



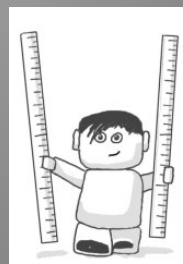
Tema 1. Estadística descriptiva. 2/2

Curso 2024-2025

Grado de Física

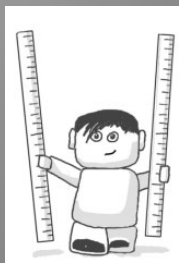
Técnicas Experimentales II

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Objetivos tema:

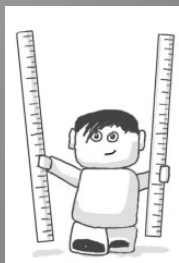
- Introducción al tema, glosario y bibliografía.
- Distribuciones de frecuencias.
- Representaciones gráficas de la información.
- Medidas características: centralización, dispersión, asimetría y apuntamiento.
- Transformaciones de variable aleatoria: lineales y no lineales.
- Muestras multivariantes.



Bibliografía:

- [VAR10] **Tratamiento de Datos Fisicos.** L. M. Varela, F. Gomez, J. Carrete. Servizo de Publicacions e Intercambio Cienti,co. Universidade de Santiago, 2010.
- [VIM12] **Vocabulario Internacional de Metrología** Conceptos fundamentales y generales, y términos asociados (VIM) (**JCGM 200:2012) 3a Ed. CEM**
- [Pena08] **Fundamentos de Estadistica,** D. Peña. Alianza Editorial, 2008.
- [Wonn87] **Introduccion a la estadistica,** T. H. Wonnacott, R. J. Wonnacott. Editorial Limusa, 1987.
- [Spie03] **Probabilidad y estadistica.** Spiegel Murray. Editorial McGrawHill, 2003.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Medidas características

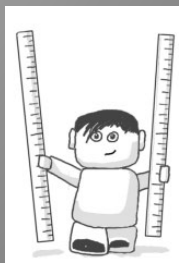
Las medidas características de una distribución de frecuencias son **una serie de parámetros asociados a la distribución que informan sobre propiedades de interés estadístico**, esto es, sobre cómo se distribuyen los resultados de la variable aleatoria. Fundamentalmente son:

- **Medidas de la posición central de la distribución**: medidas que muestran el valor alrededor del que se centran los resultados, los valores más probables, los valores centrales de la distribución … Veremos como ejemplos la media, moda, mediana y los percentiles.

- **Medidas de la dispersión de la distribución**: medidas que muestra la variabilidad que presentan los datos alrededor de sus valores centrales, o como de separados se muestran los datos de sus medidas de la posición central. Nos indican la anchura de las distribuciones de la variable aleatoria.

- **Medidas de la asimetría de los valores en la distribución**: medidas que muestran el grado de simetria alrededor de las medidas de posición centrales, esto es, si los valores se distribuyen homogéneamente o son muy asimétricos.

- **Medida de la concentración de las medidas (apuntamiento o curtosis)**: medidas que muestran la concentración de los datos en valores próximos a las medidas de la posición central (fundamentalmente a la media) o si por el contrario se acumulan en los extremos presentando colas pronunciadas.



Medidas características de centralización

Media (aritmética)(muestral): corresponde al promedio de los valores que constituyen la medida (muestra).

En el caso de una muestra de N medidas con resultados $\{ x_1, x_2, \ldots , x_{N-1}, x_N \}$ la media $\bar{x}$ será:

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

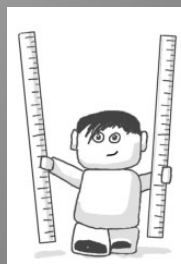
En el caso de variables discretas con valores repetidos, sean $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p\}$ los p valores diferentes de la variable aleatoria que se obtienen con frecuencias relativas $\tilde{f}_l$; $l = 1, \ldots, p$. Entonces podremos escribir:

$$\bar{x} = \sum_{j=1}^{l} \tilde{f}_l \, \tilde{x}_l$$

$$\tilde{x}_i \neq \tilde{x}_j \quad i \neq j$$

Y si agrupamos los *N* resultados en *k* clases, cada una con frecuencias relativas f_j ; $j = 1, \ldots, k$ siendo x'_j los valores de las marcas de clase, entonces:

$$\bar{x} = \sum_{j=1}^{k} f_j \, x'_j$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de centralización

En general denotaremos la media aritmética mediante

$$\bar{x} = \sum_{i=1}^{n} f_i \; x_i$$

Entendiendo que f_i son las frecuencias relativas de los valores x_i de la variable aleatoria (marca de clase, …). La media aritmética de una muestra es a su vez una variable aleatoria.(*)

También podemos evaluar la media de una función $g(x)$ de la variable aleatoria de la forma:

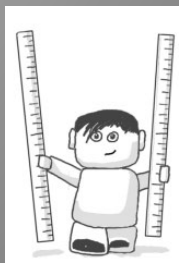
$$\overline{g(x)} = \sum_{i=1}^{n} f_i \; g(x_i)$$

Notaremos el valor medio de la variable al cuadrado $g(x) = x^2$ mediante $\overline{x^2}$:

$$\overline{x^2} = \sum_{i=1}^{n} f_i \; {x_i}^2$$

Obsérvese que en general:

$$\overline{g(x)} \neq g(\bar{x})$$



(*) Si tomamos diferentes muestras de N datos, la media muestral tendrá una cierta distribución de probabilidad que depende de la distribución de probabilidad de los datos individuales de la muestra.

Medidas características de centralización

Media geométrica: corresponde a la raíz n-ésima del productorio de los valores que constituyen la medida.
Al ser un productorio esta media se anula si un dato es cero.

En el caso de *N* medidas con resultados $\{ x_1, x_2, \ldots, x_{N-1}, x_N \}$ la media geometrica sera:

$$\bar{x}_g = \left(\prod_{i=1}^{N} x_i \right)^{1/N}$$

Y si agrupamos los *N* resultados en *k* marcas de clase, cada una con frecuencias absolutas n_i, (con *i* moviéndose en este caso de 1 a *k*), entonces:

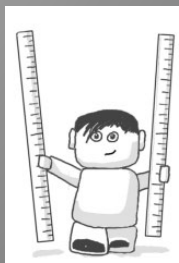
$$\bar{x}_g = \left(\prod_{i=1}^{k} {x_i}^{n_i} \right)^{1/N}$$

O tomando logaritmos:

$$\log(\bar{x}_g) = \frac{1}{N} \sum_{i=1}^{k} n_i \log(x_i)$$

$$\log(\bar{x}_g) = \sum_{i=1}^{k} f_i \log(x_i)$$

Esto establece que el logaritmo de la media geométrica es la media aritmética de una muestra cuyos datos fueran los logaritmos de los valores de la variable aleatoria con su misma frecuencia relativa.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de centralización

Media cuadrática: corresponde a la raíz cuadrada de la media de los cuadrados de los valores que constituyen la medida.

$$\bar{x}_q = \left[\frac{1}{N}\left({x_1}^2 + {x_2}^2 + \cdots + {x_N}^2\right)\right]^{1/2}$$

En el caso de *N* resultados agrupados en *k* clases, cada una con frecuencias relativas f_i, la media cuadrática es:

$$\bar{x}_q = \left[\sum_{i=1}^{k} f_i \, {x_i}^2\right]^{1/2}$$

Media armónica: corresponde al inverso de la media de los inversos de los valores que constituyen la medida.

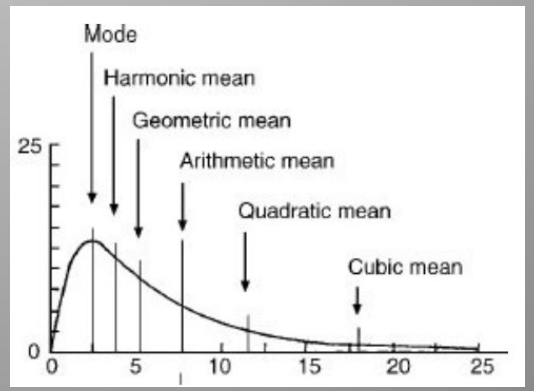
$$\frac{1}{\bar{x}_a} = \frac{1}{N}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}\right)$$

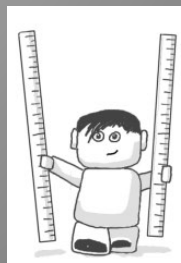
En el caso de *N* resultados agrupados en *k* clases, cada una con frecuencias relativas f_i, la media armónica es:

$$\frac{1}{\bar{x}_a} = \sum_{i=1}^{k} f_i \, \frac{1}{x_i}$$

Se puede demostrar que:

$$\bar{x}_a \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_q$$





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

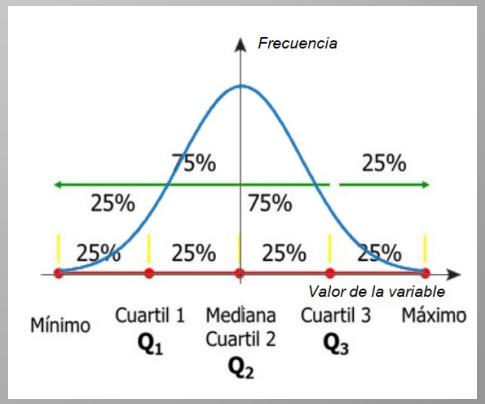
Medidas características de centralización

Moda (Md): corresponde al valor más frecuente en la distribución. Una muestra puede tener un único valor correspondiente a la moda o ser multimodal, esto es, tener una tabla de frecuencias con varios máximos de igual altura. No es una medida que aporte mucha información sobre la distribución, pues se obtiene observando el valor más frecuente, con independencia del resto de valores.

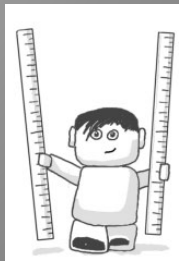
$$Md = \{x_i \mid f_i = \max(f_1, f_2, \dots, f_k)\}$$

Percentil q-ésimo (Pq): corresponde al valor de la variable aleatoria tal que la frecuencia relativa de todos los valores menores o iguales sea *q (0 ≤ q ≤ 1).*
Los percentiles permiten describir la distribución de los datos muestrales de forma detallada al mostrar los valores de la variable en los que se alcanza un determinado valor de la frecuencia acumulada.

$$q = \sum_i f_i \quad \mid x_i \leq P_q$$



Algunos percentiles tienen nombre propio:
- **Mediana** (q = 0,5) o **Q_2**.
- **Cuartiles primero o Q_1** (q = 0,25) y **tercero o Q_3** (q = 0,75).
- **Deciles primero o D_1** (q = 0,1), **segundo o D_2** (q = 0,2), …



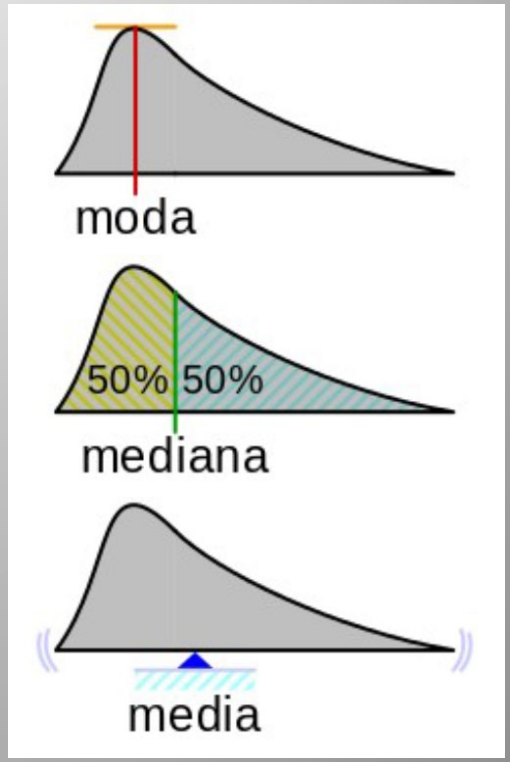
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

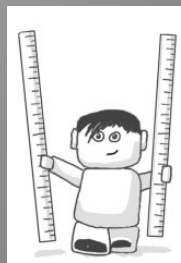
Medidas características de centralización

Mediana (Me): corresponde al valor de la variable para el que los datos de valor inferior son tan frecuentes como los de valor superior.
Corresponde con el percentil q = 0.5, tal y como hemos definido.

La mediana representa una estimación más robusta que la media aritmética para el valor central de la distribución, ya que la media es más sensible a posibles datos atípicos.

En una variable aleatoria discreta (por ejemplo) puede no existir ningún valor de la variable que coincida con el percentil *Pq* buscado. Será, por tanto, necesario establecer alguna regla de interpolación que permita establecer estos percentiles.





