5

Solución: Para calcular dichas medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a medir las diferencias. Éste es la media:

$$\overline{x} = \frac{3+3+4+4+5}{5} = 3{,}8 \text{ metros.}$$

La varianza es:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{3^2+3^2+4^2+4^2+5^2}{5} - 3{,}8^2 = 0{,}56 \text{ metros}^2,$$

y la desviación estándar es $\sigma = +\sqrt{\sigma^2} = \sqrt{0{,}56} = 0,748$ metros.

Las siguientes propiedades de la varianza (respectivamente, desviación típica) son importantes a la hora de hacer un cambio de origen y escala a una variable. En primer lugar, la varianza (respectivamente desviación típica) no se ve afectada si al conjunto de valores de la variable se le añade una constante. Si además cada observación es multiplicada por otra constante, en este caso la varianza cambia con relación al cuadrado de la constante (respectivamente la desviación típica cambia con relación al valor absoluto de la constante). Esto queda precisado en la siguiente proposición:

Proposición

Si $Y = aX + b$ entonces $\sigma_y^2 = a^2\sigma_x^2$.

Demostración: Para cada observación x_i de X, $i = 1, \ldots, n$, tenemos una observación de Y que es por definición $.y_i = ax_i + b$. Como $\overline{y} = a\overline{x} + b$, .la varianza de Y es

$$\begin{aligned}\sigma_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(ax_i + b) - (a\overline{x} + b)\right]^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n} a^2 (x_i - \overline{x})^2 = a^2\sigma_x^2.\end{aligned}$$

Las consecuencias del anterior resultado eran de esperar: Si los resultados de una medida son trasladados una cantidad b, la dispersión de los mismos no aumenta. Si estos mismos datos se multiplican por una cantidad $a < 1$, el resultado tenderá a concentrarse alrededor de su media (menor varianza). Si por el contrario $a > 1$ habrá mayor dispersión.

Observaciones.

La varianza y la desviación estándar tienen las siguientes propiedades:

- Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza ya que es función de cada una de las puntuaciones.
- La desviación típica tiene la propiedad de que en el intervalo $(\overline{x} - 2\sigma;\ \overline{x} + 2\sigma)$ se encuentran *por lo menos* el 75 % de las observaciones (es el llamado teorema de Thebycheff). Incluso si tenemos muchos datos y estos provienen de una *distribución normal*, podremos llegar al 95 % de las observaciones.
- No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central, por ejemplo, en datos nominales.

Un principio general de la inferencia estadística afirma que si pretendemos calcular de modo aproximado la varianza de una población a partir de la varianza de una muestra suya, se tiene que el error cometido es generalmente más pequeño, si en vez de considerar como estimación de la varianza de la población, a la varianza muestral consideramos lo que se denomina **cuasivarianza muestral**, s^2, que se calcula como la anterior, pero cambiando el denominador por $n - 1$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n\sigma^2}{n-1}.$$

Tipificación

La tipificación al proceso de restar la media y dividir entre su desviación típica a una variable X. De este modo se obtiene una nueva variable

$$z = \frac{X - \overline{x}}{\sigma},$$

de media 0 y desviación estándar $\sigma_z = 1$, que se denomina variable tipificada.

Esta nueva variable carece de unidades y permite hacer comparables dos medidas que en un principio no lo son, por aludir a conceptos diferentes. Así, por ejemplo, nos podemos preguntar si un *elefante* es más grueso que una *hormiga* determinada, cada uno en relación con su población. También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes. Por ejemplo si deseamos comparar el nivel académico de dos estudiantes de diferentes Universidades para la concesión de una beca de estudios, en principio sería injusto concederla directamente al que posea una nota

media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro. En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad.

Coeficiente de Variación

En el caso de que el problema sea comparar la variabilidad de dos poblaciones con medidas diferentes (variables diferentes) se puede considerar un coeficiente *adimensional* que es el coeficiente de variación. Se define del siguiente modo (como un porcentaje):

$$CV = \frac{\sigma}{\overline{x}} \cdot 100.$$

Observaciones:

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo se deba trabajar con variables positivas, de modo que $\overline{x} > 0$.

- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, entonces $CV_Y > CV_X$, ya que la desviación estándar no es sensible ante cambios de origen, pero sí la media. Lo contrario ocurre si restamos ($b < 0$)

$$CV_Y = \frac{\sigma_y}{\overline{y}} = \frac{\sigma_x}{\overline{x} + b} < \frac{\sigma_x}{\overline{x}} = CV_X$$

- Es invariante a cambios de escala. Si multiplicamos X por una constante a, para obtener $Y = aX$, entonces

$$CV_Y = \frac{\sigma_Y}{\overline{y}} = \frac{\sigma_{ax}}{a\overline{x}} = \frac{a\sigma_x}{a\overline{x}} = CV_X$$

Nota: Es importante destacar que el coeficiente de variación sirve para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones), mientras que si deseamos comparar a dos elementos de cada uno de esos conjuntos, es necesario usar los valores tipificados.

Ejemplo

Dada la distribución de edades (medidas en años) en un colectivo de 100 personas, obtener: La variable tipificada Z, los valores de la media y varianza de Z, el coeficiente de variación de Z.

x_i	n_i
2	47
7	32
15	17
30	4
	$n = 100$

Solución:

Se construye la siguiente tabla auxiliar para realizar los cálculos

x_i	n_i	$x_i n_i$	$x_i^2 n_i$
2	47	94	188
7	32	224	1568
15	17	255	3825
30	4	120	3600
	$n = 100$	693	9181

Para calcular la variable tipificada

$$Z = \frac{X - \overline{x}}{\sigma_x},$$

partimos de los datos del enunciado. Será necesario calcular en primer lugar la media y desviación típica de la variable original (X = años).

$$\begin{aligned} \overline{x} &= \frac{693}{100} = 6{,}93 \text{ años} \\ \sigma_x^2 &= \frac{9181}{100} - 6{,}93^2 = 43{,}78 \text{ años}^2 \\ \sigma_x &= \sqrt{43{,}78} = 6{,}6 \text{ años} \end{aligned}$$

A su vez los valores tipificados son

$$\begin{aligned} z_1 &= \frac{2 - 6{,}93}{6{,}6} = -0{,}745 \\ z_2 &= \frac{7 - 6{,}93}{6{,}6} = 0{,}011 \\ z_3 &= \frac{15 - 6{,}93}{6{,}6} = 1{,}22 \\ z_4 &= \frac{30 - 6{,}93}{6{,}6} = 3{,}486 \end{aligned}$$

z_i	n_i	$z_i n_i$	$z_i^2 n_i$
-0.745	47	-35.015	26.086
0.011	32	0.352	0.004
1.220	17	20.720	25.303
3.486	4	13.944	48.609
	$n = 100$	0.021	100.002

$$\begin{aligned} \overline{z} &= \frac{0{,}021}{100} \approx 0 \\ \sigma_z^2 &= \frac{100{,}002}{100} - 0 \approx 1 \\ \sigma_z &= \sqrt{1} = 1 \end{aligned}$$

Nota: El coeficiente de variación no se puede usar con variables tipificadas, aquí, por ejemplo, se tendría que

$$CV = \frac{\sigma_z}{\overline{z}} = \frac{1}{0} \to \infty.$$

Asimetría y apuntamiento

Nos vamos a plantear averiguar si los datos se distribuyen de forma simétrica con respecto a un valor central, o si bien la gráfica que representa la distribución de frecuencias es de una forma diferente del lado derecho que del lado izquierdo.

Si la simetría ha sido determinada, podemos preguntarnos si la curva es más o menos apuntada (larga y estrecha). Este apuntamiento habrá que medirlo comparándolo con cierta distribución de frecuencias que consideramos *normal* (no por casualidad es éste el nombre que recibe la distribución de referencia).

Coeficiente de asimetría

Para saber si una distribución de frecuencias es simétrica, hay que precisar con respecto a qué. Un buen candidato es la mediana, ya que para variables continuas, divide al histograma de frecuencias en dos partes de igual área. Podemos basarnos en ella para, de forma natural, decir que una distribución de frecuencias es simétrica si el lado derecho de la gráfica (a partir de la mediana) es la imagen por un espejo del lado izquierdo (figura 2).

