Medidas de dispersión

Se quiere medir la dispersión de una muestra a través de su localización. En primer lugar, definimos una medida relacionada con la media.

Ya habiendo calculado la media, $\bar{x}$ de una muestra $x_1, \ldots, x_n$, una posibilidad es calcular las **desviaciones** $x_1 - \bar{x}, \ldots, x_n - \bar{x}$. Una medida de dispersión natural puede ser la media de las desviaciones pero:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) &= \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x} \\ &= \bar{x} - \frac{1}{n}n\bar{x} \\ &= 0 \end{aligned}$$

Claro, algunas desviaciones son positivas y otras negativas.

La varianza y la desviación típica

Una medida alternativa es la varianza.

Definición 10 *Para una muestra* $x_1, \ldots, x_n$ *con media* $\bar{x}$, **la varianza** *de la muestra es*

$$s^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

La varianza está midiendo la media distancia cuadrada de los datos en torno de la media $\bar{x}$.

Observamos que si, por ejemplo las unidades de los datos son metros, entonces las unidades de la varianza son metros cuadrados. Es más natural tener una medida con las mismas unidades que la media.

Definición 11 La desviación típica *es*

$$s \stackrel{\text{def}}{=} \sqrt{s^2}.$$

Ejemplo 40 *Retomamos el Ejemplo 32 sobre los ratoncitos. Calculamos anteriomente que la media es* $5{,}33\dot{3}$. *Ahora calculamos las desviaciones.*

$$3 - 5{,}33\dot{3} \quad 6 - 5{,}33\dot{3} \quad 5 - 5{,}33\dot{3} \ldots \quad 4 - 5{,}33\dot{3}$$

Entonces, la suma de las desviaciones cuadradas es

$$(3 - 5{,}33\dot{3})^2 + \ldots + (4 - 5{,}33\dot{3})^2 = 18$$

y la varianza es $s^2 = 18/18 = 1$. *La desviación típica es también igual a* 1.

Calcular la varianza así es bastante lento. ¿Hay una manera más rápida de hacer el cálculo?

Teorema 1 *Para una muestra* $x_1, \ldots, x_n$ *con media* $\bar{x}$*, se puede expresar la varianza como*

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

Demostración

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\left[x_i^2 - 2x_i\bar{x} + \bar{x}^2\right]\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)
\end{aligned}$$

$\diamond$

Ejemplo 41 *Volvemos a los datos del Ejemplo 32.*

En este caso, la suma de los cuadrados es

$$3^2 + 6^2 + 5^2 + \ldots + 4^2 = 530$$

y entonces, usando el teorema, tenemos que la varianza es

$$s^2 = \frac{1}{18}\left(530 - 18 \times 5{,}33\dot{3}^2\right) = 1$$

igual que el resultado sacado anteriormente en el Ejemplo 40.

Ejemplo 42 *En el Ejemplo 33 calculamos anteriormente que la media ayuda por estado es de* $113{,}063 *millones. La suma de los datos cuadrados es*

$$114{,}95^2 + \ldots + 160{,}41^2 = 212180{,}009$$

Entonces, la varianza es

$$s^2 = \frac{1}{15}\left(212180{,}009 - 15 \times 113{,}063^2\right) = 1362{,}092$$

\$ millones cuadrados. La desviación típica es $s = 36{,}91$ *millones de dolares.*

Cálculo de la varianza a través de la tabla de frecuencias

Igual que con la media, es posible calcular la varianza usando la tabla de frecuencias. El resultado es exacta si los datos son discretas y aproximado si los datos son continuos.

Ejemplo 43 *Retomamos los Ejemplos 32 y 35.*

x_i	n_i	f_i	$x_i \times f_i$	$x_i^2 \times f_i$
3	1	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{9}{18}$
4	2	$\frac{2}{18}$	$\frac{8}{18}$	$\frac{32}{18}$
5	7	$\frac{7}{18}$	$\frac{35}{18}$	$\frac{175}{18}$
6	6	$\frac{6}{18}$	$\frac{36}{18}$	$\frac{216}{18}$
7	2	$\frac{2}{18}$	$\frac{14}{18}$	$\frac{98}{18}$
Total	18	1	$\frac{96}{18} = 5{,}33\dot{3}$	$\frac{530}{18} = 29{,}44\dot{4}$

La varianza es $29{,}44\dot{4} - 5{,}33\dot{3}^2 = 1$, *como calculamos antes en el Ejemplo 40.*

Fórmula general

Observamos k valores distíntos x_i con frecuencias absolutas n_i para $i = 1, \ldots, k$.

x_i	n_i	f_i	$x_i \times f_i$	$x_i^2 \times f_i$
x_1	n_1	$f_1 = \frac{n_1}{n}$	$x_1 \times f_1$	$x_1^2 \times f_1$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$x_2 \times f_2$	$x_2^2 \times f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$x_k \times f_k$	$x_k^2 \times f_k$
Total	n	1	$\bar{x}$	$\sum_{i=1}^k f_i x_i^2$

La varianza es $\sum_{i=1}^k f_i x_i^2 - \bar{x}^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2$.