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de centralización

La obtención de los percentiles se realiza ordenando los valores de la variable en orden creciente y construyendo su tabla de frecuencias absolutas acumuladas:

$$x_1 < x_2 < \cdots < x_k \qquad \{N_1, N_2, \ldots, N_{k-1}, N\};$$

A continuacion se busca el valor la variable aleatoria cuya frecuencia acumulada es más próxima a qN con q el percentil buscado y N el número de medidas.

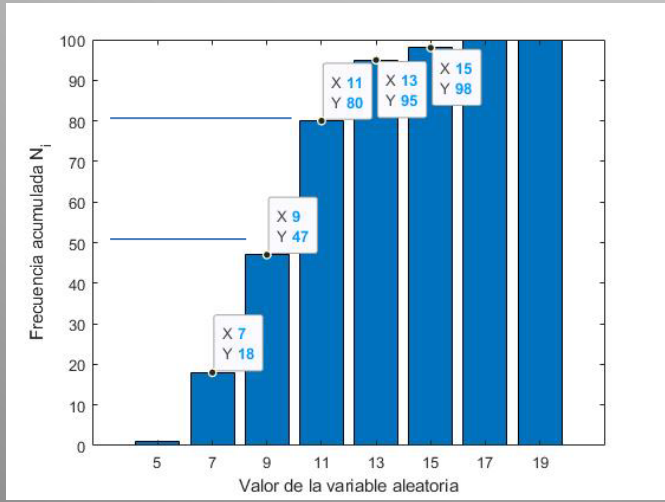
En el caso de variable discreta:

1) Si qN es entero, y coincide con una de las frecuencias absolutas acumuladas N_i para el valor x_i, entonces el valor del percentil será

$$qN = N_i\,; \qquad P_q = \frac{x_i + x_{i+1}}{2}$$

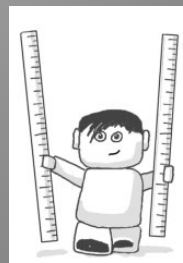
2) En caso de que no sea entero o no corresponda a ninguna frecuencia N_i, entonces estará comprendido entre dos valores x_i y x_{i+1} de la variable, con lo que se tomará el mayor de los dos

$$N_i < qN < N_{i+1}\,; \qquad P_q = x_{i+1}$$



$$P_{1/2} = 11$$

$$P_{0.8} = 12$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de centralización

x_i	n_i	f_i	N_i	F_i
0	2	0,0067	2	0,0067
1	7	0,0233	9	0,0300
2	15	0,0500	24	0,0800
3	25	0,0833	49	0,1633
4	38	0,1267	87	0,2900
5	52	0,1733	139	0,4633
6	52	0,1733	191	0,6367
7	40	0,1333	231	0,7700
8	30	0,1000	261	0,8700
9	19	0,0633	280	0,9333
10	10	0,0333	290	0,9667
11	6	0,0200	296	0,9867
12	3	0,0100	299	0,9967
13	1	0,0033	300	1,0000

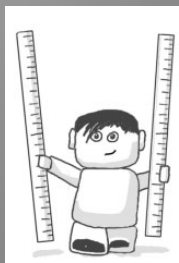
Ejemplo del Contador Geiger

En este caso $N = 300$
Los cuartiles primero y tercero $\boldsymbol{Q_1}$ (q=1/4) y $\boldsymbol{Q_3}$ (q=3/4) se obtendrán considerando los valores de qN que son 75 y 225, respectivamente. Entonces:

$\boldsymbol{Q_1}$ = 4 (comprende hasta N_i= 87)
$\boldsymbol{Q_3}$ = 7 (comprende hasta N_i = 231)

El percentil q = 14/15 corresponde a qN = 280, y

$$P_{14/15} = \frac{9 + 10}{2} = 9.5$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de centralización

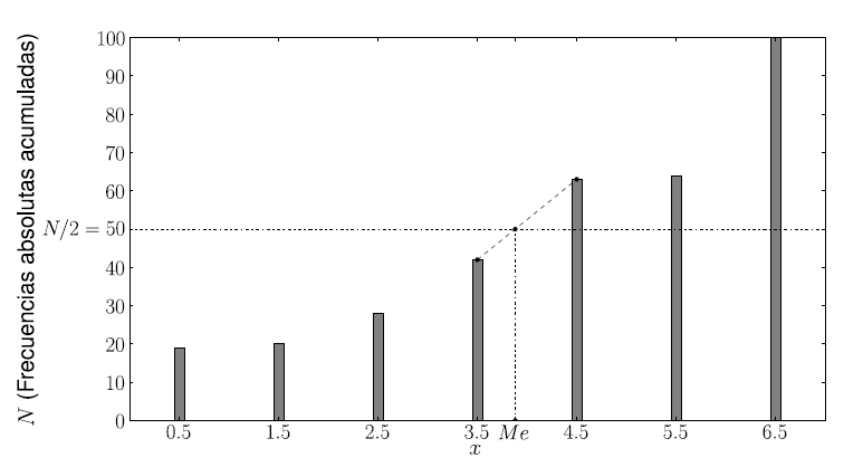
En el caso de variable continua:

Consideraremos la variable agrupada en clases $[a_0, a_1) \dots [a_{i-1}, a_i) \dots$ con marcas de clase x_i y frecuencia absoluta acumulada $N_i \;\; i = 1,2, \dots, k$.

1) Si $qN = N_i \;\; \exists i \in \{1,2, \dots, k\}$, entonces el valor del percentil P_q será el extremo superior del intervalo de la clase correspondiente a_i.

2) En otro caso, $N_i < qN < N_{i+1} \;\; \exists i \in \{1,2, \dots, k\}$ se interpola entre los dos valores de las marcas de clase x_i y x_{i+1} (*)

$$P_q = x_i + \frac{qN - N_i}{N_{i+1} - N_i} (x_{i+1} - x_i)$$

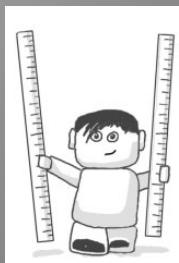


En este caso: $qN = 50$; N_i = 42; N_{i+1} = 63; x_i = 3.5; x_{i+1} =4.5; con lo que $P_{0.5} = 3.88$

Por otro lado $P_{0.63} = 5.0$

(*) Si queremos establecer una definición de P_q sin discontinuidades, entonces alternativamente

$$P_q = a_i + \frac{qN - N_i}{N_{i+1} - N_i} (a_{i+1} - a_i)$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de dispersión

Varianza muestral: corresponde al promedio del cuadrado de las desviaciones con respecto a la media muestral.

En el caso de N medidas con resultados $\{ x_1, x_2, \dots , x_{N-1}, x_N \}$ o agrupando los N resultados en k clases, cada una con frecuencias relativas $f_i \quad i = 1,2, \dots, k$ entonces (*):

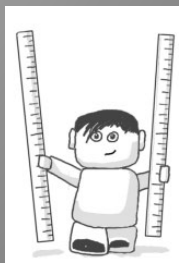
$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \quad (**)$$

$$s^2 = \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2$$

La varianza es una cantidad definida positiva. Se puede demostrar de forma sencilla que:

$$s^2 = \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i \, ({x_i}^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{k} f_i \, {x_i}^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Lo que es un importante resultado para el cálculo de la varianza muestral, que se utilizara en múltiples ocasiones (lo veréis incluso como definición en otras asignaturas).



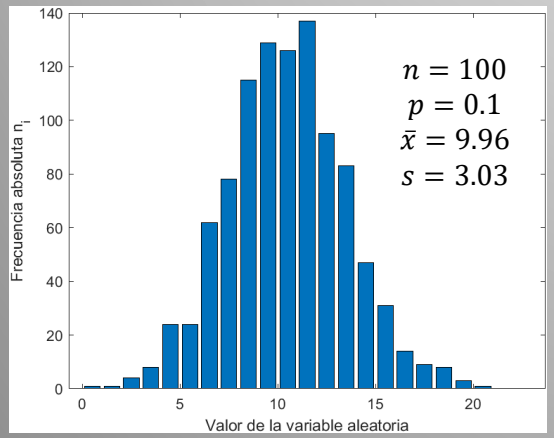
(*) N.B. Abuso de notación. El agrupamiento en clases puede modificar significativamente el valor del estadístico media y varianza muestral.
(**) Veremos más adelante la necesidad de utilizar N-1 en el denominador en lugar de N para obtener un estimador fiel de la varianza de la población.

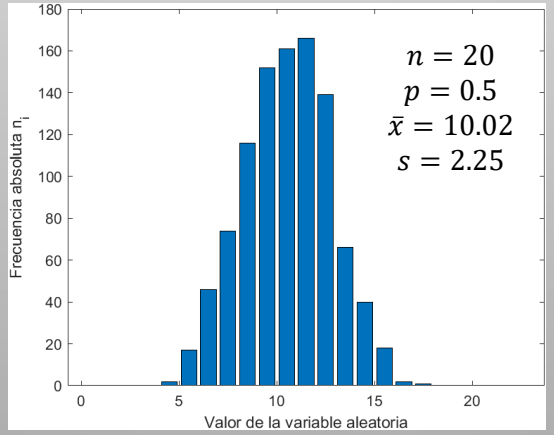
Medidas características de dispersión

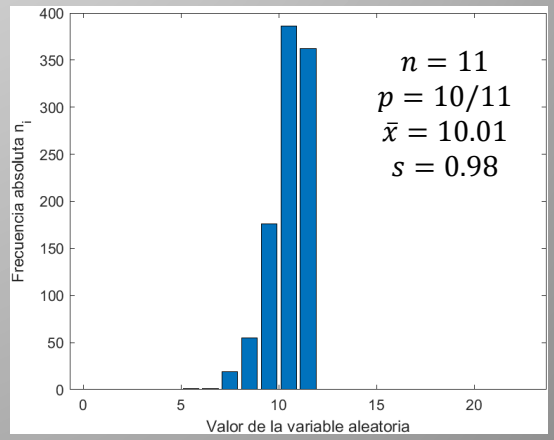
Desviación típica (o estándar) muestral: corresponde a la raíz cuadrada positiva de la varianza muestral.

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad s = \sqrt{\sum_{i=1}^{k} f_i\,(x_i - \bar{x})^2}$$

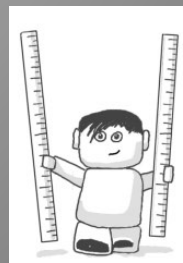
La desviación típica tiene la misma dimensión que la variable aleatoria. Es una medida del grado de dispersión de los datos respecto a la media aritmética.







Distribuciones de 1000 datos de variables aleatorias binomiales con medias iguales y distinta varianza.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de dispersión

Desigualdad de Tchebychev (Chebyshev)(Чебышёв): Establece que la frecuencia de los datos que distan más de α $(\alpha > 0)$ veces la desviación típica de la media de la distribución es inferior a $^1/_{\alpha^2}$

$$f(x_i \mid |x_i - \bar{x}| > \alpha\, s) < \frac{1}{\alpha^2}$$

Consideremos los dos conjuntos disjuntos de datos:

$$A_1 = \{x_i \mid |x_i - \bar{x}| > \alpha\, s\} \qquad A_2 = \{x_i \mid |x_i - \bar{x}| \leq \alpha\, s\}$$

Podemos escribir la desigualdad como: $f(A_1) < \frac{1}{\alpha^2}$

$$s^2 = \sum_{x_i \epsilon A_1} f_i\,(x_i - \bar{x})^2 + \sum_{x_i \epsilon A_2} f_i\,(x_i - \bar{x})^2 \geq \sum_{x_i \epsilon A_1} f_i\,(x_i - \bar{x})^2 > \sum_{x_i \epsilon A_1} f_i\,\alpha^2 s^2 = \alpha^2 s^2 f(A_1)$$

Es importante darse cuenta de que esta desigualdad es aplicable a cualquier distribución de datos de varianza finita y es válida **aunque desconozcamos la distribución de los datos**.
A partir de la desigualdad, podemos acotar la frecuencia de los datos en los siguientes intervalos:

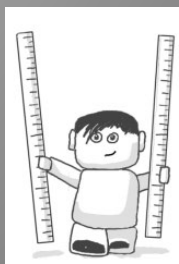
$$f(\bar{x} - 2s, \bar{x} + 2s) > 0.75$$

$$f(\bar{x} - 3s, \bar{x} + 3s) > 0.889$$

$$f(\bar{x} - 4s, \bar{x} + 4s) > 0.938$$

$$f(\bar{x} - ks, \bar{x} + ks) > 1 - 1/k^2$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



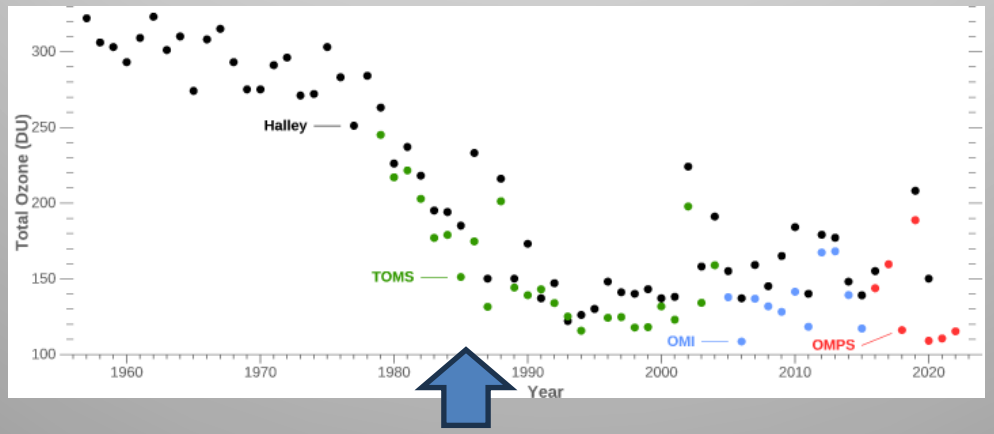
Medidas características de dispersión

Se denomina rango intercuartilico a la distancia comprendida entre el primer y el tercer cuartil de la variable aleatoria esto es, al valor $R_I = P_{0.75} - P_{0.25}$

Se considera un **dato atipico leve** el que que se encuentra a más de $1.5\ R_I$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$ y se considera un **dato atipico extremo** el que aparece a mas de $3\ R_I$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$.