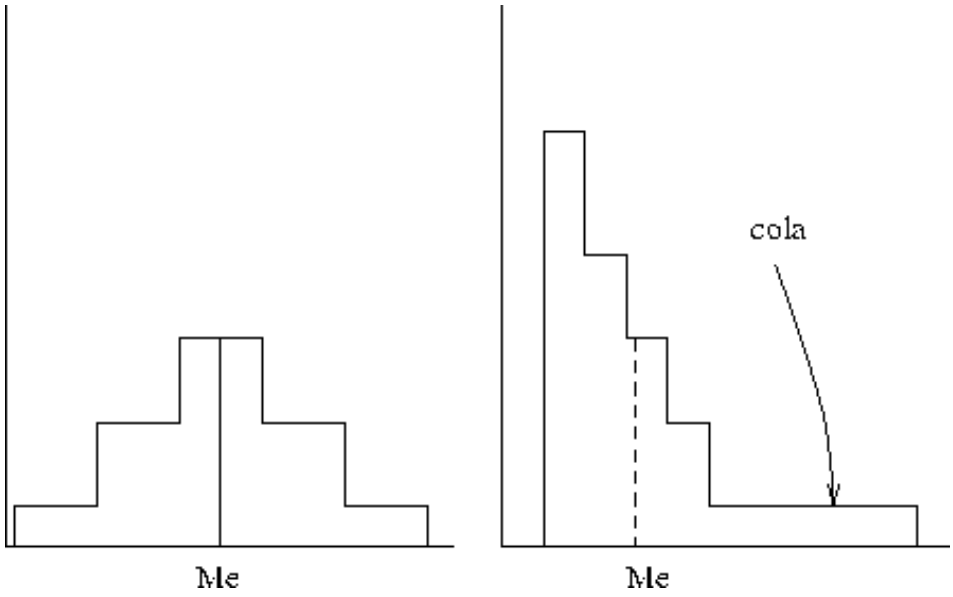


Figura 2: Distribuciones de frecuencias simétricas y asimétricas

Observación: Si una variable es continua simétrica y unimodal, coinciden la media, la mediana y la moda.

Dentro de los tipos de asimetría posible, se pueden destacar los dos fundamentales (figura 3):

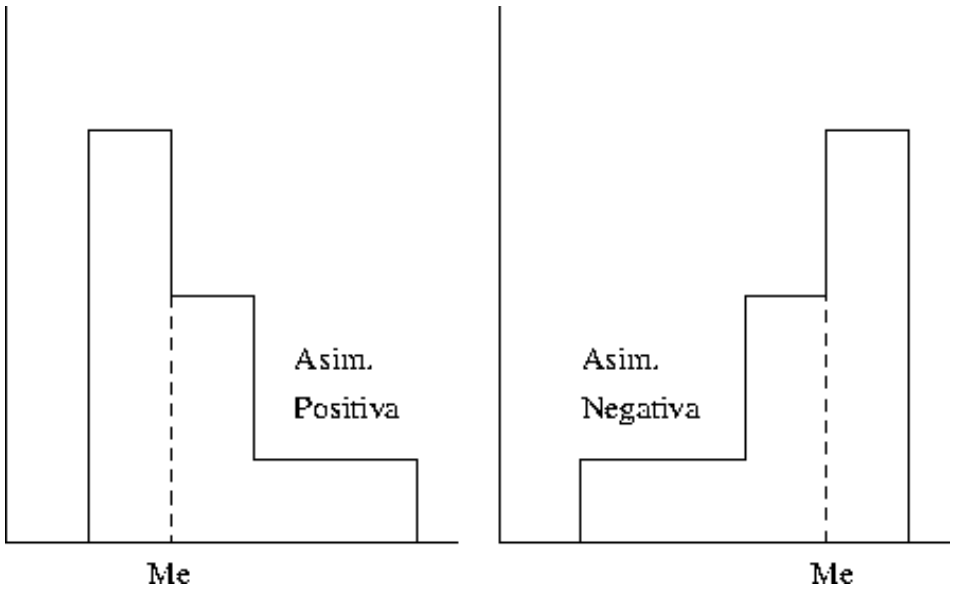


Figura 3: Asimetría positiva y asimetría negativa

Asimetría positiva: Si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en la parte derecha hay frecuencias más pequeñas (en la cola).

Asimetría negativa: Cuando la cola está en el lado izquierdo.

Se puede definir un **coeficiente de asimetría**. Previamente, se tiene que definir el momento de orden $p \in \mathbb{N}$

$$\mu_p = \frac{1}{n} \sum_{i=1}^{n} x_i^p,$$

y el el momento *central* de orden $p \in \mathbb{N}$

$$m_p = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^p ,$$

entonces el coeficiente de asimetría se define como

$$\gamma_1 = \frac{m_3}{m_2 \sqrt{m_2}} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^3}{\sigma^3}$$

Apoyándonos en este índice, se dice que hay asimetría positiva si $\gamma_1 > 0$, y que la asimetría es negativa si $\gamma_1 < 0$.