¿Porqué?

$$\begin{aligned} \sum_{i=1}^k f_i x_i^2 - \bar{x}^2 &= \sum_{i=1}^k \frac{n_i}{n} x_i^2 - \bar{x}^2 \\ &= \frac{1}{n} \left(\sum_{i=1}^k n_i x_i^2 - n\bar{x}^2 \right) \end{aligned}$$

Ejemplo 44 *Volvemos a los Ejemplos 22 y 36.*

Clase	*Centro* x_i	n_i	f_i	$x_i \times f_i$	$x_i^2 \times f_i$
$[10, 20)$	15	6	,1875	2,8125	42,1875
$[20, 30)$	25	7	,21875	5,46875	136,71875
$[30, 40)$	35	8	,25	8,75	306,25
$[40, 50)$	45	6	,1875	8,4375	379,6875
$[50, 60)$	55	4	,125	6,875	378,125
$[60, 70)$	65	1	,03125	2,01325	132,03125
Total		32	1	34,357	1375

Estimamos la varianza y desviación típica con $s^2 \approx 1375 - 34{,}357^2 = 194{,}597$ *y* $s \approx 13{,}950$.

La varianza y desviación típica exacta son 201,25 *y* 14,19 *respectivamente.*

La regla de Chebyshev

Es una regla que pone un límite sobre la dispersión de la mayoría de los datos en torno de la media.

Teorema 2 *Para cualquier conjunto de datos, la proporción de datos que distan menos de m desviaciones típicas de la media es como mínimo*

$$1 - \frac{1}{m^2}$$

Demostración

Creeme, soy el profesor. $\diamond$

Dice, por ejemplo, que por lo menos 75 % de las observaciones están a menos de dos desviaciones típicas de la media y por lo menos, 88,8̇ % de las observaciones están a menos de 3 desviaciones típicas de la media.

Es una regla muy conservador.

Ejemplo 45 *Volvemos al Ejemplo 32 sobre los ratoncitos. En el Ejemplo 35 calculamos, la media como* $5{,}33\dot{3}$ *y en el Ejemplo 40, demostramos que la desviación típica es* 1*. Luego, la regla de Chebyshev dice que por los menos un* $75\,\%$ *de los datos están contenidos en el intervalo* $[3{,}33\dot{3}, 7{,}33\dot{3}]$ *y que el intervalo*

$$5{,}33\dot{3} \pm 3 \times 1 = [2{,}33\dot{3}, 8{,}33\dot{3}]$$

contiene por lo menos un $88{,}88\dot{8}\,\%$ *de los datos.*

Pero, ordenando los datos (o mirando la tabla de frecuencias que construimos en el Ejemplo 35) vemos que sólo hay un dato con valor de 3 *y que todos los demás datos son menores o iguales a* 7*. Entonces el primer intervalo contiene* $\frac{17}{18} \times 100\,\% \approx 94{,}4\,\%$ *de los datos y el segundo intervalo contiene todos los datos en la muestra.*

Una regla empírica

Una **regla empírica** dice que si la distribución de los datos es más o menos simétrica y unimodal, (es decir con una distribución **normal**) entonces aproximadamente un 68 % de los datos caerán dentro de ± 1 desviaciones típicas de la media, 95 % dentro de ± 2 desviaciones y 99,7 % dentro de ± 3 desviaciones típicas de la media.

El coeficiente de variación

Es otra medida de variabilidad que tiene la ventaja de ser sin unidades.

Definición 12 *Para una muestra de datos con media $\bar{x}$ y desviación típica s, se define el coeficiente de variación como*

$$CV = \frac{s}{|\bar{x}|}.$$

Si cambiamos la escala de medir la variable, el coeficiente de variación no cambia. No obstante, si la media es igual a cero, el coeficiente de variación no existe.

La cuasi varianza y la cuasi desviación típica

De vez en cuando se define la varianza (y desviacion típica) con un divisor de $n-1$ en lugar de n, es decir

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

En este caso, el resultado se llama la **cuasi varianza** (y **cuasi desviación típica**).