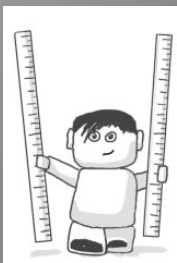
!Los datos atipicos no deber rechazarse! Es importante revisar lo que ocurre con estos datos, evaluar sus incertidumbres, posibles errores en el proceso de medida… Pero no se puede eliminar un dato o un conjunto de datos por separarse de la media o de los valores esperados de acuerdo a un modelo.

Joseph Farman
Geofísico Británico



Farman, J., Gardiner, B. & Shanklin, J. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interaction. *Nature* **315**, 207–210 (1985). https://doi.org/10.1038/315207a0

Por qué diez años de datos que indicaban el agujero en la capa de ozono de la Antártida no fueron considerados antes por los científicos.

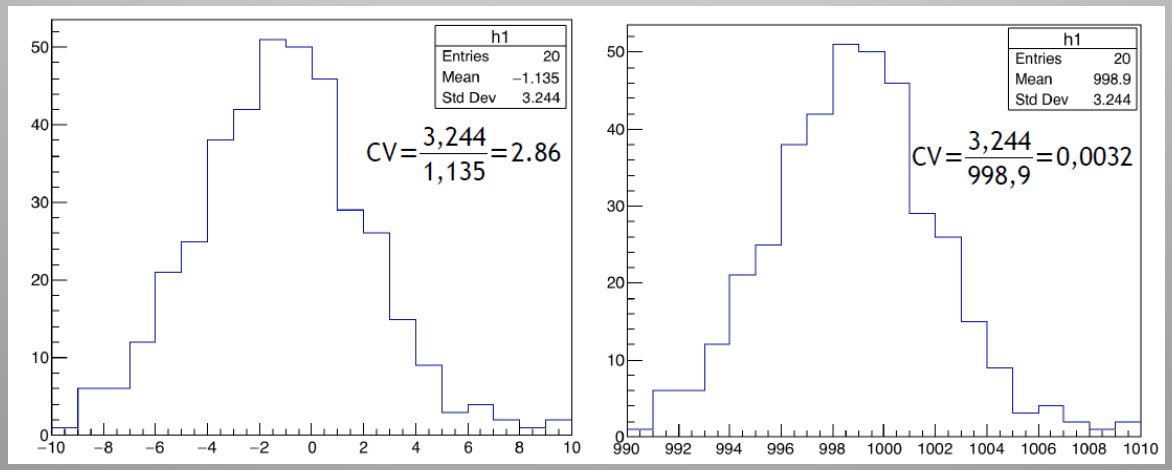


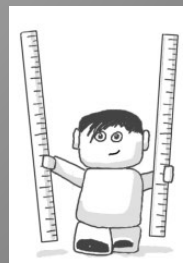
Medidas características de dispersión

Coeficiente de variacion de Pearson (CV). Corresponde a una medida relativa de la desviación de la distribución frente a su valor medio, esto es, de la variabilidad relativa de los datos de la muestra. Se define como:

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

Al contrario que la varianza o la desviación típica, el coeficiente de variación de Pearson (adimensional) evalúa la desviación de la muestra pesada por su valor central, es un índice de la desviación relativa de la muestra (i.e. relación ruido/señal) (N.B. no es aplicable a distribuciones con media nula).





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de dispersión

Otras definiciones de interés como medidas características de la dispersión de una distribución de frecuencias de una variable aleatoria son:

Desviación media con respecto a la media:

$$DM_{\bar{x}} = \sum_{i=1}^{k} f_i \, |x_i - \bar{x}|$$

Desviación media con respecto a la mediana:

$$DM_{Me} = \sum_{i=1}^{k} f_i \, |x_i - Me|$$

Coeficiente de variación media (respecto a la media):

$$CV_{M\bar{x}} = \frac{DM_{\bar{x}}}{|\bar{x}|}$$

Coeficiente de variación media (respecto a la mediana):

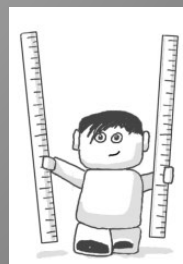
$$CV_{Me} = \frac{DM_{Me}}{|\bar{x}|}$$

Recorrido de la variable:

$$R = \max\{x_i\} - \min\{x_i\}$$

Recorrido semi-intercuartílico:

$$R_{SI} = \frac{1}{2} R_I = \frac{P_{3/4} - P_{1/4}}{2}$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características muestrales

Definimos como momento de orden *j* de una distribucion de frecuencias f_i con respecto al punto (o valor) *c* como:

$$m_j(c) = \sum_{i=1}^{k} f_i \, (x_i - c)^j$$

Si el punto $c = 0$ se denominan momentos respecto al origen, $m_j(0)$, y cuando $c = \bar{x}$ **momentos centrales** o momentos respecto a la media $m_j(\bar{x})$.

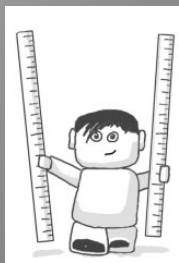
Muchas de las medidas características muestrales se pueden obtener utilizando los momentos de distinto orden de la distribución de frecuencias. Por ejemplo:

$$m_1(0) = \sum_{i=1}^{k} f_i \, (x_i)^1 = \bar{x}$$

$$m_1(\bar{x}) = \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^1 = 0$$

$$m_2(\bar{x}) = \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2 = s^2$$

$$m_2(\bar{x}) = m_2(0) - (m_1(0))^2$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

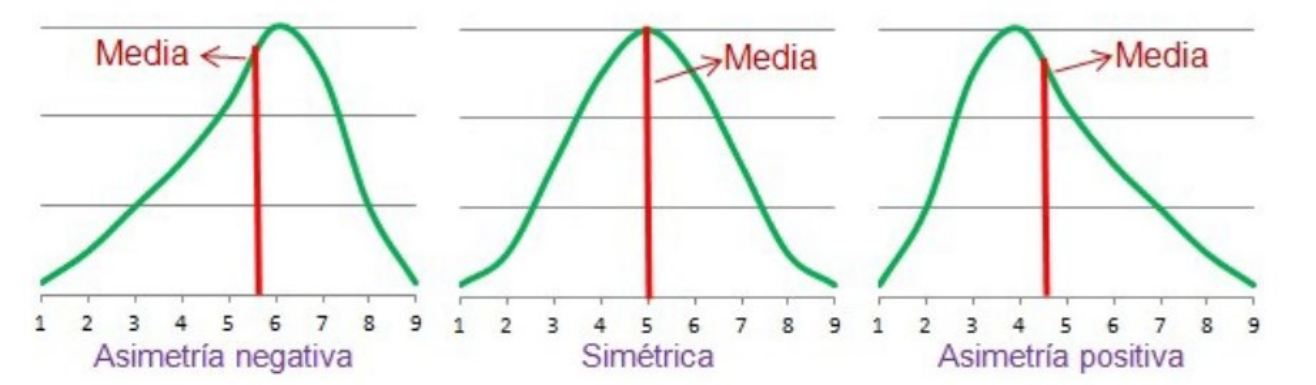
Medidas características de asimetría

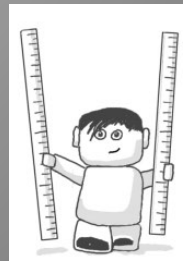
Coeficiente de asimetría de Fisher:
Definido como el cociente del momento central de orden 3 de la distribución de datos sobre la desviación típica al cubo (adimensional).

$$A_F = \frac{1}{s^3}\sum_{i=1}^{k} f_i\,(x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{s^3}$$

Cuando $A_F > 0$ la distribución será asimétrica positiva (derecha), con una mayor importancia en el sumatorio de los valores alejados mayores que la media. En el caso de $A_F < 0$ los valores menores y alejados de la media contribuyen más (izquierda) y se denomina asimétrica negativa.

Si la distribución es simétrica, entonces sabemos que $A_F = 0$. El recíproco no es cierto, es decir, aunque $A_F = 0$ la distribución puede ser simétrica o no. En ocasiones se denota al coeficiente de asimetría de Fisher como γ_1.





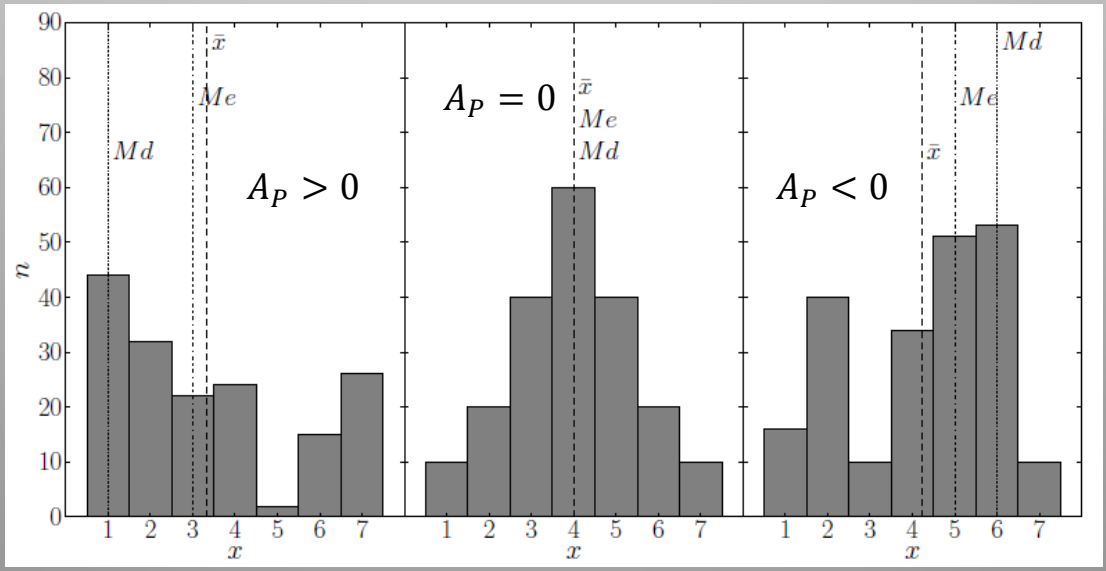
Medidas características de asimetría

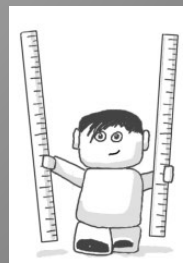
Coeficiente de asimetría de Pearson:
Definido como la media menos la moda dividido por la desviación típica (adimensional).

$$A_P = \frac{\bar{x} - Md}{s}$$

En este caso si $A_P > 0$ la distribución será asimétrica positiva (izquierda en la figura inferior), con una moda menor que la media. En el caso de $A_P < 0$ (derecha en la figura inferior) la moda es mayor que la media y la asimetría se denomina negativa.

No está definido en caso de distribuciones multimodales.