Se puede observar que es un índice *adimensional*, sin unidades de medida. Por otro lado, sucede que diferencias importantes entre la media y la moda o la media y la mediana indican asimetría.

Coeficiente de curtosis o apuntamiento

Se define el coeficiente de curtosis de Fisher como:

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3,$$

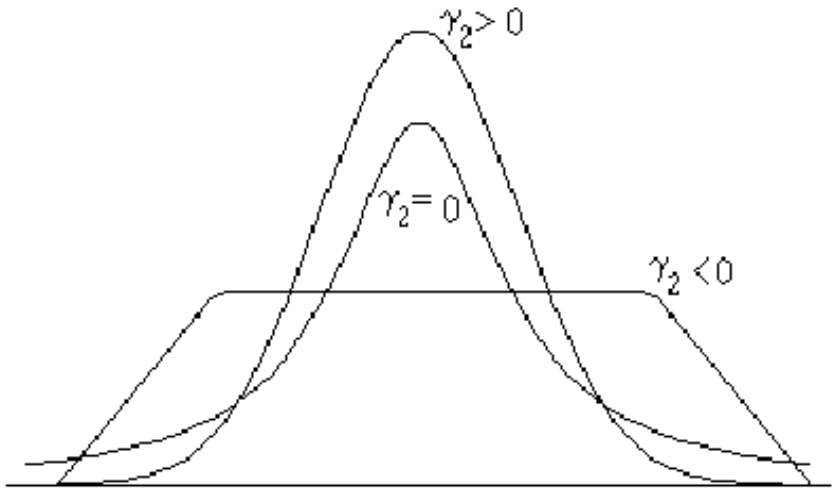
donde m_4 es el momento de cuarto orden. Es éste un coeficiente adimensional, invariante ante cambios de escala y de origen. Sirve para medir si una distribución de frecuencias es muy apuntada o no. Para decir si la distribución es larga y estrecha, hay que tener un patrón de referencia. El patrón de referencia es la distribución normal o gaussiana para la que se tiene

$$\frac{m_4}{\sigma^4} = 3, \text{ lo que implica que } \gamma_2 = 0.$$

Dependiendo del valor alcanzado por γ_2 se puede clasificar una distribución en

leptocúrtica: Si $\gamma_2 > 0$ (más apuntada que la normal)

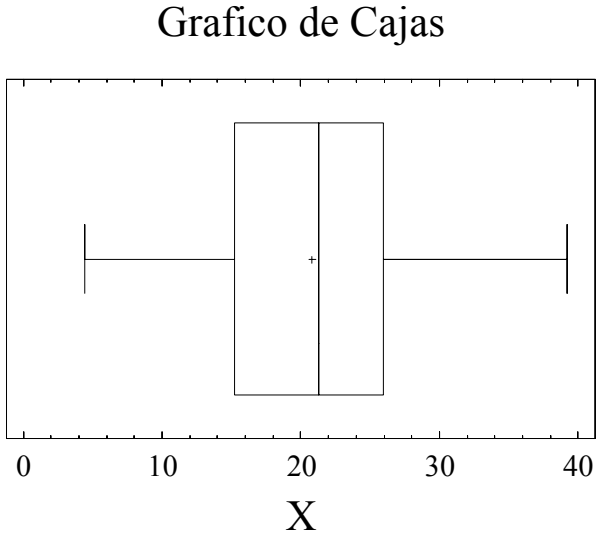
platicúrtica: Si $\gamma_2 < 0$ (más aplastada que la normal).



Otros tipos de métodos gráficos

Diagramas de Cajas (Box-Plots)

Los datos pueden quedar caracterizados por la mediana, el primer y el tercer cuartil. Con todos ellos se puede construir un diagrama en forma de caja



En los extremos de la caja se ponen los cuartiles (1^o y 3^o) y dentro de la caja se pone la mediana. En las *patas* se ponen los siguientes valores: $Q_1 - 1{,}5\,(Q_3 - Q_1)$ y $Q_3 +$

$1{,}5\,(Q_3 - Q_1)$ respectivamente. Es decir, el primer cuartil más o menos vez y media el rango intercuartílico.

Observando la forma de la caja, la posición de la mediana y la longitud de las *patas* se obtiene bastante información sobre la distribución de los datos, la simetría de los mismos, la dispersión y la tendencia central de los mismos.

Cuando aparecen observaciones anómalas (*outliers*) se sitúan como puntos aislados después de los extremos de las *patas.*

Nota: Se consideran observaciones otuliers aquellas que están situadas a una distancia de la media mayor que 3 desviaciónes estándar (σ), es decir o mucho mayores que la media o mucho menores.

Gráfico de tallo y hojas

Es un procedimiento semigráfico para presentar la información que resulta adecuado para un número pequeño de datos. El procedimiento que se emplea es el siguiente:

- Se redondean los datos a dos o tres cifras significativas
- Se colocan los datos en una tabla con dos columnas separadas con una línea. A la izquierda de la línea se ponen, formando el tallo, los dígitos de las decenas o de las centenas (con 3 cifras) y decenas. A la derecha se ponen las unidades en ambos casos formando las hojas. Cada tallo define una clase y sólo se escribe una vez. El número de hojas representa la frecuencia de esa clase.

Ejemplo: Salida de Statgraphics:

```
DATOS:  114, 125, 114, 124, 142, 152, 133, 113, 172, 127, 135,
161, 122, 127, 134, 147

Stem-and-Leaf Display for x1: unit = 1,0   1|2 represents 12,0

      3    11|344
      8    12|24577
```

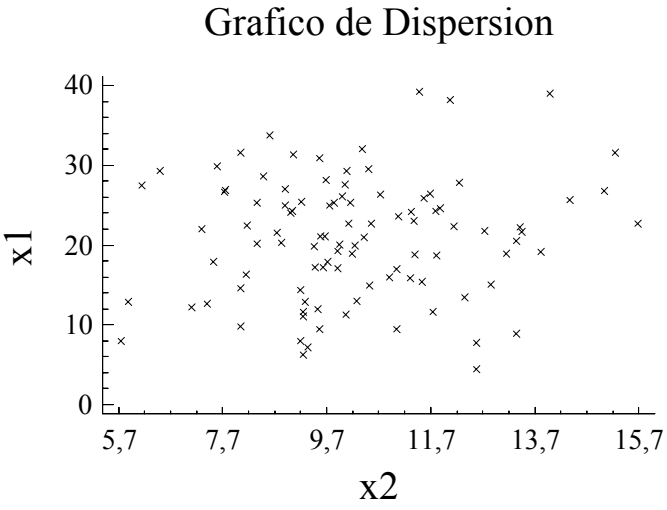
```
8     13|345
5     14|27
3     15|2
2     16|1
1     17|2
```

```
The StatAdvisor
---------------

   This display shows a frequency tabulation for x1.  The range of the
data has been divided into 7 intervals (called stems), each
represented by a row of the table.  The stems are labeled using
one or more leading digits for the data values falling within that
interval. On each row, the individual data values are represented
by a digit (called a leaf) to the right of the vertical line.
This results in a histogram of the data from which you can recover
at least two significant digits for each data value.  If there are
any points lying far away from most of the others (called outside
points), they are placed on separate high and low stems.  In this
case, there are no outside points.  Outside points are illustrated
graphically on the box-and-whisker plot, which you can access via
the list of Graphical Options.  The leftmost column of numbers are
depths, which give cumulative counts from the top and bottom of
the table, stopping at the row which contains the median.
```

Diagramas de dispersión

En ocasiones se trata con dos variables al mismo tiempo (X, Y). En un primer análisis exploratorio se trataría de reflejar posibles relaciones entre las dos variables, por ejemplo lineales. Es decir cuando aumentan los valores de una también lo hacen los de la otra de manera lineal. Se pueden considerar los gráficos de dispersión, donde en abscisas se ponen los valores de una variable (X_2) y en ordenadas los de la otra (X_1). Considero el programa en Statgraphics donde se muestra un ejemplo donde aparece una relación lineal entre dos variables.



Diagramas de datos temporales

Por último, se pueden considerar gráficos de datos temporales. En abcisas se ponen los instantes de tiempo y en ordenadas los valores de la serie. Se pueden observar tendencias(crecientes o decrecientes) en las series , ciclos y comportamientos estacionales. El siguiente ejemplo es el de un índice medido durante todos los días de un mes.