La razón técnica es que así, la cuasi varianza es un estimador insesgado de la varianza poblacional. (Ver **Estadística 1**).

Es importante observar que automaticamente en Statgraphics, se calcula la cuasi varianza.

Los cuartiles y el rango intercuartílico

Igual que con la media, la varianza y desviación típica son muy sensibles a datos atípicos. Una medida más robusta de la dispersión de los datos es el rango intercuartílico.

Definimos algunos conceptos básicos:

Definición 13 *Si tenemos una muestra ordenada* $x_1 \leq x_2 \leq \ldots \leq x_n$*, entonces el* **rango** *de los datos es la distancia*

$$R = x_n - x_1$$

entre el datos más grande y el dato más pequeño.

Obviamente, el rango es muy sensible a datos atípicos.

Recordamos ahora que la mediana (o segundo cuartil) separa la muestra en dos partes. De manera semejante, definimos las cuartiles.

Definición 14 *El* **primer cuartil**, Q_1, *es la mediana de la primera mitad de la muestra. El* **tercer cuartil**, Q_3, *es la mediana de la segunda mitad de la muestra.*

Los cuartiles dividen la muestra en 4 partes.

Definición 15 *El rango intercuartílico es la diferencia*

$$RI = Q_3 - Q_1.$$

Obviamente el cálculo de los cuartiles depende de si el tamaño de la muestra es un número par o impar.

Supongamos que tenemos un número impar de datos.

Ejemplo 46 *Calculamos la mediana de los pagos de ayuda social (Ejemplo 33) en varios estados de EE.UU. en el Ejemplo 38.*

Los 15 datos ordenados eran

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

y la mediana es 113,66*.*

Hay 7 datos menores que la mediana y su mediana es el cuarto dato $Q_1 = 91{,}95$*. Hay 7 datos mayores que la mediana y su mediana es* $Q_3 = 160{,}41$

Entonces, el rango intercuartílico es $Q_3 - Q_1 = 68{,}46$ *millones de dolares.*

Ahora vemos un ejemplo con un número par de datos.

Ejemplo 47 *Volvemos al Ejemplo 39.*

Tenemos 18 *datos.*

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

Calculamos anteriormente que la mediana es $Q_2 = 5$. *Dividiendo la muestra por la mitad, tenemos dos partes de* 9 *datos. La mediana de la primera mitad es el quinto dato* $Q_1 = 5$ *y la mediana de la segunda mitad es el dato 14,* $Q_3 = 6$.

El rango intercuartílico es $Q_3 - Q_1 = 1$.

Ejemplo 48 *Supongamos que tenemos una muestra de* 9 *datos.*

$$-60\ 4\ 7\ 10\ 12\ 15\ 20\ 25\ 48$$

La mediana es $Q_2 = 12$.

Hay cuatro datos menores y cuatro mayores que la mediana. La mediana de los primeros cuatro datos es $Q_1 = \frac{4+7}{2} = 5{,}5$ *y la mediana de los últimos cuatro datos es* $Q_3 = 22{,}5$.

El rango intercuartílico es $Q_3 - Q_1 = 17$.

Cálculo de la mediana y los cuartiles mediante el diagrama de tallo y hojas

Si construimos un diagrama de tallo y hojas, ya ordenamos los datos, lo que facilita el cálculo rápido de la mediana y los cuartiles.

Ejemplo 49 *Volvemos al Ejemplo 29 sobre emisiones de óxido. Añadimos una columna de frecuencias acumuladas en el tallo.*

4	0	2 2 3 4
11	0	5 6 6 6 7 7 9
18	1	0 0 2 2 3 4 4
29	1	5 5 5 5 5 7 7 7 8 8 9
34	2	0 1 1 2 3
39	2	5 7 7 9 9
40	3	4
44	3	5 6 7 8
46	4	1 2
47	4	5