Medidas características de asimetría

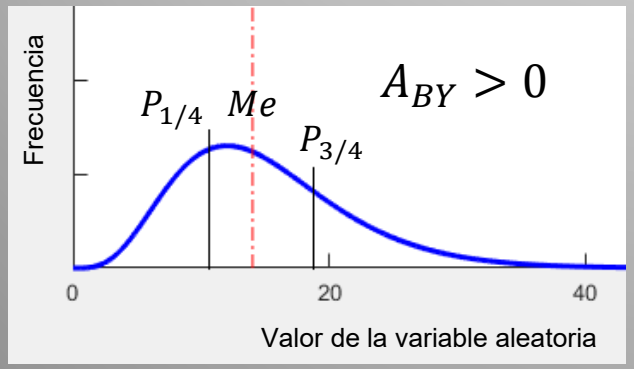
Coeficiente de asimetría de Bowley-Yule:

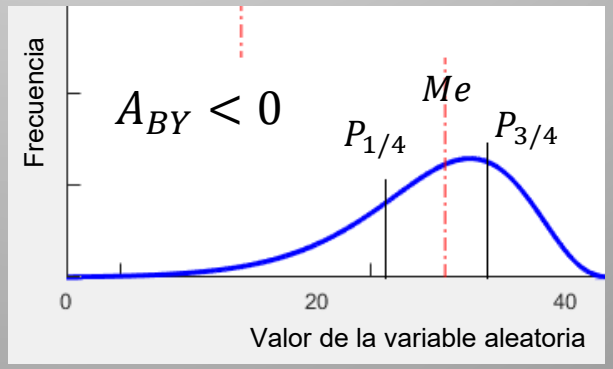
Definido como la diferencia de las distancias de la mediana (segundo cuartil) al primer y tercer cuartil dividido por la distancia intercuartílica.

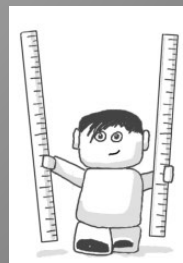
$$A_{BY} = \frac{(P_{3/4} - Me) - (Me - P_{1/4})}{P_{3/4} - P_{1/4}} = \frac{P_{3/4} + P_{1/4} - 2\,Me}{P_{3/4} - P_{1/4}}$$

La asimetría es positiva cuando $(P_{3/4} - Me) > (Me - P_{1/4})$

mientras que es negativa cuando $(P_{3/4} - Me) < (Me - P_{1/4})$







USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Medidas características de curtosis

Coeficiente de apuntamiento o curtosis:

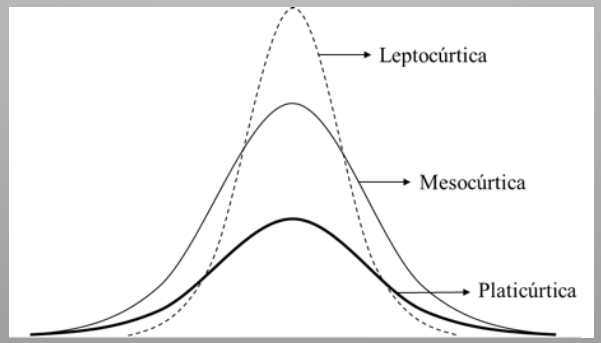
Definido como el cociente del momento de cuarto orden respecto a la media dividido por el cuadrado de la varianza.

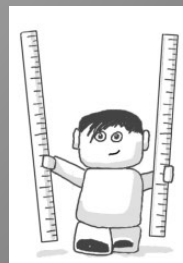
$$g = \frac{m_4(\bar{x})}{s^4} = \frac{1}{s^4} \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^4$$

También denominado a veces β_2 nos indica la forma de la distribución de datos en torno a la media o grado de apuntamiento. Suele tomarse como referencia el coeficiente de apuntamiento de la distribución normal o gaussiana $g = 3$ y por tanto también se define alternativamente el coeficiente g_2:

$$g_2 = g - 3$$

- Distribuciones leptocúrticas (alto apuntamiento) $g > 3$ ó $g_2 > 0$
- Distribuciones normo o mesocúrticas (apuntamiento gaussiano) $g = 3$ ó $g_2 = 0$
- Distribuciones platicúrticas (bajo apuntamiento) $g < 3$ ó $g_2 < 0$





Transformaciones de la variable aleatoria

Cuando realizamos una medida en muchos casos debemos evaluar el valor de una magnitude indirecta mediante la transformación functional de los valores obtenidos directamente. En general, nos planteamos qué le ocurre a la variable aleatoria X cuando la transformamos en una nueva variable Y tal como

$$Y = h(X)$$

$Y = a + b\,X$ $\quad$ $Y = X^p$ $\quad$ $Y = \log(X)$ $\quad$ $Y = \exp(X)$ $\quad$ ……

Consideremos un caso particularmente simple con sólo dos valores de la variable aleatoria

$x_1 = 0;\quad f_1 = 0.5$ $\quad$ $x_2 = 2;\quad f_2 = 0.5$ $\quad$ $\bar{x} = 1;\quad s_x = 1$

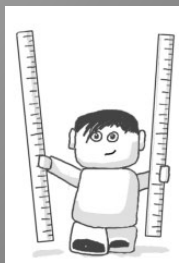
Tomando simplemente una nueva variable como el cuadrado de la primera $Y = X^2$

$y_1 = 0;\quad f_1 = 0.5$ $\quad$ $y_2 = 4;\quad f_2 = 0.5$ $\quad$ $\bar{y} = 2;\quad s_y = 2$

Obviamente

$$\bar{y} \neq (\bar{x})^2;\quad s_y \neq (s_x)^2$$

En general la transformación de la variable aleatoria $Y = h(X)$ genera una nueva variable con una distribución diferente de la original y las medias y desviaciones típicas no se transforman de acuerdo a la misma función h.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Transformaciones de la variable aleatoria

Consideremos una transformación lineal de la variable aleatoria

$$Y = a + b\,X$$

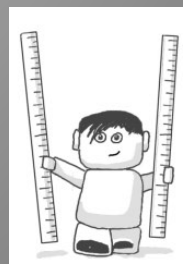
Partimos de una variable X con valores $\{x_1, x_2, \dots, x_k\}$ y frecuencias relativas $\{f_1, f_2, \dots, f_k\}$, obteniendo la variable Y con valores $\{a + b\,x_1, a + b\,x_2, \dots, a + b\,x_k\}$ y frecuencias relativas $\{f_1, f_2, \dots, f_k\}$

$$\bar{y} = \sum_{i=1}^{k} f_i\,(a + b\,x_i) = a + b\,\bar{x}$$

$$s_y{}^2 = \sum_{i=1}^{k} f_i\,(y_i - \bar{y})^2 = \sum_{i=1}^{k} f_i\,(a + bx_i - a - b\,\bar{x})^2 = b^2 \sum_{i=1}^{k} f_i\,(x_i - \bar{x})^2 = b^2\ s_x{}^2$$

$$s_y = |b|\,s_x$$

Observemos que en el caso de una traslación $b = 0$ la varianza de la nueva variable no se modifica y solo se transforma la media. En el caso de un cambio de escala sin traslación $a = 0$ se modifican tanto la media como la varianza.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Transformaciones de la variable aleatoria

Si consideramos una transformación monótona de la variable $Y = h(X)$ (i.e. creciente), entonces tendremos que

$$x_1 < x_2 < \cdots < x_k$$

$$y_i = h(x_i) \quad \Downarrow$$

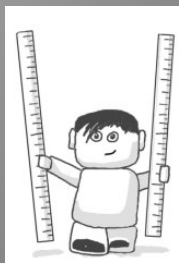
$$y_1 < y_2 < \cdots < y_k$$

Podremos afirmar que los percentiles de la nueva variable se transforman de acuerdo a la ley de transformación de la variable, es decir:

$$P_q(y) = h\left[P_q(x)\right]$$

En una transformación lineal se cumple la condición de monotonía y por tanto:

$$y_i = a + b\, x_i \qquad P_q(y) = a + b\, P_q(x)$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Transformaciones de la variable aleatoria

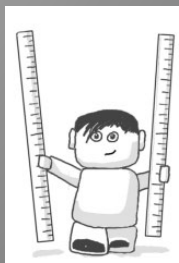
Si consideramos una transformación funcional de la variable $Y = h(X)$ podremos estudiar algunas aproximaciones respecto a la transformación de la media y varianza. Desarrollando esta función en serie de Taylor entorno a la media de X

$$h(z) = h(\bar{x}) + h'(\bar{x})(z - \bar{x}) + \frac{1}{2}h''(\bar{x})\,(z - \bar{x})^2 + \cdots$$

$$\bar{y} = \sum_{i=1}^{k} f_i\, h(x_i) = h(\bar{x}) \sum_{i=1}^{k} f_i + h'(\bar{x}) \sum_{i=1}^{k} f_i\,(x_i - \bar{x}) + \frac{1}{2}h''(\bar{x}) \sum_{i=1}^{k} f_i\,(x_i - \bar{x})^2 + \cdots$$

$$\bar{y} = h(\bar{x}) + \frac{1}{2}h''(\bar{x})\, {s_x}^2 + \cdots$$

Si los términos $\frac{1}{2}h''(\bar{x})\, {s_x}^2 + \cdots$ son despreciables entonces se puede considerar $\bar{y} \approx h(\bar{x})$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Transformaciones de la variable aleatoria

En el caso de la varianza de la nueva variable transformada teniendo en cuenta que:

$$h(z) = h(\bar{x}) + h'(\bar{x})(z - \bar{x}) + \frac{1}{2}h''(\bar{x})\,(z - \bar{x})^2 + \cdots$$

$$\bar{y} = h(\bar{x}) + \frac{1}{2}h''(\bar{x})\,{s_x}^2 + \cdots$$

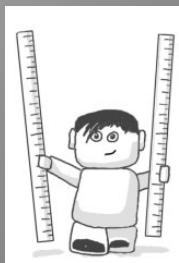
$${s_y}^2 = \sum_{i=1}^{k} f_i\,(y_i - \bar{y})^2 = \sum_{i=1}^{k} f_i \left(h(\bar{x}) + h'(\bar{x})(x_i - \bar{x}) + \frac{1}{2}h''(\bar{x})\,(x_i - \bar{x})^2 + \cdots - h(\bar{x}) - \frac{1}{2}h''(\bar{x})\,{s_x}^2 - \cdots \right)^2$$

$${s_y}^2 = \sum_{i=1}^{k} f_i \left(h'(\bar{x})(x_i - \bar{x}) + \frac{1}{2}h''(\bar{x})\,[(x_i - \bar{x})^2 - {s_x}^2] + \cdots \right)^2$$

$${s_y}^2 = \left(h'(\bar{x})\right)^2 \sum_{i=1}^{k} f_i\,(x_i - \bar{x})^2 + h'(\bar{x})\,h''(\bar{x}) \sum_{i=1}^{k} f_i\,(x_i - \bar{x})^3 + \cdots$$

Si los consideramos que los términos de orden superior son despreciables, entonces a primer orden obtenemos

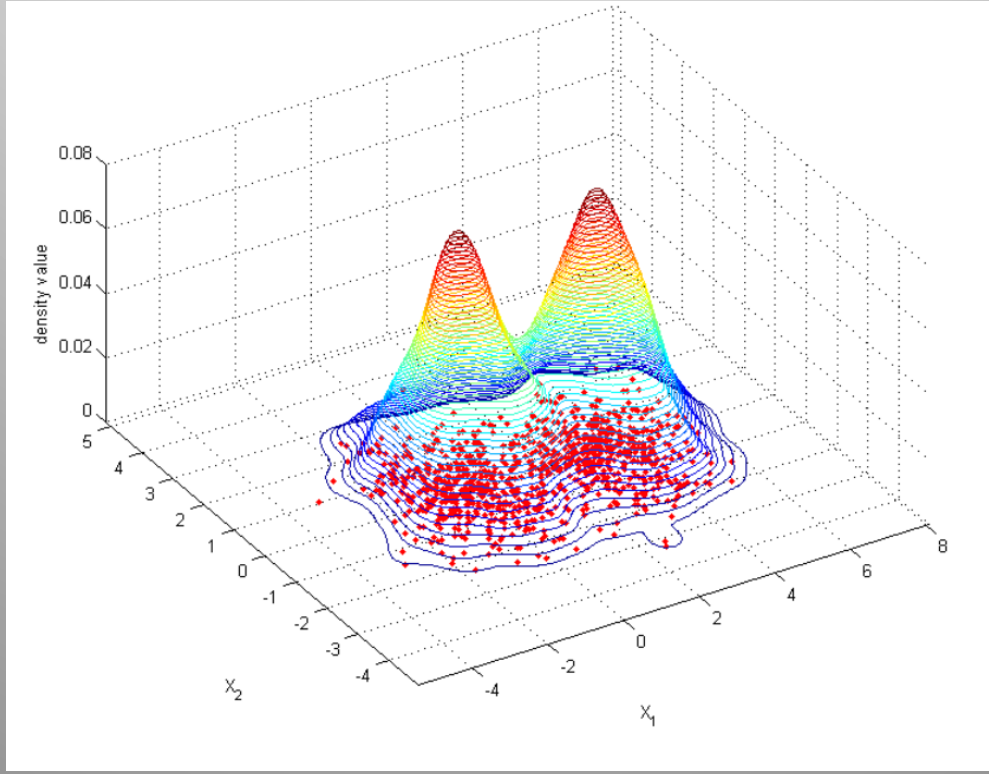
$${s_y}^2 = \left(h'(\bar{x})\right)^2 {s_x}^2 + \cdots$$

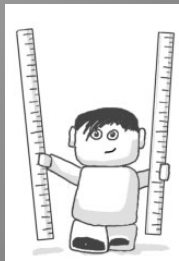


USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones de frecuencias multivariantes

Al estudiar **varias variables aleatorias** obtenidas simultáneamente en un experimento surgen **nuevas e importantes propiedades que muestran las relaciones entre estas variables**. Para poder caracterizarlas introduciremos las muestras y distribuciones de frecuencia multivariantes, ejemplificando con pares de dos variables, pero entendiendo que estos resultados se pueden extender a cualquier número de ellas.





