Descripción numérica de una muestra de datos cuantitativos

Los gráficos considerando anteriormente, nos proporciona una idea de la localización, la dispersión y la asimetría de la distribución. Ahora buscamos medidas numéricas de estas cantidades.

Medidas de localización

La medida más utilizada es la media (aritmética).

Definición 8 *Supongamos que tenemos una muestra* $x_1, \ldots, x_n$*. Entonces,* **la media** *(aritmetica) es*

$$\begin{aligned} \bar{x} &\stackrel{\text{def}}{=} \frac{1}{n}(x_1 + \ldots + x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Se puede usar la media con muestras de datos continuos o discretas.

Ejemplo 32 *Los siguientes datos son los números de críos nacidos conjuntamente para 18 parejas de ratones campestres.*

3	6	5	6	5	7
5	7	6	6	6	5
5	5	4	5	6	4

La media de estos datos es

$$\frac{1}{18}(3 + \ldots + 4) = \frac{96}{18} \approx 5{,}333$$

ratoncitos por pareja.

Ejemplo 33 *Los siguientes datos son los pagos de ayuda social anuales en millones de dolares en 15 estados de los EE.UU. en el año 1990.*

114,95	56,79	115,15	171,75	65,96
39,62	112,28	92,43	164,20	179,37
121,99	113,66	91,95	95,43	160,41

La media es

$$\frac{1}{15}(114{,}95 + \ldots + 160{,}41) \approx 113{,}063$$

millones de dolares por estado.

A menudo, se quiere calcular la media a través de una tabla de frecuencias previamente hecha. Esto es fácil si los datos son discretas.

Ejemplo 34 *Retomamos el Ejemplo 20 sobre las cartas recibidas diariamente por el estadístico. Tenemos la siguiente tabla de frecuencias.*

Número x_i	*Frec. absoluta* n_i	*Frec. relativa* f_i
0	3	0,1
1	3	0,1
2	9	0,3
3	12	0,4
4	3	0,1
> 4	0	0
Total	30	1

Añadimos otra columna a la tabla para representar los valores de $x_i \times f_i$.

x_i	n_i	f_i	$x_i \times f_i$
0	3	0,1	0
1	3	0,1	0,1
2	9	0,3	0,6
3	12	0,4	1,2
4	3	0,1	0,4
> 4	0	0	0
Total	30	1	2,3

La cantidad media de cartas diarias recibidas es 2,3.

Una fórmula general

Supongamos que observamos k valores distíntos x_i con frecuencias absolutas n_i para $i = 1, \ldots, k$. Entonces se construye la siguiente tabla

x_i	n_i	f_i	$x_i \times f_i$
x_1	n_1	$f_1 = \frac{n_1}{n}$	$x_1 \times f_1$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$x_2 \times f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$x_k \times f_k$
Total	n	1	$\bar{x}$

Se tiene

$$\begin{aligned} \bar{x} &= \sum_{i=1}^{k} f_i x_i \\ &= \sum_{i=1}^{k} \frac{n_i}{n} x_i \\ &= \frac{1}{n} \sum_{i=1}^{k} n_i x_i \end{aligned}$$

Ejemplo 35 *Construimos una tabla de frecuencias para los datos del Ejemplo 32*

x_i	n_i	f_i	$x_i \times f_i$
3	1	$\frac{1}{18}$	$\frac{3}{18}$
4	2	$\frac{2}{18}$	$\frac{8}{18}$
5	7	$\frac{7}{18}$	$\frac{35}{18}$
6	6	$\frac{6}{18}$	$\frac{36}{18}$
7	2	$\frac{2}{18}$	$\frac{14}{18}$
Total	18	1	$\frac{96}{18} \approx 5{,}333$

y el resultado es igual a la media calculada directamente.

Si los datos son continuos, sólo se puede aproximar la media a través de la tabla de frecuencias.

En este caso, se aproxima suponiendo que todos los datos en un intervalo están en el centro del intervalo.

Ejemplo 36 *Volvemos al Ejemplo 22. Tenemos la siguiente tabla de frecuencias.*

Clase	*Centro* x_i	n_i	f_i	$x_i \times f_i$
$[10, 20)$	15	6	,1875	2,8125
$[20, 30)$	25	7	,21875	5,46875
$[30, 40)$	35	8	,25	8,75
$[40, 50)$	45	6	,1875	8,4375
$[50, 60)$	55	4	,125	6,875
$[60, 70)$	65	1	,03125	2,01325
Total		32	1	34,357

Se estima la media en 34,357*.*

Volviendo al Ejemplo 22 y haciendo el cálculo mediante los datos originales, se tiene

$$\bar{x} = \frac{1}{32}(42{,}1 + \ldots + 34{,}2) = 33{,}175$$

es el resultado exacto.

El problema con la media

Ejemplo 37 *Supongamos que tenemos una muestra de 99 datos iguales a* 1 *y un* **dato atípico** *igual a* 101*.*

Entonces la media de esta muestra es 2 *que no es una buena medida de la localización de la mayoría de los datos.*

La media se ve muy afectada por la presencia de datos atípicos.

Una medida alternativa y robusta a atípicos es **la mediana**.

Definición 9 *Supongamos que se tiene una muestra de datos ordenados;* $x_1 \leq x_2 \leq \ldots \leq x_n$*. Entonces, si* n *es un número impar, la mediana es* $x_{\frac{n+1}{2}}$ *y si* n *es un número par, la mediana es* $\frac{x_{\frac{n}{2}}+x_{\frac{n+2}{2}}}{2}$*.*

Es decir que la mediana es el dato (o el promedio de los dos datos) más centrico de la muestra.

Ejemplo 38 *Ordenamos los datos en el Ejemplo 33.*

39,62	56,79	65,96	91,95	92,43
95,43	112,28	113,66	114,95	115,15
121,99	160,41	164,20	171,75	179,37

La mediana es 113,66*.*

Ejemplo 39 *Volvemos al Ejemplo 32.*

En primer lugar, ordenamos los datos en la muestra.

3	4	4	5	5	5
5	5	5	5	6	6
6	6	6	6	7	7

$n = 18$ *es un número par y entonces, la mediana es* $\frac{x_9+x_{10}}{2} = \frac{5+5}{2} = 5$*.*

Otras medidas de localización

La **media geométrica** de una muestra $x_1, \ldots, x_n$, se define como

$$\sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Sufre de los mismos problemas como la media aritmetica pero además, si algún de los datos es negativo, puede que no exista.

Un intento para evitar los efectos de atípicos es calcular la **media recortada** de la muestra quitando el valor más alto y el valor más pequeño (o 2 de cada lado etc.) Es un método razonable pero ¿cuántos datos se deben quitar?