Tallo		*Unidades*
Hoja		*Decimales*
1|1	=	1,1

Tenemos 47 datos y entonces, la mediana es

$$x_{24} = 1{,}7.$$

El primer quartil es la mediana de los primeros 23 datos,

$$Q_1 = x_{12} = 1{,}0.$$

El tercer quartil es la mediana de los últimos 23 datos, es decir

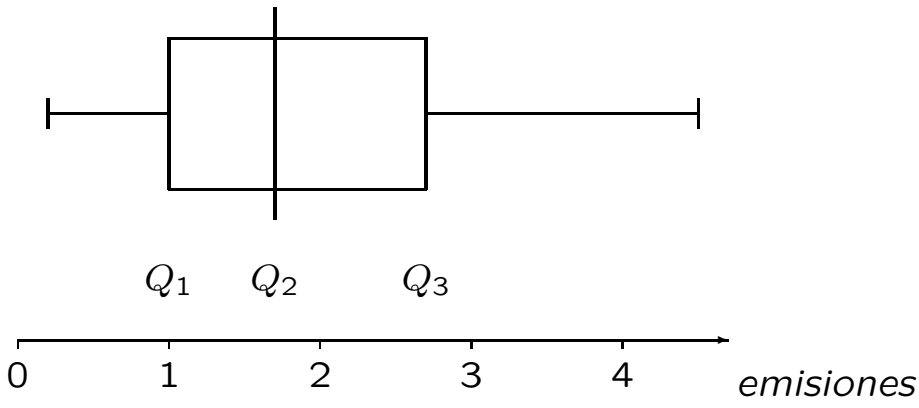
$$Q_3 = x_{36} = 2{,}7$$

El rango intercuartílico es $2{,}7 - 1{,}0 = 1{,}7$.

El diagrama de caja

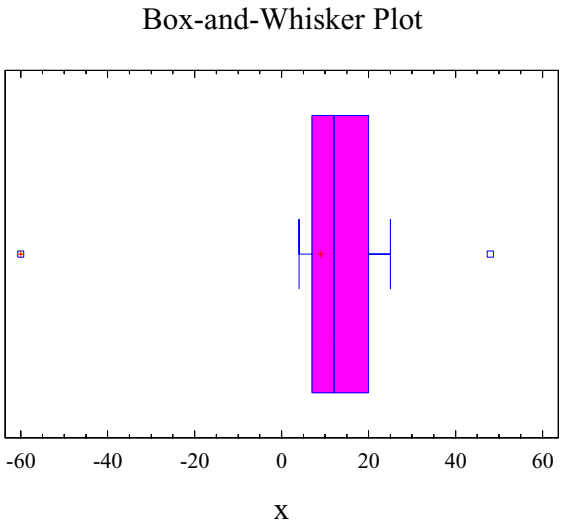
Es una manera visual de ver la mediana, cuartiles, rango y posibles datos atípicos.

Ejemplo 50 *Vemos un diagrama de caja para los datos sobre emisiones del Ejemplo 49.*



La distribución es un poco asimétrica a la derecha pero no parecen datos atípicos.

Ejemplo 51 *Volvemos al Ejemplo 48. Aquí se ve un dato atípico y un dato atípico extremo.*



Fórmula general

- La caja central es la región entre el primer y tercer cuartil.
- Se añade una recta vertical para la mediana.
- Se extiende la recta horizontal a la izquierda hasta el punto más pequeño menos de 1,5 rangos intercuartílicos del primer cuartil. Se marca el final con una linea vertical.
- Se extiende la linea a la derecha hasta el punto más grande menos de 1,5 rangos intercuartílicos del tercer cuartil. Se marca el final
- Datos entre 1,5 y 3 rangos intercuartílicos más bajos (altos) de Q_1 (Q_3) son datos atípicos, indicados en Statgraphics con un cuadro.
- Datos más de 3 rangos intercuartílicos más bajos (altos) de Q_1 (Q_3) son atípicos extremos marcados con un cuadro y una cruz.

Medidas de asimetría y curtosis

Son otros medidas relacionadas con la media y varianza.

Definición 16 *Para una muestra* $x_1, \ldots, x_n$, **el coeficiente de asimetría** *es*

$$CA = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$$

El coeficiente de asimetría vale (aproximadamente) 0 si la distribución es simétrica, es positiva si la distribución es asimétrica a la derecha y es negativa si la distribución es asimétrica a la izquierda.

Definición 17 *El* **coeficiente de apuntamiento** *es*

$$CAp = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4}$$

Mide la cantidad de curtosis o apuntamiento de la distribución relativa a la distribución normal.

Ejemplo 52 *Retomamos los datos sobre emisiones y hacemos los cálculos en Statgraphics.*

```
Summary Statistics for emisiones

Count = 47
Average = 1,8617
Median = 1,7
Variance = 1,33154
Standard deviation = 1,15393
Range = 4,3
Lower quartile = 1,0
Upper quartile=2,7
Interquartile range = 1,7
Skewness = 0,617837
Kurtosis = -0,446736
Coeff. of variation = 61,9823%
```