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones de frecuencias multivariantes

Si consideramos dos variables estadisticas X e Y con valores posibles $\{x_1, x_2, \dots, x_k\}$ y $\{y_1, y_2, \dots, y_l\}$,cada muestra del experimento aleatorio multivariante proporcionara un par (x_i, y_j). Se denomina **frecuencia absoluta**, n_{ij}, al **número de veces** que aparece el par (x_i, y_j) en el experimento y se denomina **frecuencia relativa**, f_{ij}, a **la fracción de veces** que se observa el par (x_i, y_j) en el total de las medidas realizadas. Podremos construir una tabla de distribución de las frecuencias absolutas observadas:

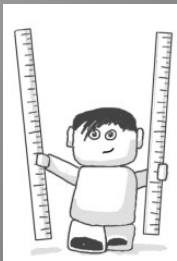
X/Y	y_1	y_2	...	y_j	...	y_l
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}
...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}
...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}		n_{kj}	...	n_{kl}

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N$$

$$f_{ij} = \frac{n_{ij}}{N}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1$$

El conjunto de las frecuencias (absolutas o relativas) observadas en un experimento aleatorio constituyen su **distribucion de frecuencias multivariante** (absolutas o relativas).



Distribuciones marginales de frecuencias multivariantes

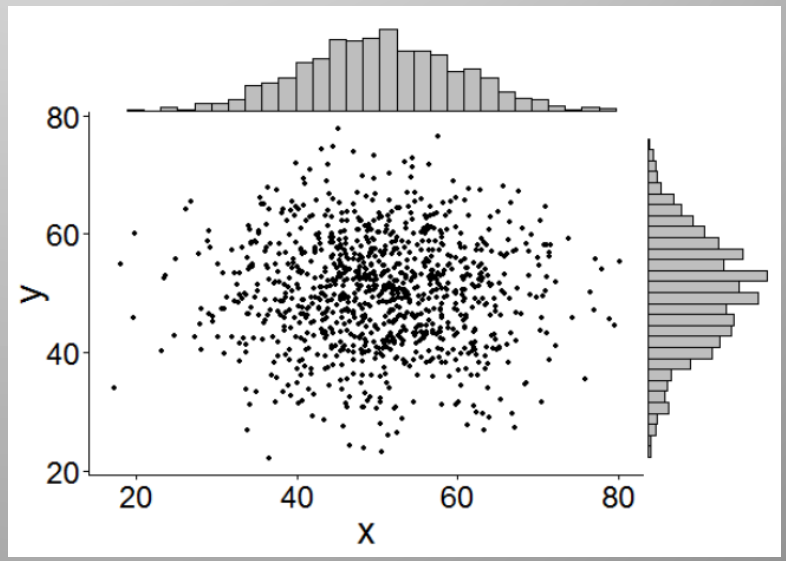
La **distribución de frecuencia marginal** para una variable aleatoria $\boldsymbol{X}$ dentro de una distribución multivariante se corresponde con la distribución de frecuencias asociada a los posibles valores de la variable $\boldsymbol{X}$ $(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_k\})$ independientemente del valor que tomen el resto de variables aleatorias.

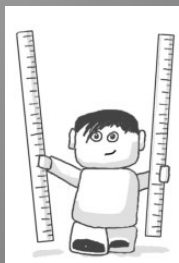
Asi, la **distribucion de frecuencia absoluta marginal** para la variable $\boldsymbol{X}$ de la distribucion bivariante $\boldsymbol{X}, \boldsymbol{Y}$, con valores posibles $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_k\}$ e $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_l\}$, será la serie de valores n_{x_i} definidos como:

$$\sum_{j=1}^{l} n_{ij} = n_{x_i} \qquad i = 1,2, \dots, k$$

Análogamente para la variable $\boldsymbol{Y}$

$$\sum_{i=1}^{k} n_{ij} = n_{y_j} \qquad j = 1,2, \dots, l$$





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones marginales de frecuencias multivariantes

Las **distribuciones de frecuencias relativas marginales** se definen como las series de valores $\{f_{x_1}, f_{x_2}, \dots, f_{x_k}\}$ y $\{f_{y_1}, f_{y_2}, \dots, f_{y_l}\}$ que cumplen:

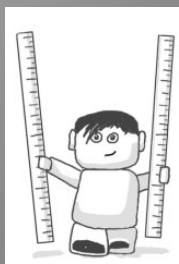
$$f_{x_i} = \frac{1}{N}\sum_{j=1}^{l} n_{ij} = \frac{n_{x_i}}{N} \qquad i = 1,2,\dots,k \qquad f_{x_i} = \sum_{j=1}^{l} f_{ij}$$

$$f_{y_j} = \frac{1}{N}\sum_{i=1}^{k} n_{ij} = \frac{n_{y_j}}{N} \qquad j = 1,2,\dots,l \qquad f_{y_j} = \sum_{i=1}^{k} f_{ij}$$

Y verifican las propiedades:

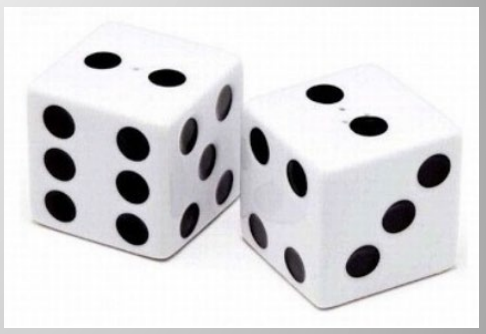
$$N = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{l} n_{y_j}$$

$$1 = \sum_{j=1}^{l}\sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} f_{x_i} = \sum_{j=1}^{l} f_{y_j}$$



Distribuciones multivariantes

Ejemplo de distribución multivariante (lanzamiento de dos dados):

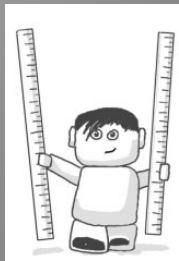


D1\D2	1	2	3	4	5	6	
1	25	19	33	32	35	27	$n_{D1=1}$ =171
2	22	30	30	19	27	22	$n_{D1=2}$ =150
3	30	27	23	37	23	31	$n_{D1=3}$ =171
4	22	24	39	27	24	49	$n_{D1=4}$ =185
5	32	28	25	15	14	30	$n_{D1=5}$ =144
6	42	43	12	34	28	20	$n_{D1=6}$ =179

$n_{D2=1}$ =173 $n_{D2=3}$ = 162 $n_{D2=5}$ = 151

$n_{D2=2}$ =171 $n_{D2=4}$ =164 $n_{D2=6}$ = 179

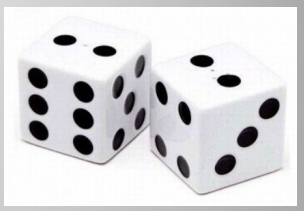
$$N = 1000$$

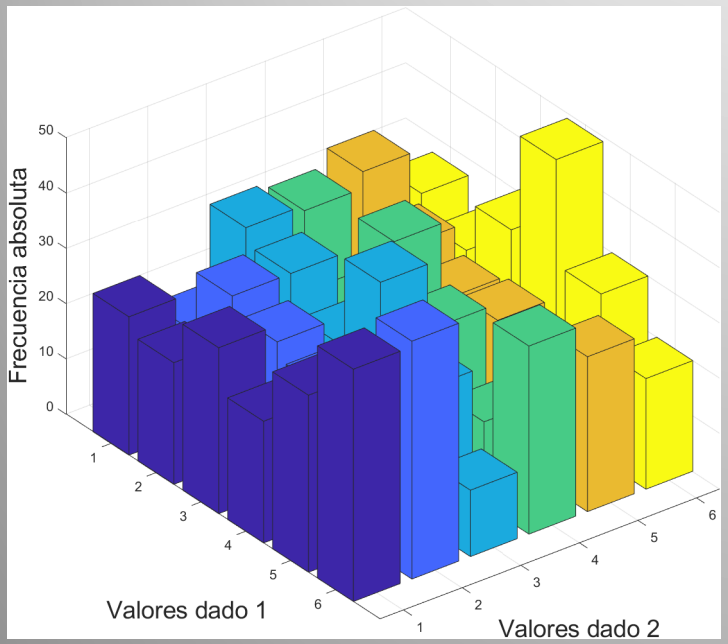


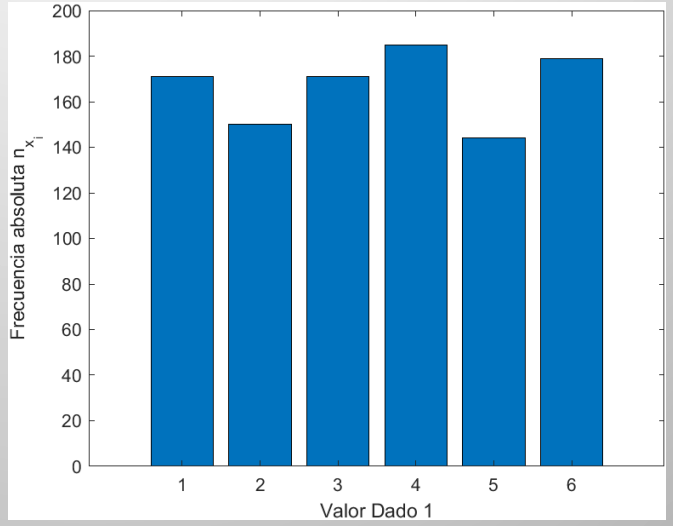
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

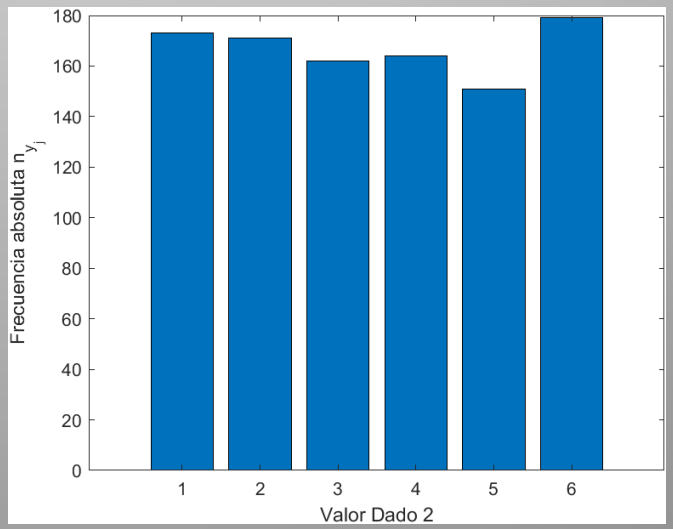
Distribuciones multivariantes

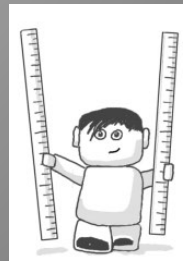
Distribuciones de frecuencia marginales del ejemplo de dos dados.











USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: frecuencia condicionada

Para una distribución multivariante, se definen los valores **de frecuencia absoluta condicionada a $Y = y_j$ para el valor x_i de la variable X** como el número de resultados obtenidos x_i en la variable X condicionadas a que aparezca el valor y_j en la variable Y,

$$n(x_i \mid Y = y_j) = n(x_i \mid y_j) = n_{ij}$$

Análogamente para Y

$$n(y_j \mid X = x_i) = n(y_j \mid x_i) = n_{ij}$$

La **frecuencia relativa condicionada a $Y = y_j$ para el valor x_i de la variable X** , se define como la frecuencia absoluta condicionada dividida por la frecuencia marginal de la condición n_{y_j}

$$f(x_i \mid y_j) = \frac{n_{ij}}{n_{y_j}} = \frac{f_{ij}}{f_{y_j}}$$

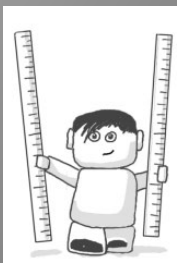
corresponden a las frecuencias relativas de aparicion de $X = xi$ **restringidas únicamente al subconjunto de datos donde $Y = y_j$**, y no respecto al total de las observaciones.

Análogamente para Y

$$f(y_j \mid x_i) = \frac{n_{ij}}{n_{x_i}} = \frac{f_{ij}}{f_{x_i}}$$

corresponden a las frecuencias relativas de aparicion de $Y = y_j$ **restringidas únicamente al subconjunto de datos donde $X = xi$**, y no respecto al total de las observaciones.
En general, por tanto:

$$f(x_i \mid y_j) \neq f(y_j \mid x_i)$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: frecuencia condicionada

frecuencia absoluta condicionada

$$n(x_i \mid Y = y_j) = n(x_i \mid y_j) = n_{ij}$$

$$n(y_j \mid X = x_i) = n(y_j \mid x_i) = n_{ij}$$

Frecuencia relativa condicionada

$$f(x_i \mid y_j) = \frac{n_{ij}}{n_{y_j}} = \frac{f_{ij}}{f_{y_j}}$$

$$f(y_j \mid x_i) = \frac{n_{ij}}{n_{x_i}} = \frac{f_{ij}}{f_{x_i}}$$

X/Y	y_1	y_2	...	y_j	...	y_l
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}
x_2	n_{21}	n_{22}	...	n_{2j}	..	n_{2l}
...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	..	n_{il}
...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}		n_{kj}	...	n_{kl}

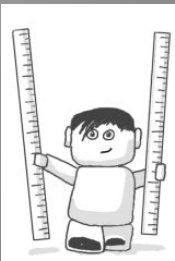
Frecuencia relativa marginal para x_i

$$f_{x_i} = \frac{1}{N}\sum_{j=1}^{l} n_{ij} = \frac{n_{x_i}}{N}$$

$$f(x_i \mid y_j) \neq f(y_j \mid x_i)$$

Frecuencia relativa marginal para y_j

$$f_{y_j} = \frac{1}{N}\sum_{i=1}^{k} n_{ij} = \frac{n_{y_j}}{N}$$

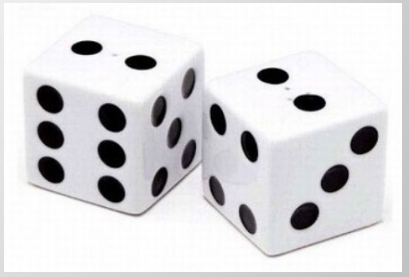


Distribuciones multivariantes: frecuencia condicionada

D1\D2	1	2	3	4	5	6	
1	25	19	33	32	35	27	$n_{D1=1}$ =171
2	22	30	30	19	27	22	$n_{D1=2}$ =150
3	30	27	23	37	23	31	$n_{D1=3}$ =171
4	22	24	39	27	24	49	$n_{D1=4}$ =185
5	32	28	25	15	14	30	$n_{D1=5}$ =144
6	42	43	12	34	28	20	$n_{D1=6}$ =179

$n_{D2=1}$ =173 $n_{D2=3}$ = 162 $n_{D2=5}$ = 151

$n_{D2=2}$ =171 $n_{D2=4}$ =164 $n_{D2=6}$ = 179



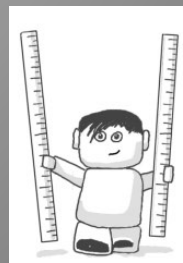
La frecuencia relativa de obtener D1=4 condicionada por que D2=5 se puede obtener como:

$$f(D1 = 4 | D2 = 5) = \frac{n(D1 = 4, D2 = 5)}{n(D2 = 5)} = \frac{24}{151} = 0.1589$$

La frecuencia relativa de obtener D2=5 condicionada por que D1=2 se puede obtener como:

$$f(D2 = 5 | D1 = 4) = \frac{n(D1 = 4, D2 = 5)}{n(D1 = 4)} = \frac{24}{185} = 0.1297$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Distribuciones multivariantes: frecuencia condicionada

Las frecuencias relativas condicionadas de una variable bivariante cumplen las siguientes propiedades:

$$\sum_{i=1}^{k} f(x_i \mid y_j) = \frac{\sum_{i=1}^{k} n_{ij}}{n_{y_j}} = 1 \qquad \sum_{j=1}^{l} f(y_j \mid x_i) = \frac{\sum_{j=1}^{l} n_{ij}}{n_{x_i}} = 1$$

A su vez partiendo de la definición tendremos:

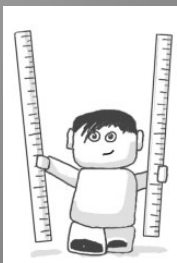
$$f(x_i \mid y_j) = \frac{n_{ij}}{n_{y_j}} = \frac{f_{ij}}{f_{y_j}} \qquad f(y_j \mid x_i) = \frac{n_{ij}}{n_{x_i}} = \frac{f_{ij}}{f_{x_i}}$$

$$f_{ij} = f(x_i \mid y_j)\ f_{y_j} = f(y_j \mid x_i)\ f_{x_i}$$

Podremos, entonces, también escribir:

$$f_{x_i} = \sum_{j=1}^{l} f_{ij} = \sum_{j=1}^{l} f(x_i \mid y_j)\ f_{y_j} \qquad f_{y_j} = \sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} f(y_j \mid x_i)\ f_{x_i}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Distribuciones multivariantes: independencia estadística

Diremos que se cumple la independencia estadística de los eventos $X = x_i$ e $Y = y_j$ cuando se verifica que:

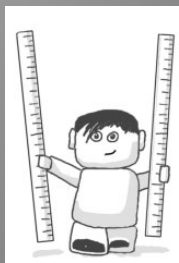
$$f(x_i \mid y_j) = f_{x_i}$$

Como consecuencia de esta afirmación, tendremos que:

$$f(x_i \mid y_j) = f_{x_i} \Leftrightarrow \frac{f_{ij}}{f_{y_j}} = f_{x_i} \Leftrightarrow f_{ij} = f_{x_i} f_{y_j} \Leftrightarrow \frac{f_{ij}}{f_{x_i}} = f_{y_j} \Leftrightarrow f(y_j \mid x_i) = f_{y_j}$$

Para dos eventos estadísticamente independientes, **la frecuencia relativa del par (x_i, y_j) es igual a la multiplicación de la frecuencia relativa marginal de x_i por la frecuencia relativa marginal de y_j .**

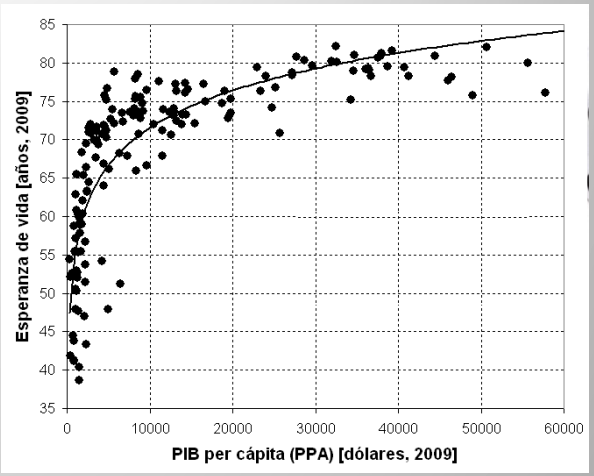
Si esta condición se cumple para todo par de eventos de las variables X e Y diremos que éstas dos variables son estadísticamente independientes.



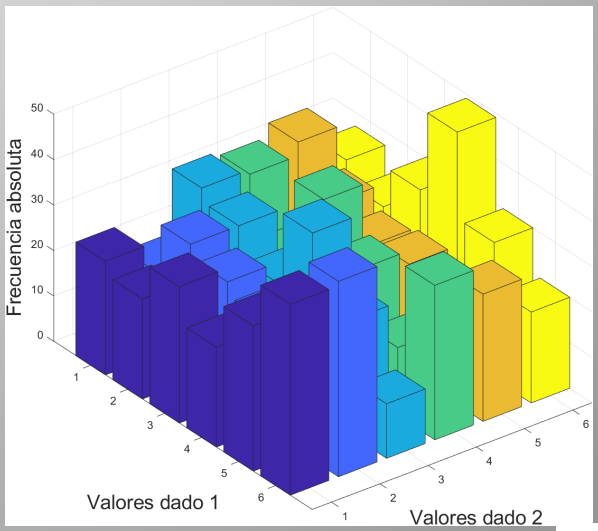
Distribuciones multivariantes: representación gráfica

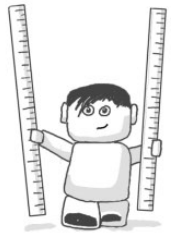
En el caso multivariante, las representaciones gráficas de las distribuciones de frecuencia más importantes son:

● **Diagramas de dispersión** (*scatter plot*): representación sobre ejes coordinados del conjunto de datos mediante un marcador utilizando los valores de la variable como coordenadas $\{(x_i, y_i)\}\ i = 1,2, \dots, N$



● **Diagramas de frecuencia multivariantes** entre los que se encuentran los **diagramas de barras**, **histogramas**, … tal y como se vio en el caso unidimensional. En este caso, las frecuencias se asignan a pares $\{(x_i, y_i)\}\ i = 1, \dots, k\ j = 1, \dots, l$ de los posibles valores de las variables aleatorias $\boldsymbol{X}$ e $\boldsymbol{Y}$ que se colocan sobre una superficie, con barras proporcionales a su frecuencia relativa o absoluta.
Como en el caso unidimensional pueden construirse para las distribuciones de probabilidad simples o acumuladas, sobre casos discretos u organizando en marcas de clase las variables continuas, ...





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

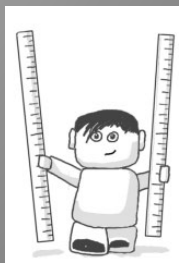
Consideremos una variable estadística bidimensional $(\boldsymbol{X}, \boldsymbol{Y})$, con posibles valores de la variable $\boldsymbol{X}$ $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_k\}$ y de la variable $\boldsymbol{Y}$ $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_l\}$, (ya sea porque se trata de variables discretas o agrupadas en intervalos de clase), definimos como **momento de ordenes r y s respecto al punto (c, d) como:**

$$m_{r,s}(c,d) = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} \ (x_i - c)^r \left(y_j - d\right)^s$$

Siendo f_{ij} la frecuencia relativa del par $\left(x_i, y_j\right)$

Cuando $(\boldsymbol{c}, \boldsymbol{d}) = (\boldsymbol{0}, \boldsymbol{0})$ se denominan momentos respecto al origen, y en el caso de $(\boldsymbol{c}, \boldsymbol{d}) = (\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ se denominan momentos respecto a las medias de la distribución o momentos centrales.

Veremos a continuación medidas características muestrales multivariantes que se construyen utilizando los momentos de distinto orden de la distribución de frecuencias multivariante.

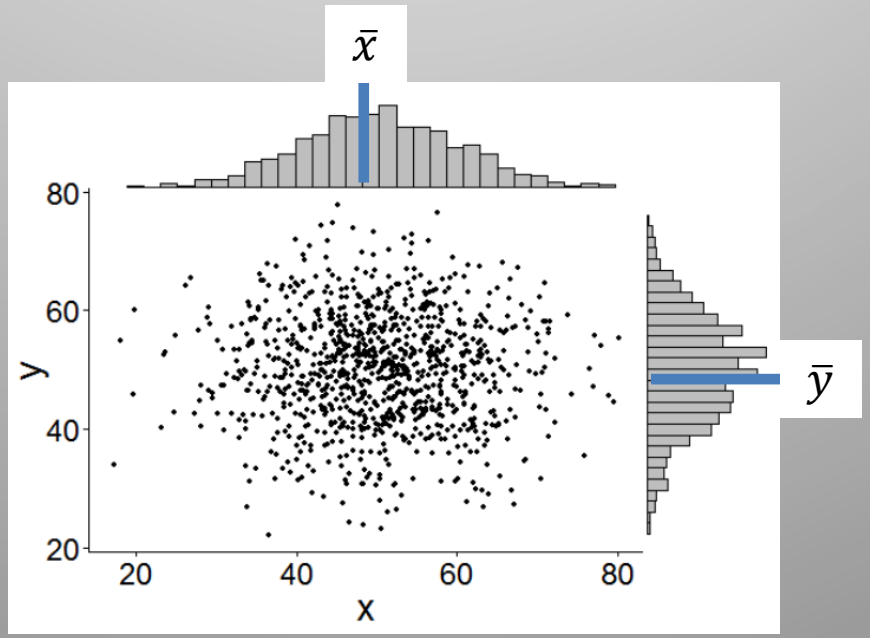


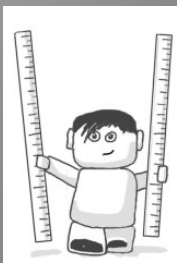
Distribuciones multivariantes: medidas características

La **media aritmética de cada variable estadística** que compone la **variable bidimensional** (X, Y), se define como el **momento de orden uno respecto al origen para la variable promediada:**

$$\bar{x} = m_{1,0}(0,d) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ (x_i - 0)^1 (y_j - d)^0 = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ x_i = \sum_{i=1}^{k} f_{x_i}\, x_i$$

$$\bar{y} = m_{0,1}(c,0) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ (x_i - c)^0 (y_j - 0)^1 = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ y_j = \sum_{j=1}^{l} f_{y_j}\, y_j$$





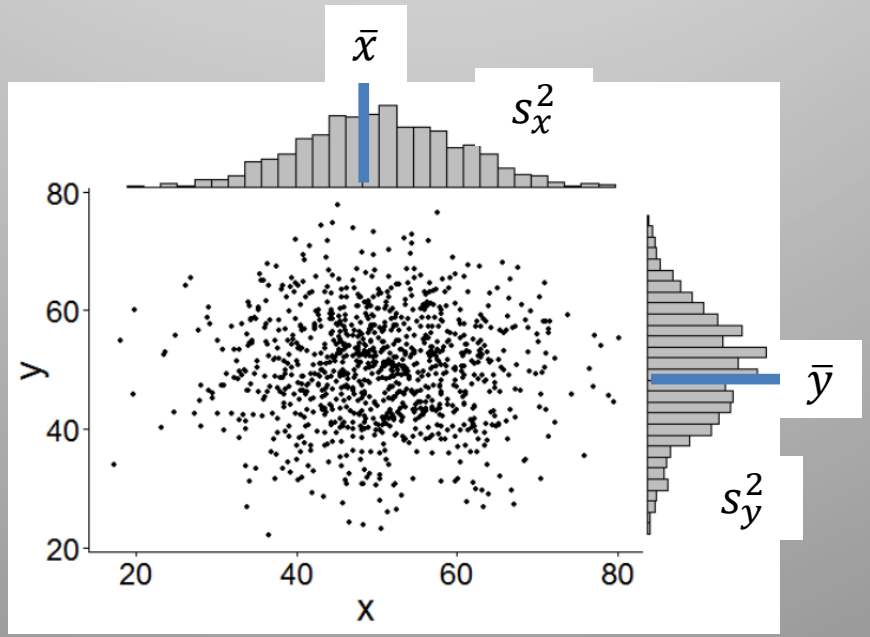
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

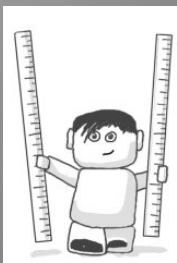
Distribuciones multivariantes: medidas características

La **varianza de cada variable estadística** que compone la **variable bidimensional** $(\boldsymbol{X}, \boldsymbol{Y})$, se define como el **momento de orden dos respecto al valor medio para cada variable:**

$$s_x^2 = m_{2,0}(\bar{x}, d) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ (x_i - \bar{x})^2 (y_j - d)^0 = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_{x_i}(x_i - \bar{x})^2$$

$$s_y^2 = m_{0,2}(c, \bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ (x_i - c)^0 (y_j - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\ (y_j - \bar{y})^2 = \sum_{j=1}^{l} f_{y_j}(y_j - \bar{y})^2$$





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

La **covarianza** de una **variable aleatoria bidimensional** (X, Y), se define como el **momento de orden 1,1 respecto al valor medio de ambas variables:**

$$\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} \ (x_i - \bar{x})^1 (y_j - \bar{y})^1$$

La covarianza puede desarrollarse como:

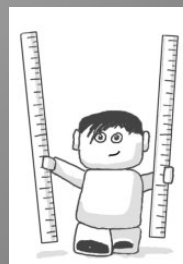
$$\text{cov}(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} \ (x_i - \bar{x}) \ (y_j - \bar{y}) =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} x_i y_j - \bar{y} \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} x_i - \bar{x} \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} y_j + \bar{x}\bar{y}$$

$$\text{cov}(x, y) = \overline{xy} - \bar{x}\bar{y}$$

Donde denota el valor medio del producto de las variables x e y, que puede extenderse a la definición del valor medio de cualquier función $g(x, y)$ de las variables aleatorias como:

$$\overline{g(x, y)} = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} \ g(x_i, y_j)$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

La **covarianza** tiene las siguientes propiedades:

$$\text{cov}(x, y) = \text{cov}(x, y)$$

$$\text{cov}(x, x) = s_x^2 \qquad \text{cov}(y, y) = s_y^2$$

$$\text{cov}(a + bx, c + dy) = bd\ \text{cov}(x, y) \quad a, b, c, d \in \mathbb{R}$$

Si las dos variables aleatorias son independientes, esto es, se cumple que:

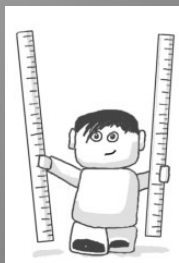
$$f_{ij} = f_{x_i} f_{y_j} \quad \forall i, j$$

Entonces:

$$\text{cov}(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} x_i y_j - \bar{x}\bar{y} = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{x_i} f_{y_j} x_i y_j - \bar{x}\bar{y} = 0$$

X,Y variables aleatorias independientes $\Rightarrow \text{cov}(x, y) = 0$

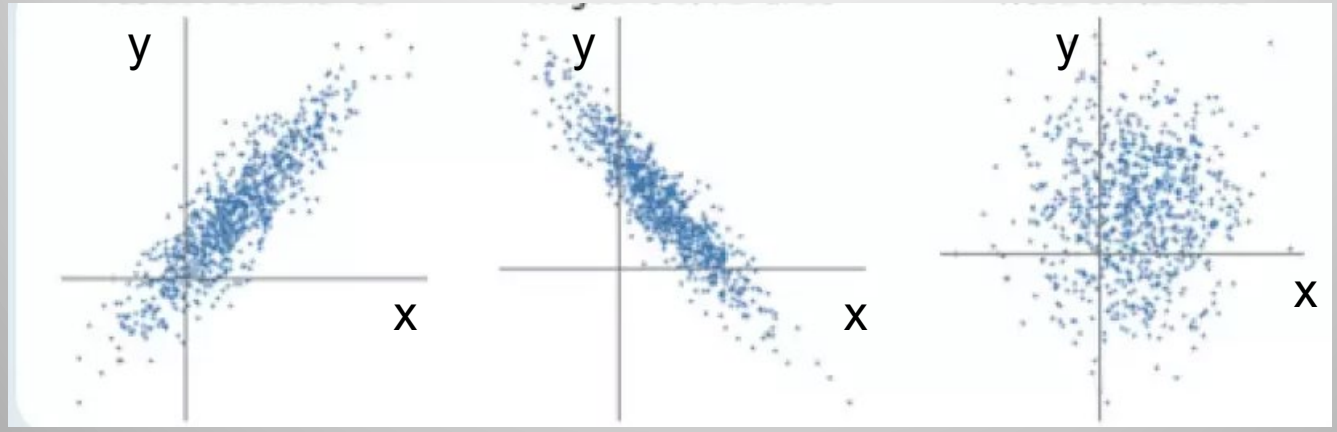
$\text{cov}(x, y) = 0 \nRightarrow$ X,Y variables aleatorias independientes



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

La **covarianza** indica el grado de correlación entre las variables aleatorias:



$\text{cov}(x, y) > 0$

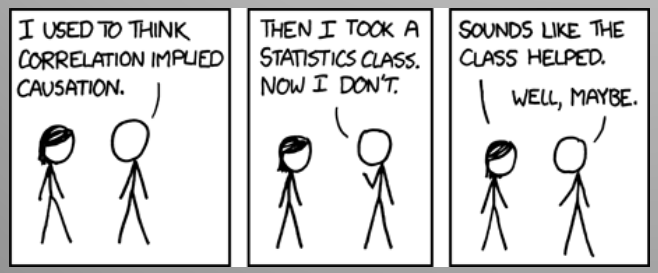
Correlación positiva

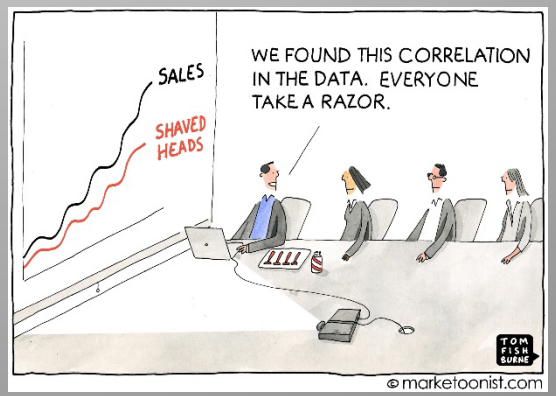
$\text{cov}(x, y) < 0$

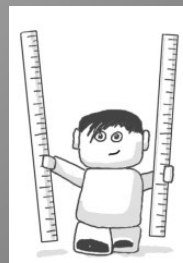
Correlación negativa

$\text{cov}(x, y) \approx 0$

Ausencia de correlación







USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

Como la covarianza no es invariante ante cambios de escala (ver transformacion lineal), se introduce el **coeficiente de correlación lineal entre variables estadísticas** $(\boldsymbol{X}, \boldsymbol{Y})$ como:

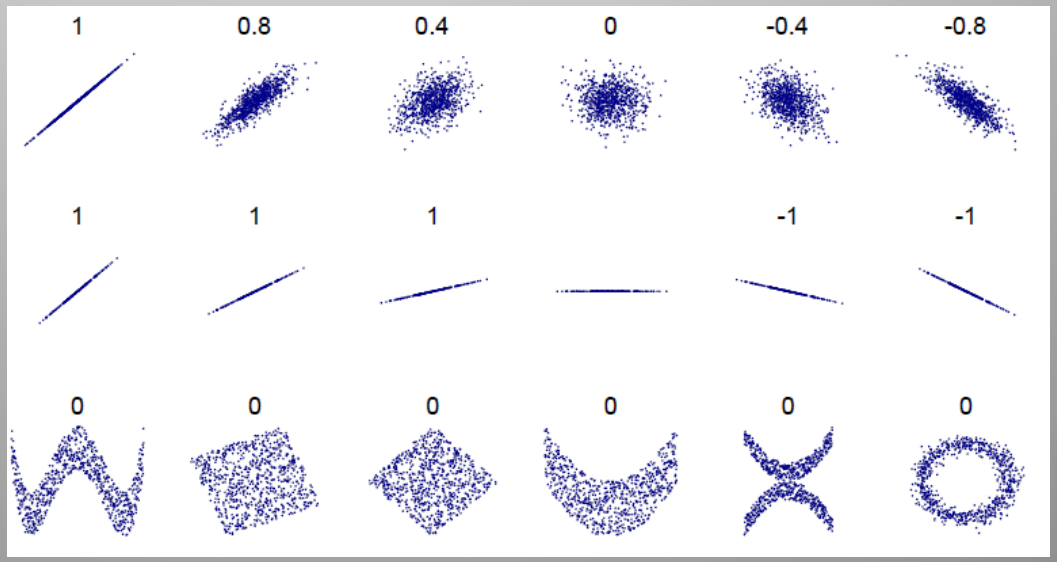
$$r(x,y) = \frac{\mathrm{cov}(x,y)}{s_x\, s_y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\,(x_i - \bar{x})\,(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{k} f_{x_i}(x_i - \bar{x})^2}\,\sqrt{\sum_{j=1}^{l} f_{y_j}(y_j - \bar{y})^2}}$$

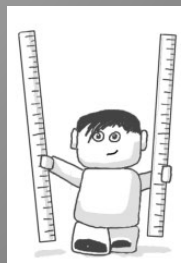
El coeficiente de correlación lineal es **adimensional** e **invariante ante cambios de escala**.

$$-1 \le r(x,y) \le 1$$

Cuando $r(x,y) = \pm 1$ las dos variables están linealmente correlacionadas.

En cambio cuando $r(x,y) \approx 0$ las variables no muestran correlación.





USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

Veamos que se verifica que el **coeficiente de correlación lineal entre variables estadísticas** (X, Y) está acotado entre -1 y 1:

$$-1 \leq r(x, y) \leq 1$$

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} \left[(x_i - \bar{x}) - \delta \left(y_j - \bar{y} \right) \right]^2 \geq 0$$

Para cualquier $\delta \in \mathbb{R}$

Si desarrollamos esta expresión:

$$S = s_x^2 + \delta^2 s_y^2 - 2\,\delta\,\mathrm{cov}(x, y) \geq 0$$

Basta tomar:

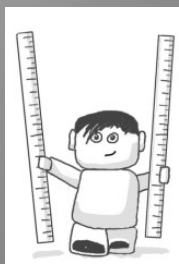
$$\delta = \frac{\mathrm{cov}(x, y)}{s_y^2}$$

Para obtener:

$$s_x^2 s_y^2 \geq [\mathrm{cov}(x, y)]^2$$

Y por lo tanto:

$$1 \geq \left| \frac{\mathrm{cov}(x,y)}{s_x\, s_y} \right|$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Distribuciones multivariantes: medidas características

Podemos ver que si **coeficiente de correlación lineal entre variables estadísticas** $(\boldsymbol{X}, \boldsymbol{Y})$ es -1 o 1 existe una relación lineal entre ambas variables:

Si consideramos una variable aleatoria de la forma:

$$s^2(x + \lambda y) = s_x^2 + \lambda^2 s_y^2 + 2\,\lambda\,\mathrm{cov}(x, y)$$

Supongamos que (se demuestra igual en el caso -1):

$$r(x, y) = 1 \;\Rightarrow\; \mathrm{cov}(x, y) = s_x s_y$$

Por lo tanto:

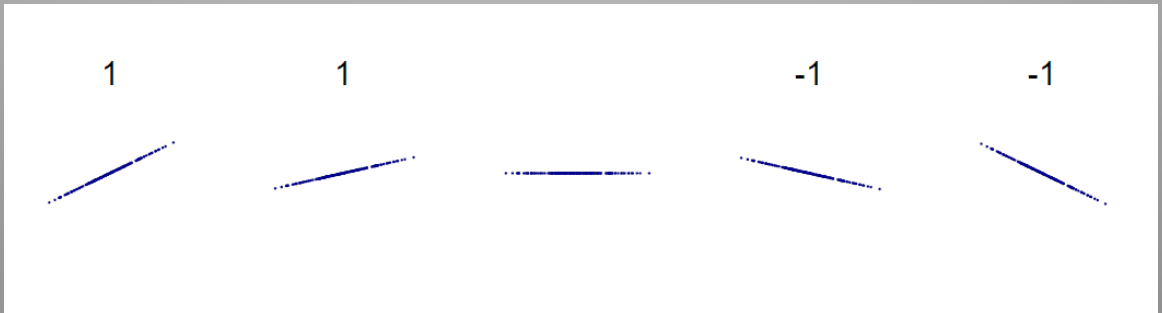
$$s^2(x + \lambda y) = s_x^2 + \lambda^2 s_y^2 + 2\,\lambda\, s_x s_y = \left(s_x + \lambda\, s_y\right)^2$$

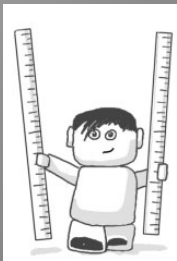
Basta tomar:

$$\lambda = -{}^{s_x}/_{s_y}$$

Para obtener que:

$$s^2(x + \lambda y) = 0$$





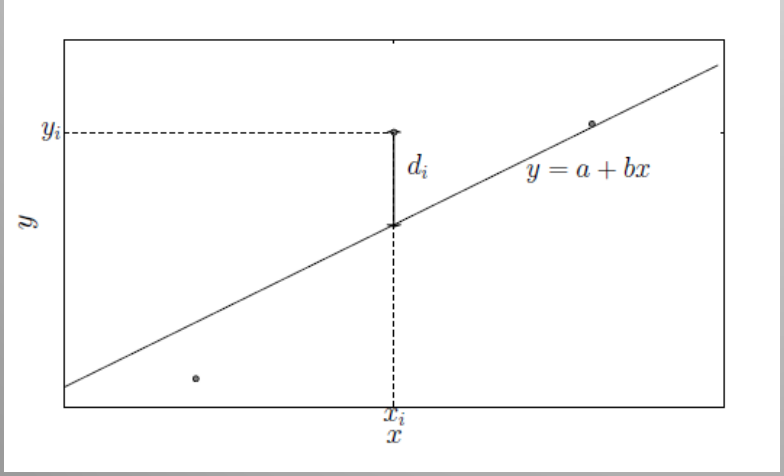
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Regresión lineal y covarianza

Consideremos una muestra de N datos de una variable aleatoria bivariante $(\boldsymbol{X}, \boldsymbol{Y})$ de la forma $\{(x_1, y_1); (x_2, y_2); \ldots; (x_N, y_N)\}$ sobre la que consideramos la hipótesis de que existe una cierta relación lineal que las describe:

$$Y = a + bX$$

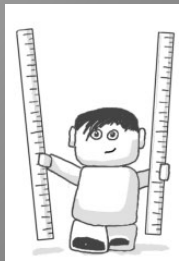
Los datos están sujetos a ciertas fuentes de incertidumbre, de modo que, si aceptamos la hipótesis de su descripción mediante una regresión lineal, intentaremos encontrar los valores de los parámetros a y b más verosímiles teniendo en cuenta los resultados experimentales.



Estableceremos como medida de verosimilitud:

$$\mathcal{L}(a, b) = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

Pensando que los valores de a y b más verosímiles son los que hacen menor esta función.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Regresión lineal y covarianza

Si buscamos el mínimo de la función de verosimilitud (función objetivo), en este caso:

$$0 = \frac{\partial}{\partial a}\mathcal{L}(a,b) \;\Rightarrow\; \sum_{i=1}^{N} y_i = a\,N + b\sum_{i=1}^{N} x_i$$

$$0 = \frac{\partial}{\partial b}\mathcal{L}(a,b) \;\Rightarrow\; \sum_{i=1}^{N} x_i y_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2$$

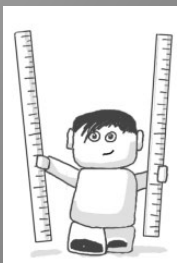
Estas dos ecuaciones para los parámetros de la regresión lineal pueden ser escritas como:

$$\bar{y} = a + b\bar{x}$$

$$\overline{xy} = a\,\bar{x} + b\,\overline{x^2}$$

Podemos determinar la solución para la pendiente de la regresión b, mediante:

$$\overline{xy} = (\bar{y} - b\bar{x})\,\bar{x} + b\,\overline{x^2}$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Regresión lineal y covarianza

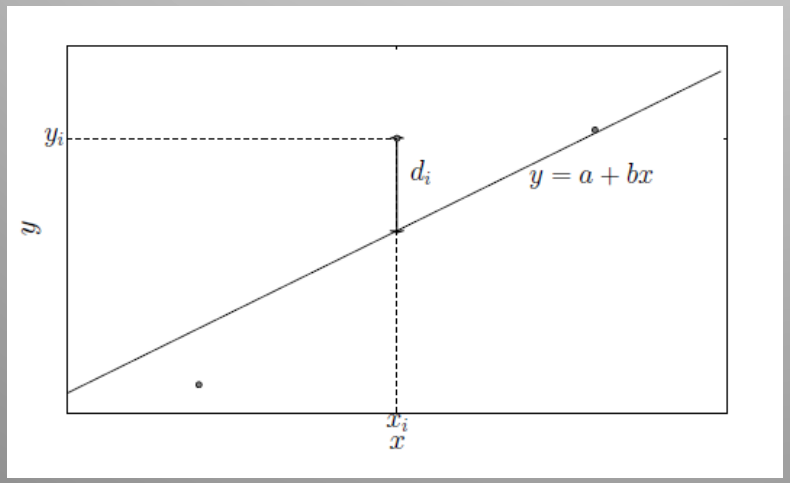
Esto arroja como parámetros más verosímiles de la regresión:

$$b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\mathrm{cov}(x,y)}{s_x^2} \qquad a = \bar{y} - \frac{\mathrm{cov}(x,y)}{s_x^2}\,\bar{x}$$

Observemos que la covarianza nos indica el signo de la pendiente y determina su valor junto con la varianza de la abcisa. Si la covarianza es nula, entonces la pendiente de la recta es nula.

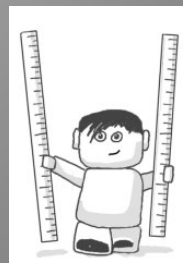
El valor del intercepto en el origen depende tanto de la covarianza como de los valores medios de abcisas y ordenadas.

Un cambio de origen de coordenadas no modifica la covarianza ni la varianza de las coordenadas x e y, por tanto, no altera el valor de la pendiente b, mientras que sí altera el valor del parámetro a.



$$a = \bar{y} - \frac{\mathrm{cov}(x,y)}{s_x^2}\,\bar{x}$$

$$b = \frac{\mathrm{cov}(x,y)}{s_x^2}$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Matriz de covarianza

Si consideramos tres variables aleatorias X, Y, Z, podremos definir los valores medios y varianzas de modo análogo al caso bidimensional:

$$\bar{x} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\, x_i \qquad \bar{y} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\, y_j \qquad \bar{z} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\, z_k$$

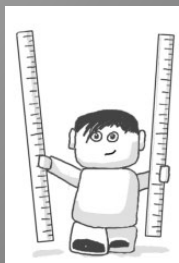
$$s_x^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\,(x_i - \bar{x})^2 \qquad s_y^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\,(y_j - \bar{y})^2 \qquad s_z^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}\,(z_k - \bar{z})^2$$

$$\mathrm{cov}(x,y) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}(x_i - \bar{x})(y_j - \bar{y}) \qquad \mathrm{cov}(x,z) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}(x_i - \bar{x})(z_k - \bar{z})$$

$$\mathrm{cov}(y,z) = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} f_{ijk}(z_k - \bar{z})(y_j - \bar{y})$$

Esto nos permite definir una matriz real y simétrica, denominada matriz de covarianzas:

$$M = \begin{pmatrix} s_x^2 & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(x,y) & s_y^2 & \mathrm{cov}(y,z) \\ \mathrm{cov}(x,z) & \mathrm{cov}(y,z) & s_z^2 \end{pmatrix}$$



Matriz de covarianza

En el caso de tener N variables aleatorias $X_1, \dots, X_N$, podremos definir los valores medios y varianzas de modo general:

$$\overline{x_k} = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} f_{i_1 i_2 \dots i_k \dots i_N} \; x_{i_k} \qquad k = 1, .., N$$

$$s_{x_k}^2 = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} f_{i_1 i_2 \dots i_k \dots i_N} \; \left(x_{i_k} - \overline{x_k}\right)^2 \qquad k = 1, .., N$$

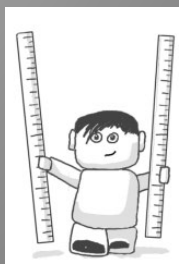
$$\mathrm{cov}(x_k, x_p) = \sum_{i_1} \sum_{i_2} \dots \sum_{i_N} f_{i_1 i_2 \dots i_k \dots i_p \dots i_N} \; \left(x_{i_k} - \overline{x_k}\right)\left(x_{i_p} - \overline{x_p}\right) \quad k, p = 1, .., N$$

La matriz de covarianzas multidimensional será:

$$M = \begin{pmatrix} s_{x_1}^2 & \mathrm{cov}(x_1, x_2) & \dots & \mathrm{cov}(x_1, x_N) \\ \mathrm{cov}(x_2, x_1) & s_{x_2}^2 & \dots & \mathrm{cov}(x_2, x_N) \\ \mathrm{cov}(x_3, x_1) & \mathrm{cov}(x_3, x_2) & \dots & \mathrm{cov}(x_3, x_N) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(x_N, x_1) & \mathrm{cov}(x_N, x_2) & \dots & s_{x_N}^2 \end{pmatrix}$$

Esta matriz tiene siempre un determinante positivo que nos permite definir la varianza generalizada s_g^2.

$$s_g^2 = \det(M) \geq 0$$



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA