

Técnicas experimentales II

Estadística

Héctor Alvarez Pol

Departamento de Física de Partículas
Universidade de Santiago de Compostela

Técnicas Experimentales II - G1031225

Grado en Física - G1031V01

Curso 2018 / 2019

1. Estadística descriptiva

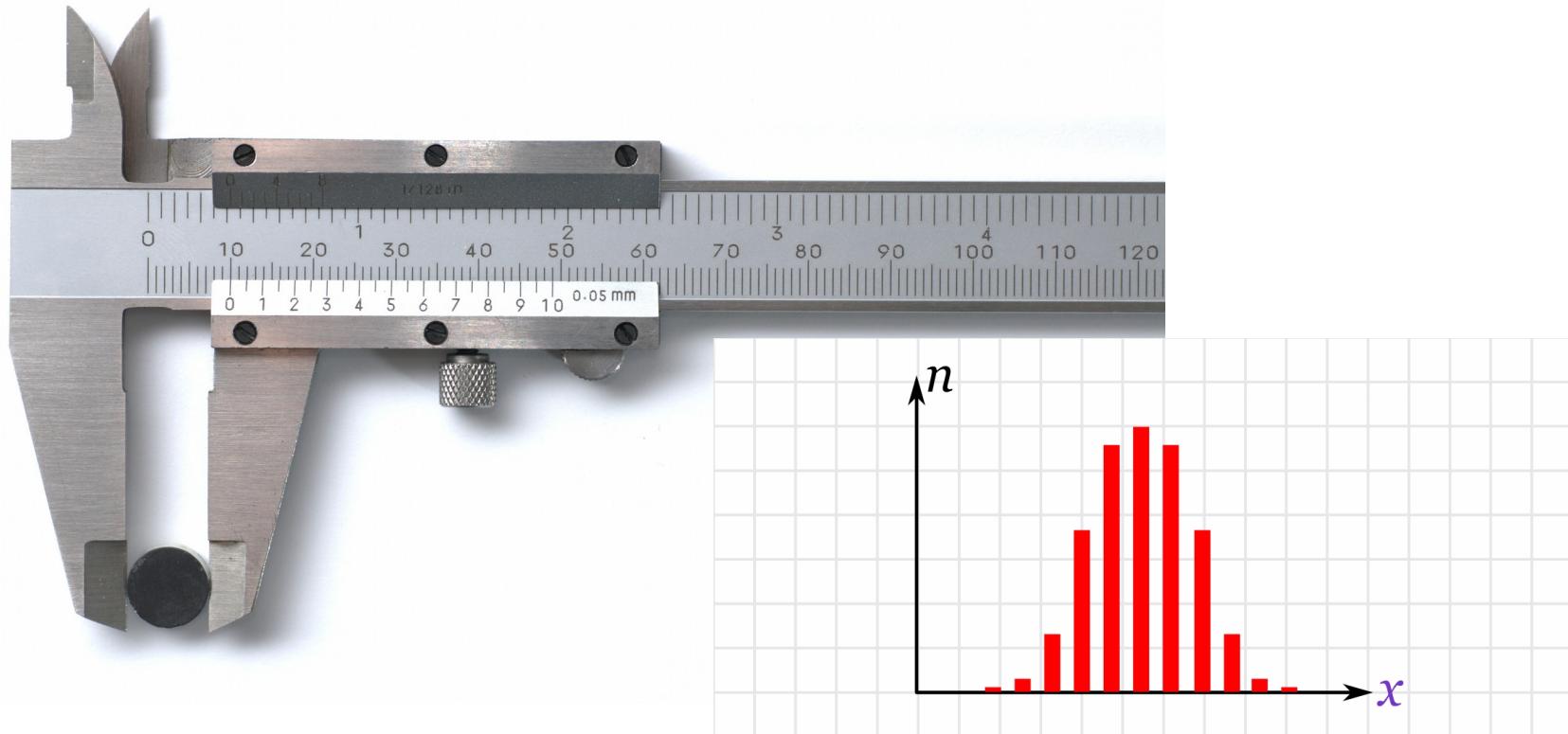
- Introducción al tema, glosario y bibliografía.
- Distribuciones de frecuencias.
- Representaciones gráficas de la información.
- Medidas características: centralización, dispersión, asimetría y apuntamiento.
- Transformaciones de variable aleatoria: lineales y no lineales.
- Muestras multivariantes.

Bibliografía: estadística descriptiva

- [VAR10] **Tratamiento de Datos Físicos.** L. M. Varela, F. Gómez, J. Carrete. Servizo de Publicacións e Intercambio Científico. Universidade de Santiago, 2010.
- [SMEP72] **Statistical Methods in Experimental Physics,** *W.T. Eadie, D. Drijard, F.E. James, M. Roos, B. Sadoulet.* North-Holland Publishing Company. CERN, Geneva.
- [Peña08] **Fundamentos de Estadística,** D. Peña. Alianza Editorial, 2008.
- [Wonn87] **Introducción a la estadística,** T. H. Wonnacott, R. J. Wonnacott. Editorial Limusa, 1987.
- [Spie03] **Probabilidad y estadística.** Spiegel Murray. Editorial McGrawHill, 2003.
- [Bev80] **Data reduction and error analysis for the physical sciences.** P.R Bevington, D.K Robinson. McGraw-Hill, 1980.

Introducción: estadística descriptiva

- **Medir consiste en determinar una magnitud accesible a la experimentación mediante la comparación** de una cantidad de una cierta magnitud con otra de su misma clase que se adopta como **patrón**.
- Cuando se realizan medidas, se obtienen resultados numéricos que corresponden a los distintos procesos de medida, distribuidos de acuerdo con la forma en la que el mensurando se distribuye o con su dependencia con variables asociadas al proceso de medida.

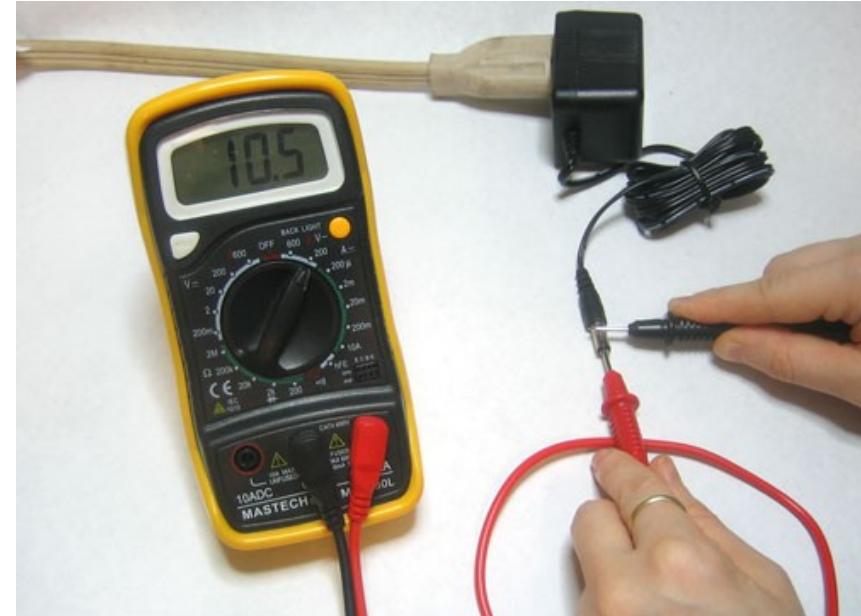


Introducción: estadística descriptiva

Ejemplos de procesos de medida:

1) **Medida de voltajes con un polímetro analógico:** la medida de la diferencia de potencial se ve afectada por errores en la calibración del instrumento de medida, errores de paralelaje en la posición del operador en su visión del marcador del instrumento, variaciones en la humedad del aire que pueden afectar a la resistencia interna del aparato de medida, ...

2) **Medida del conteo de radiación alpha mediante un detector de Geiger-Müller:** el conteo por unidad de tiempo se podrá ver afectado por errores del operador y del instrumento de medida del tiempo, ... pero independientemente de esto, cada contaje puede diferir del anterior reflejando la distribución de probabilidad del mensurando.



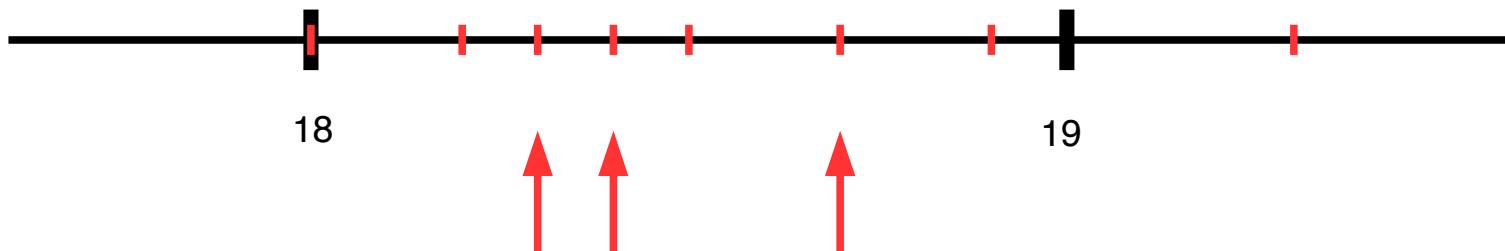
Introducción: estadística descriptiva

Ejemplo: **Mediciones.**

- 1) El resultado de la medición puede no ser un valor constante. La medida con un termómetro de la temperatura de una muestra criogénica que devuelve los valores en kelvin da como resultados:
18,2, 18,7, 18,9, 18,4, 18,4, 18,5, 18,0, 18,3, 19,3, 18,3, 18,7 .

Nos anticiparemos a la definición de la media de una serie de medidas:

$$\bar{T} = (1/n) \sum_{i=1}^n T_i = (1/11) \sum_{i=1}^{11} T_i = 203,7/11 = 18,52 \text{ K}$$



Introducción: estadística descriptiva. Glosario

Medida: conjunto de operaciones que conllevan a determinar el valor de una cantidad.

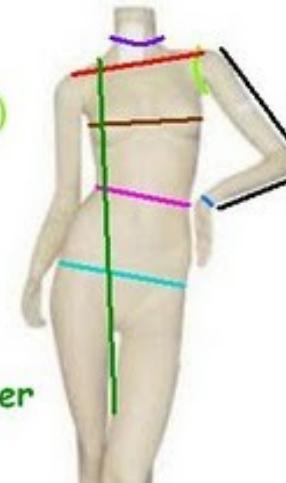
Principio de medida: base científica de una medida.

Método de medida: secuencia lógica de operaciones, descritas genéricamente, usadas en la realización de medidas.

Procedimiento de medida: conjunto de operaciones, descritas de forma específica, utilizadas en la realización de medidas particulares de acuerdo con un método de medida establecido.

Cantidad de influencia o influyente: cantidad que no es el mensurando pero que afecta al resultado de la medida de modo determinante.

morado: contorno de cuello
rojo: ancho de espalda
cafe: contorno de pecho
verde: contorno de hombro (sisa)
rosa: contorno de cintura
celeste: contorno de cadera
negro: largo de manga (medir de la misma manera que en la imagen)
azul: contorno de muñeca
verde pasto: largo total (puede ser hasta la cintura o como desee)



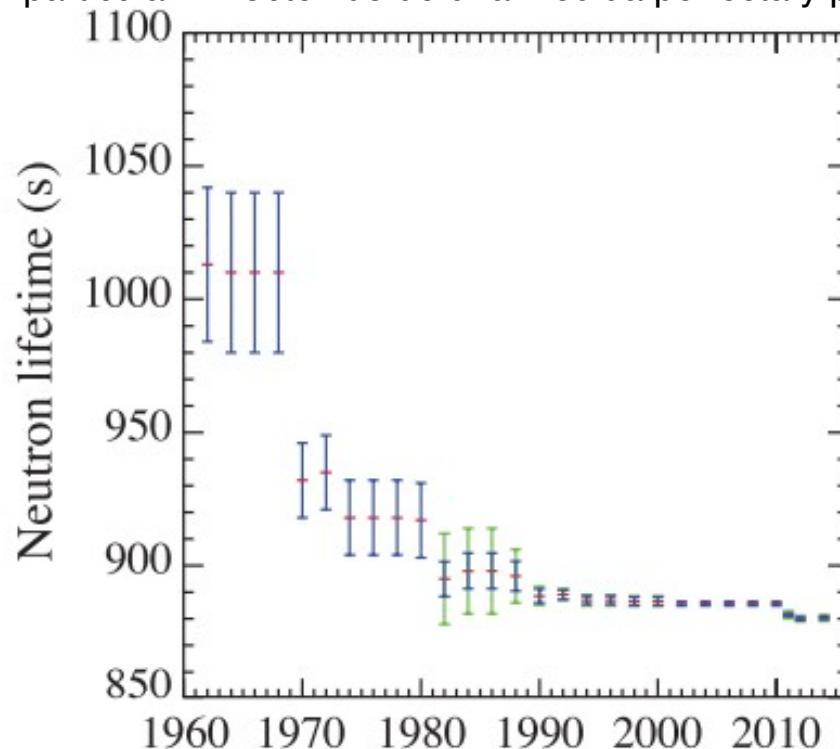
Introducción: estadística descriptiva. Glosario

Cantidad (medible): atributo de un fenómeno, cuerpo o sustancia que se puede distinguir cualitativamente y determinar cuantitativamente.

Mensurando: cantidad particular sujeta a medida.

Valor (de una cantidad o mensurando): magnitud de una cantidad particular expresado como una unidad de medida multiplicada por un número.

Valor real (de una cantidad o mensurando): valor consistente con la definición de una cantidad particular. El obtenido de una medida perfecta y por su naturaleza, indeterminado.



Valor real convencional: valor atribuido para una cantidad particular, aceptando, en ocasiones por convenio, que tenga una incertidumbre apropiada a su uso.

← **Valor real convencional actual.**

← **Quizá el valor real esté aquí!**

<http://pdg.lbl.gov/2015/reviews/rpp2015-rev-history-plots.pdf>

Introducción: estadística descriptiva. Glosario

Precisión (o fidelidad): grado de coincidencia entre los resultados independientes que se obtienen de una serie de medidas realizadas bajo idénticas condiciones de medida.

Exactitud de la medida: cercanía del acuerdo entre el resultado de una medida y el valor real del mensurando (es un concepto cualitativo).

Veracidad (justeza o justicia de una medida): grado de coincidencia entre el resultado de nuestra medida y el valor real convencional.

Sesgo o desviación (de una medida): diferencia entre el valor obtenido en la medida y el valor real del mensurando (en algún caso también se emplea para el valor real convencional).

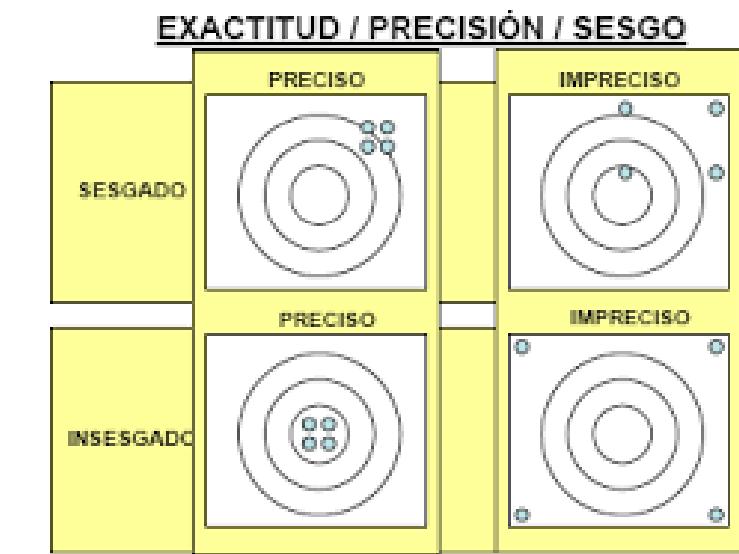
■ Veracidad



■ Exactitud



■ Precisión



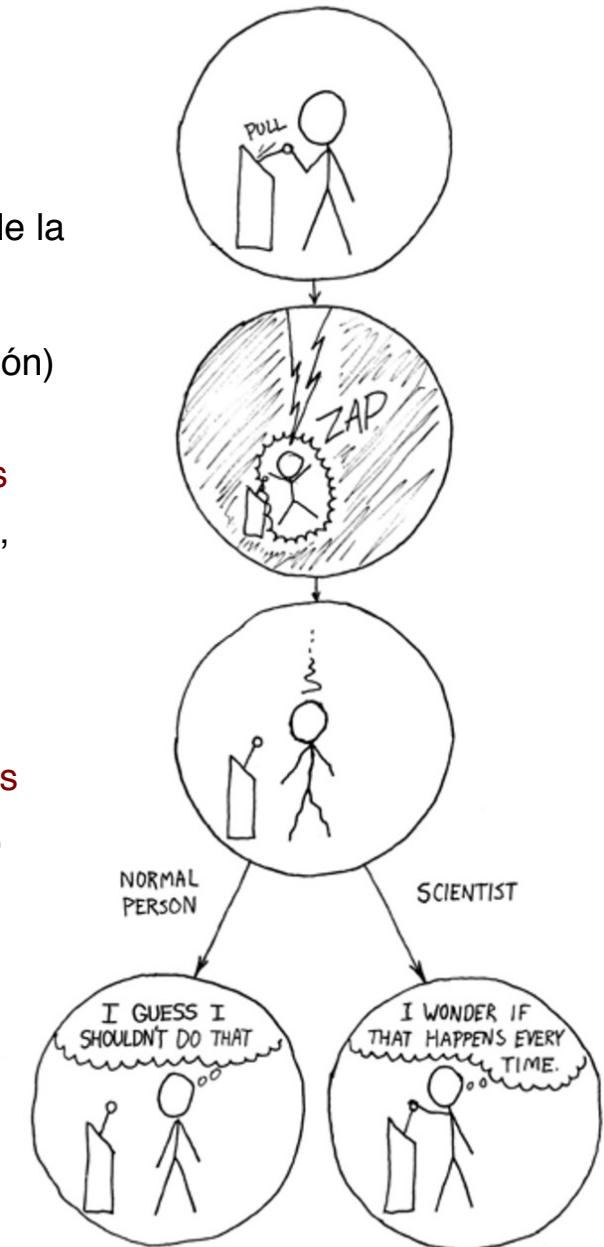
Introducción: estadística descriptiva. Glosario

Resultado de una medida: valores atribuibles al mensurando, obtenidos mediante una medida.

Resultados corregidos/no corregidos: aquellos después/antes de la corrección por errores sistemáticos.

Repetibilidad (de los resultados de una medida): cercanía (precisión) del acuerdo entre los resultados de sucesivas medidas independientes del mismo mensurando realizadas **bajo las mismas condiciones de medida** (operador, tiempo, aparatos, lugar, método, ...).

Reproducibilidad (de los resultados de una medida): cercanía (precisión) del acuerdo entre los resultados de medidas independientes del mismo mensurando realizadas **bajo condiciones cambiantes de medida** (operador, tiempo, aparatos, lugar, método, ...). **Las leyes físicas deben ser reproducibles.**



<https://xkcd.com/>

Introducción: estadística descriptiva. Glosario

Población: conjunto de todos los elementos, fenómenos, individuos o unidades estadísticas sobre los que se puede manifestar una propiedad objeto de medida. Puede ser finito o infinito.

Muestra: subconjunto de la población escogido para su análisis estadístico.

Experimento aleatorio: aquel en el que los resultados posibles, conocidos de antemano, pueden sucederse sin que sea posible conocer el resultado concreto de cada intento, pudiéndose repetir en condiciones idénticas un número indeterminado de veces.

Variables (cantidades o mensurandos) cuantitativas o estadísticas: corresponden a valores numéricos asignables a los sucesos aleatorios. Pueden ser discretas o continuas.

Variables (cantidades o mensurandos) cualitativas: son las que se corresponden a valores no numéricos a los sucesos aleatorios.



Introducción: estadística descriptiva. Glosario

La manera de considerar los resultados de una medida física en el que obtenemos valores de un mensurando a través de un experimento aleatorio es a través de la definición de variable aleatoria.

Variable aleatoria: sobre un espacio muestral (Ω) de un determinado fenómeno aleatorio en el que se defina la función real X que asocia cada suceso posible (A) a un valor de la recta real, X se denomina **variable aleatoria unidimensional**:

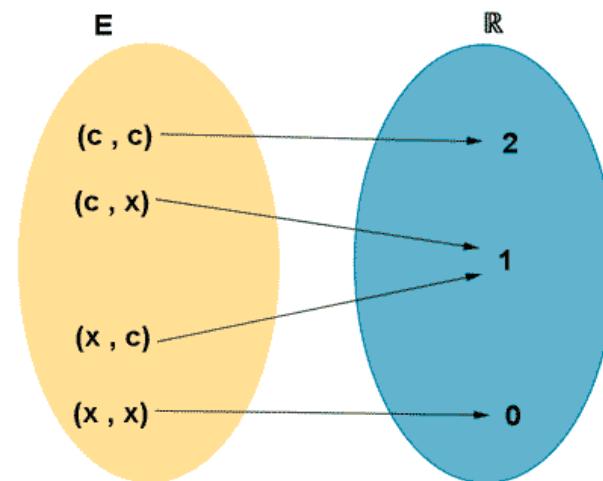
$$X : \Omega \rightarrow \mathbb{R}$$

$$A \rightarrow X(A)$$

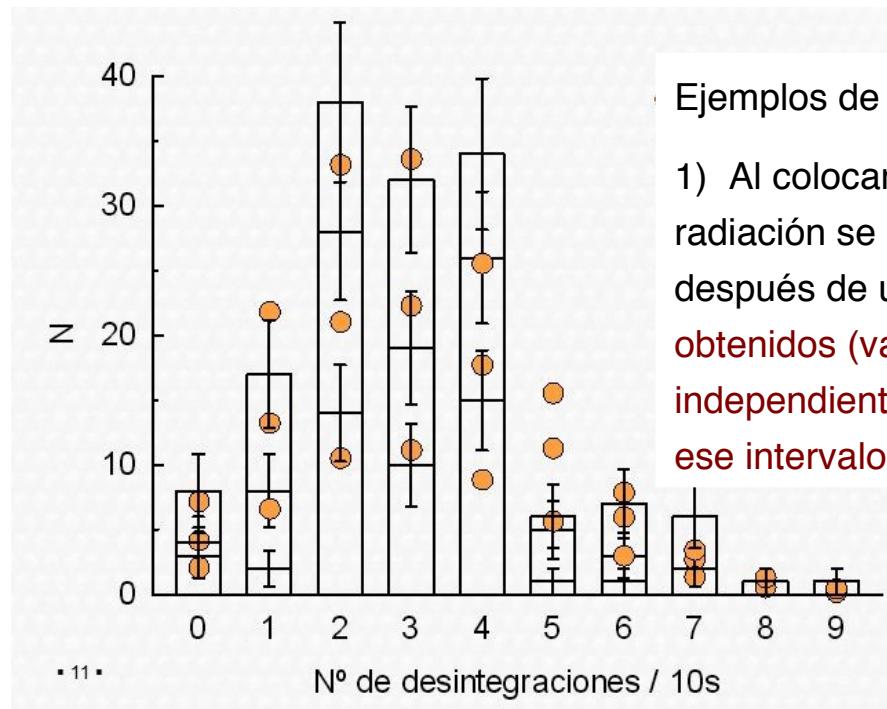
Así, tras la realización de un experimento aleatorio de N ensayos de medición, tendremos N valores

$$\{x_i\}_{i=1}^N = \{x_1, x_2, \dots, x_N\}$$

Se puede extender fácilmente a variables aleatorias multidimensionales en los que varias características se estudien o resulten de cada experimento aleatorio (el resultado independiente de lanzar dos monedas, estudiar la distribución de momentos y posiciones de una partícula, ...).

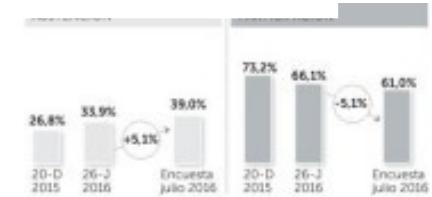
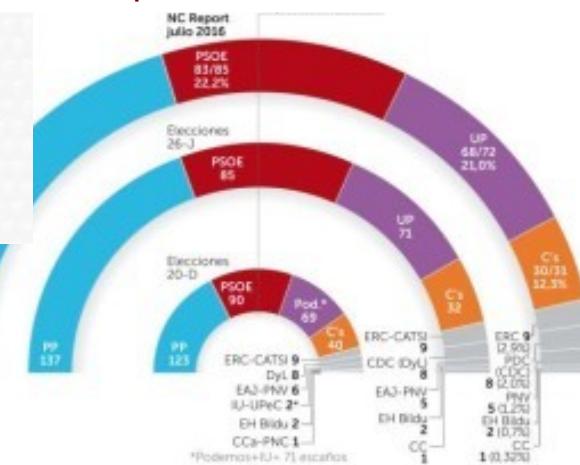


Introducción: estadística descriptiva. Glosario



Ejemplos de experimentos aleatorios:

1) Al colocar una muestra radiactiva enfrente de un contador de radiación se obtienen diferentes conteos al parar el contador después de un intervalo de 10 segundos. **Cada uno de los conteos obtenidos (variables cuantitativa discreta)** constituye una medida independiente del experimento aleatorio de conteo de radiación en ese intervalo de tiempo.



	18/29	30/44	45/64	65 y +	Total
PP	11.6	15.2	21.0	29.2	29.2
PSOE	10.3	12.2	13.0	15.9	13.1
UP	14.8	16.0	11.2	8.7	12.4
C's	8.4	8.4	6.5	6.4	7.3
Otros	5.8	5.3	7.7	3.0	5.6
Blanco	0.6	0.8	0.3	0.4	0.5
Abs./Nulo	48.4	42.2	40.2	36.4	41.0

2) En un estudio de intención de voto se pregunta telefónicamente a un grupo de personas, cuyos teléfonos se escogen al azar, por su edad, sexo, estado civil e intención de voto político (en caso de que sean mayores de edad). **Los resultados de este experimento aleatorio constituyen una variable estadística aleatoria de cuatro dimensiones**. La muestra (las personas contactadas) constituyen un subconjunto de la población completa con derecho a voto.

Distribución de frecuencias

Para describir los resultados de un experimento, podemos detallar todos los valores obtenidos después de los **N** ensayos de medición o bien podemos observar cuales son idénticos (en caso de que pertenezcan a un conjunto discreto de posibles resultados) o cuales están comprendidos en intervalos determinados en los que podamos organizar nuestros datos. En el caso de encontrar **k** posibles resultados, podremos ordenarlos y numerarlos: $x_1, x_2, \dots, x_i, \dots, x_k$

Se denomina **frecuencia absoluta**, (n_i), asociada al valor **i**-esimo de los resultados de la variable aleatoria, al número de veces que se observa el valor en el total de las medidas realizadas.

La suma de las frecuencias absolutas sobre todos los resultados es **N**: $\sum_{i=1}^k n_i = N$

Se denomina **frecuencia relativa**, (f_i), asociada al valor **i** de los resultados de la variable aleatoria, a la fracción de veces que se observa el valor en el total de las medidas realizadas: $f_i = \frac{n_i}{N}$

La suma de las frecuencias relativas sobre todos los resultados es 1: $\sum_{i=1}^k f_i = 1$

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

Distribución de frecuencias. Frecuencias acumuladas

Resulta relevante definir la **frecuencia acumulada absoluta (o relativa)** como el número de veces (o fracción de las veces) que se observan valores inferiores a uno dado en el total de las medidas realizadas.

La **frecuencia acumulada absoluta** hasta el valor i -esimo será:

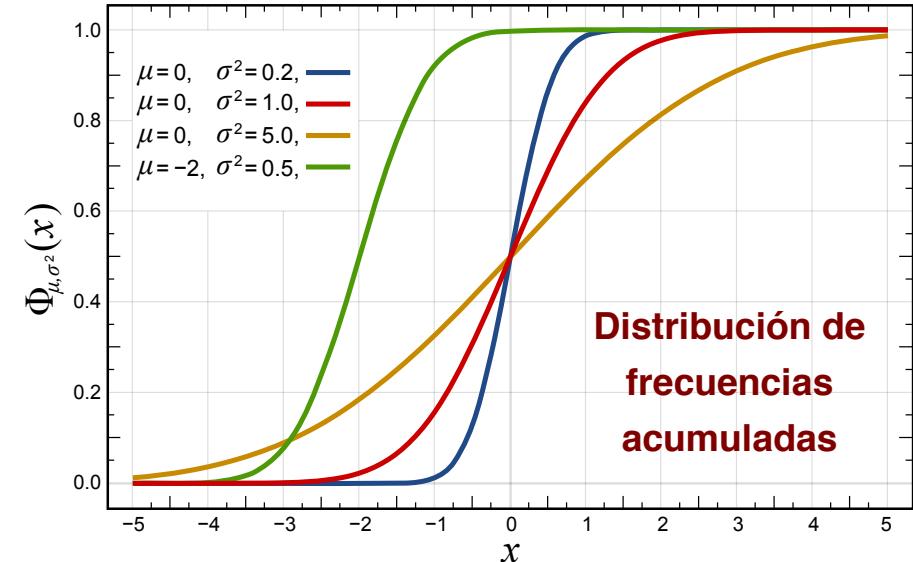
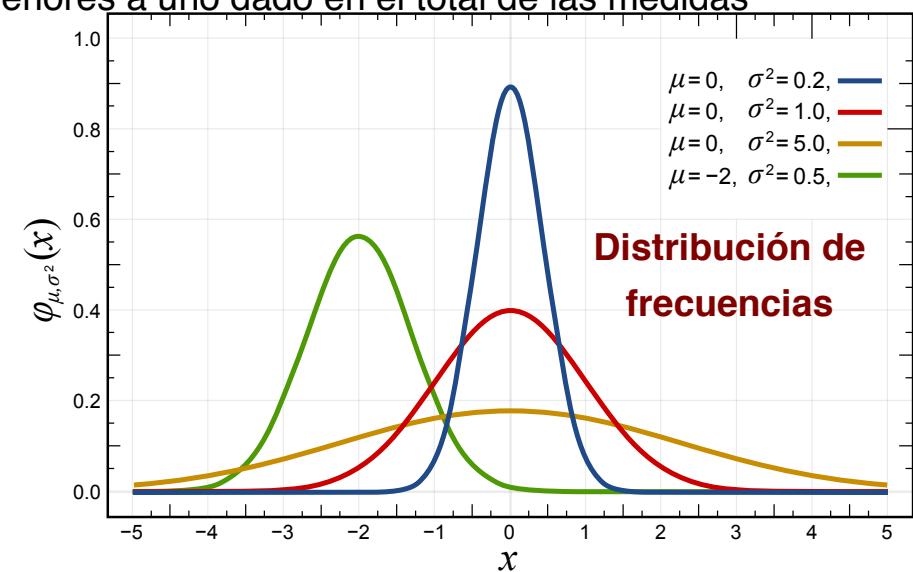
$$N_i = \sum_{j=1}^i n_j \quad N_k = \sum_{j=1}^k n_j = N$$

La **frecuencia acumulada relativa** hasta el valor i -esimo será:

$$F_i = \sum_{j=1}^i f_j = \frac{N_i}{N} \quad F_k = \sum_{j=1}^k f_j = 1$$

Las **frecuencias acumuladas son funciones monótonas crecientes** que cumplen:

$$\begin{aligned} 0 &\leq n_i \leq N & 0 &\leq f_i \leq 1 \\ N_i &= N_{i-1} + n_i & F_i &= F_{i-1} + f_i \end{aligned}$$



Distribución de frecuencias. Marcas de clase

En el caso de variable aleatoria continua (o de variable discreta con un gran número de posibles resultados), se **agruparán los resultados en una serie de intervalos disjuntos**, cuya unión cubra todo el posible recorrido de la variable. A los sucesos dentro de cada clase se les etiqueta por un suceso característico, el valor central del intervalo o, en general, una **marca de clase**.

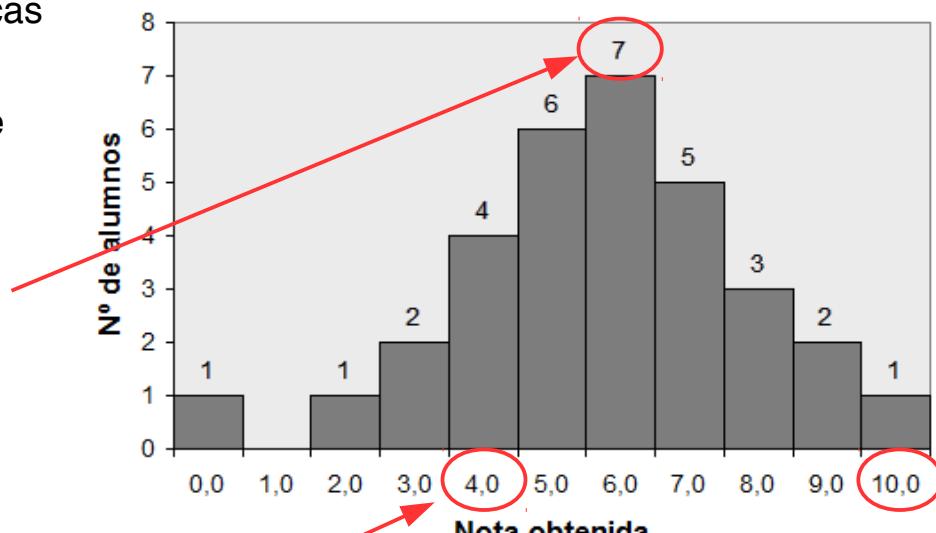
- Los resultados anteriores referidos a frecuencias o frecuencias acumuladas tanto relativas como absolutas, se generalizan sumando las frecuencias de todos los resultados que pertenezcan a una marca de clase determinada. Así, **la frecuencia absoluta para una marca de clase corresponderá al número de resultados que pertenezcan a su clase asociada**.
- No existen, sin embargo, normas referentes a como realizar la agrupación, aunque si ciertas reglas para optimizar la selección.

- La operación de agrupar los resultados en marcas de clase **conlleva una perdida de información**, asociada a la agrupación de resultados dentro de una marca sin detallar su posición dentro del intervalo o marca al que se adscribe.

**¿Los 7 resultados son exactamente un 6.0?
¿Habrá algún 6.5? ¿Están homogéneamente distribuidos entre 6.0 y 7.0?**

¿Ambigüedad en las marcas de clase?

Histograma de las notas de 32 alumnos



Distribución de frecuencias. Marcas de clase

Ejemplo: (1.8 en [VAR10]) Datos de una desintegración radiactiva medidas por un contador Geiger en el que se han efectuado medidas cada 30 segundos.

7	5	3	6	8	4	5	7	6	4	6	3	4	4	5	7	5	4	3	4
9	8	9	7	0	3	5	8	7	8	4	6	5	5	7	4	7	3	5	2
3	5	3	8	4	9	4	10	3	5	5	8	6	7	6	5	6	6	2	9
8	9	5	9	6	5	5	7	3	7	6	4	6	9	4	5	7	6	5	8
4	8	4	5	7	8	7	6	5	4	5	7	8	3	9	6	1	6	1	5
7	5	3	9	8	1	6	4	7	8	5	6	11	9	7	4	5	10	7	4
6	4	6	10	7	6	2	13	3	6	0	8	1	6	8	1	11	6	8	3
5	6	9	4	10	7	6	7	9	6	3	7	5	12	7	8	6	3	5	6
2	7	5	6	7	5	5	2	4	6	9	2	5	10	2	9	5	5	7	4
2	6	7	8	4	5	7	6	6	7	5	4	3	2	6	8	7	1	6	5
10	8	3	2	8	4	6	3	3	8	4	5	6	7	8	6	9	8	3	2
11	2	6	5	5	7	9	8	5	2	4	6	6	3	5	4	6	4	4	5
7	5	6	7	4	10	6	7	4	5	8	7	5	5	4	6	3	8	6	6
12	10	5	6	12	3	11	4	10	4	5	4	9	8	3	6	8	7	5	2
3	5	10	7	9	6	7	4	11	7	6	1	11	2	5	9	4	8	5	6

Distribución de frecuencias. Marcas de clase

Ejemplo: (1.8 en [VAR10]) Datos de una desintegración radiactiva medidas por un contador Geiger en el que se han efectuado medidas cada 30 segundos.

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

x_i : resultados de la variable aleatoria x

n_i : frecuencia absoluta

f_i : frecuencia relativa

N_i : frecuencia acumulada absoluta

F_i : frecuencia acumulada relativa

$$\sum_{i=1}^k n_i = 300$$

Distribución de frecuencias. Marcas de clase

- Para optimizar la selección de las marcas de clase en un conjunto de **N** medidas, **se suele elegir un número de clases del orden del entero más proximo a la raíz de N.**

En el ejemplo anterior (1.8 en [VAR10]), con **N = 300**, el número de divisiones óptimo estaría alrededor de 17 (un número superior a las divisiones que aparecían de forma natural en el ejemplo, 14. En este caso no tiene sentido realizar divisiones diferentes a la propia naturaleza del problema).

- **Las clases en las que se dividen los resultados se suelen tomar de la misma amplitud**, salvo en casos en los que la agrupación de valores sugiera otras soluciones.
- **No puede haber ambigüedad en las clases**: un resultado siempre debe ser atribuible a una clase y solo a una. Para ello deben definirse intervalos semiabiertos por un extremo y cerrados por el otro.

Con esto tendremos una división del recorrido de la variable aleatoria en subintervalos o clases:

$$\{ [a_0, a_1), [a_1, a_2), [a_2, a_3), [a_3, a_4), \dots, [a_{k-1}, a_k) \}$$

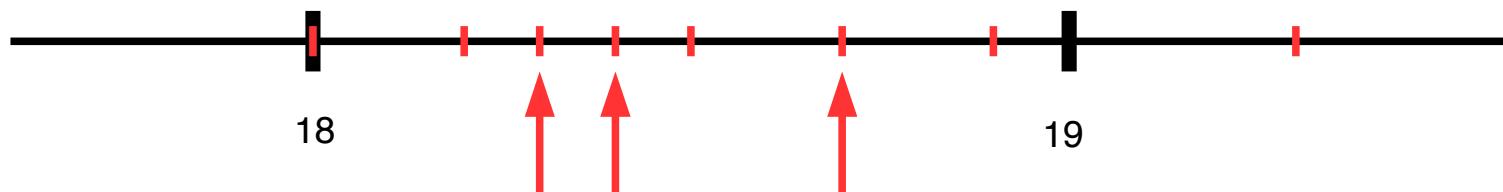
A los que asociaremos marcas de clase: $\{ x_1, x_2, x_3, x_4, \dots, x_{k-1}, x_k \}$ en las que habitualmente se toma el centro de cada intervalo como marca asociada: $x_i = 0.5 (a_{i-1} + a_i)$

- Se establecen las frecuencias de los resultados pertenecientes a cada clase (contando).

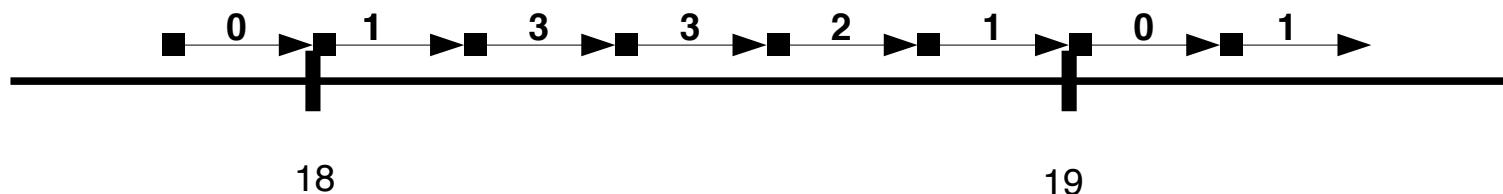
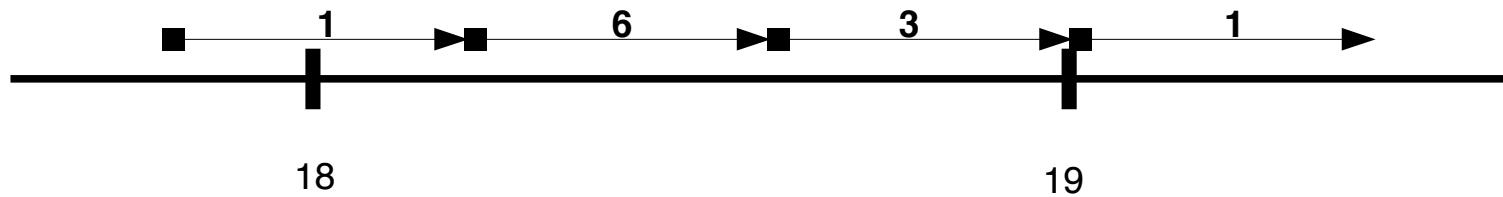
Distribución de frecuencias. Marcas de clase

Ejemplos de división en marcas de clase:

- 1) Volvemos a los datos de un ejemplo anterior: La medida con un termómetro de la temperatura de una muestra criogénica que devuelve los valores en kelvin da como resultados: 18,2, 18,7, 18,9, 18,4, 18,4, 18,5, 18,0, 18,3, 19,3, 18,3, 18,7 .



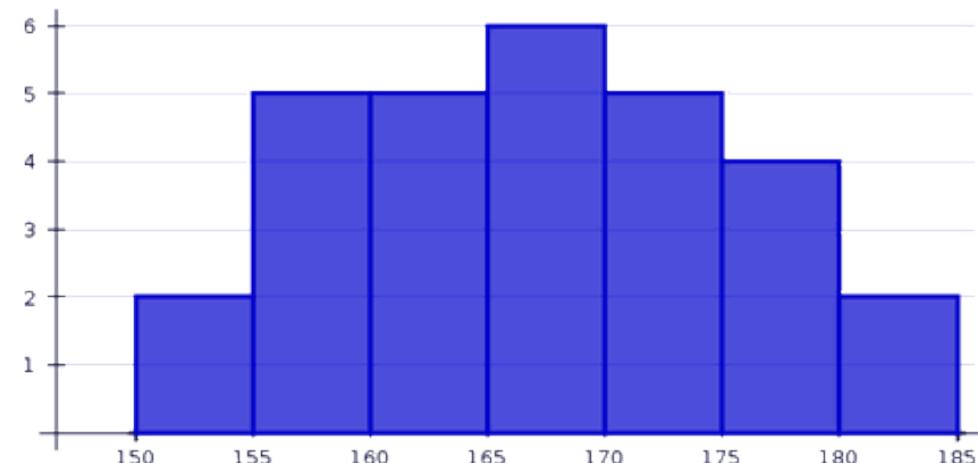
Al tratarse de 11 datos, 3 o 4 divisiones serían lo adecuado, aunque también pueden darse casos en los que un mayor o menor número de divisiones mejoran la comprensión de la distribución.



Representación gráfica

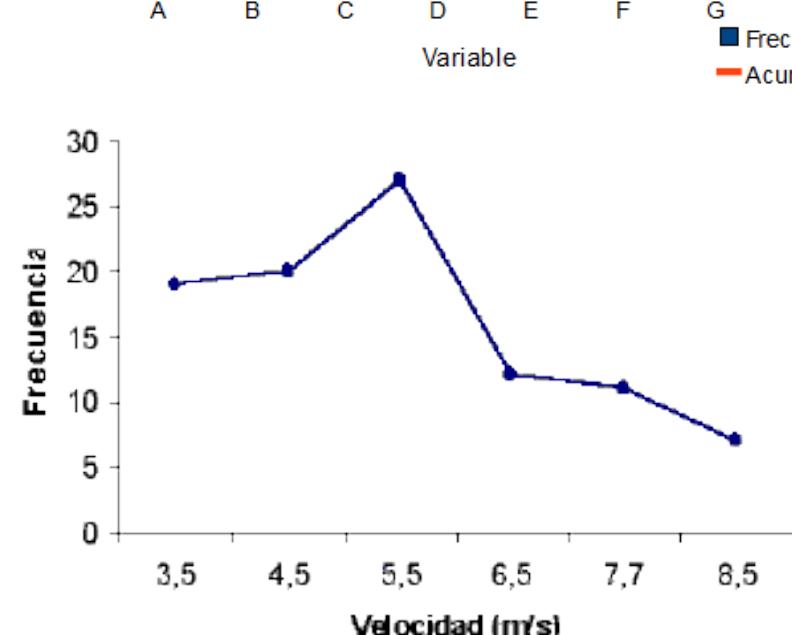
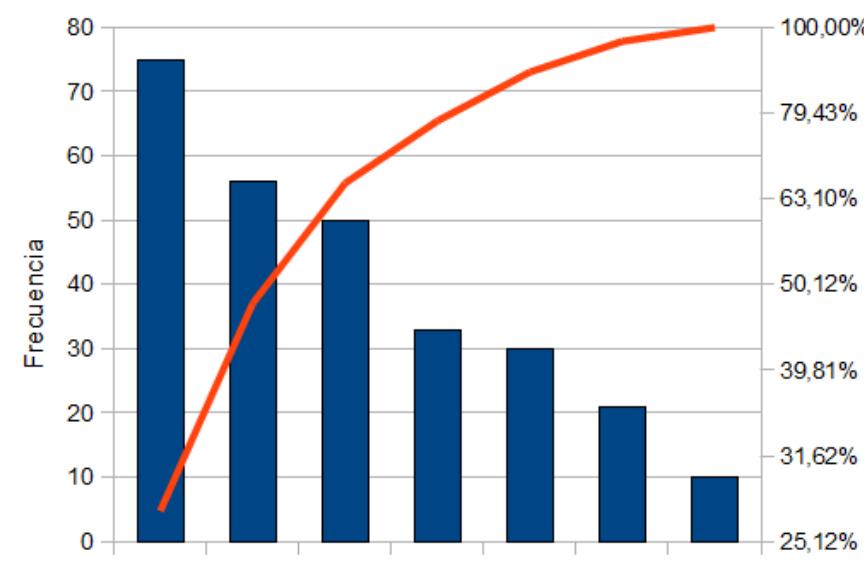
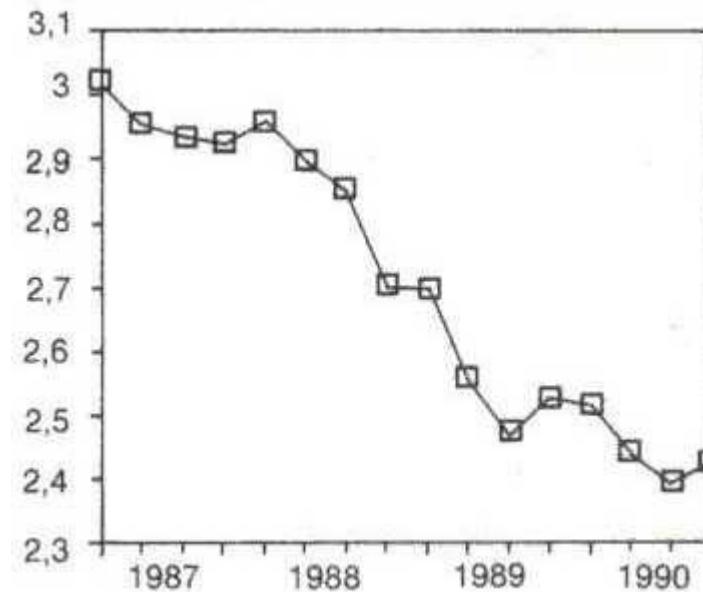
La información contenida en los resultados de un experimento aleatorio puede interpretarse con mayor facilidad utilizando representaciones gráficas de las distribuciones de frecuencia. Las más importantes representaciones gráficas son:

- **Diagramas de barras:** asocia a cada marca de clase o valor de la variable una barra de altura igual a la frecuencia con la que la marca o el valor aparecen. Puede corresponder a frecuencias absolutas o relativas, ordinarias o acumuladas.
- **Histogramas:** asocia a cada clase de la variable un rectángulo cuya base sea la amplitud del intervalo y cuya altura es proporcional a la frecuencia con que aparecen los resultados en la clase.



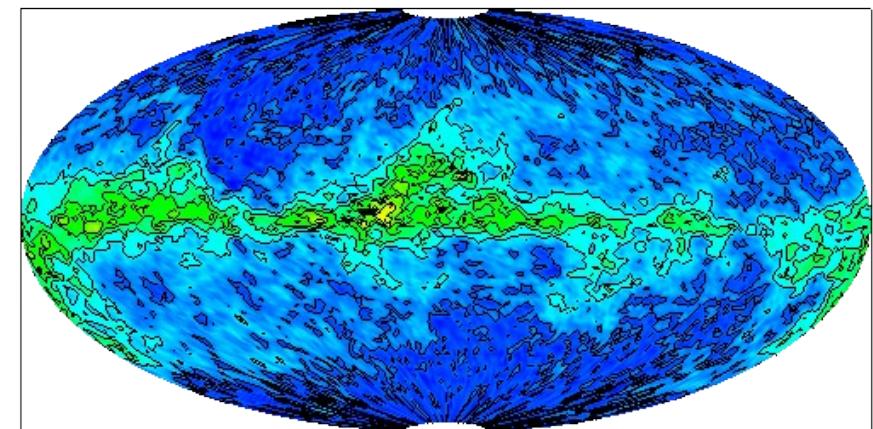
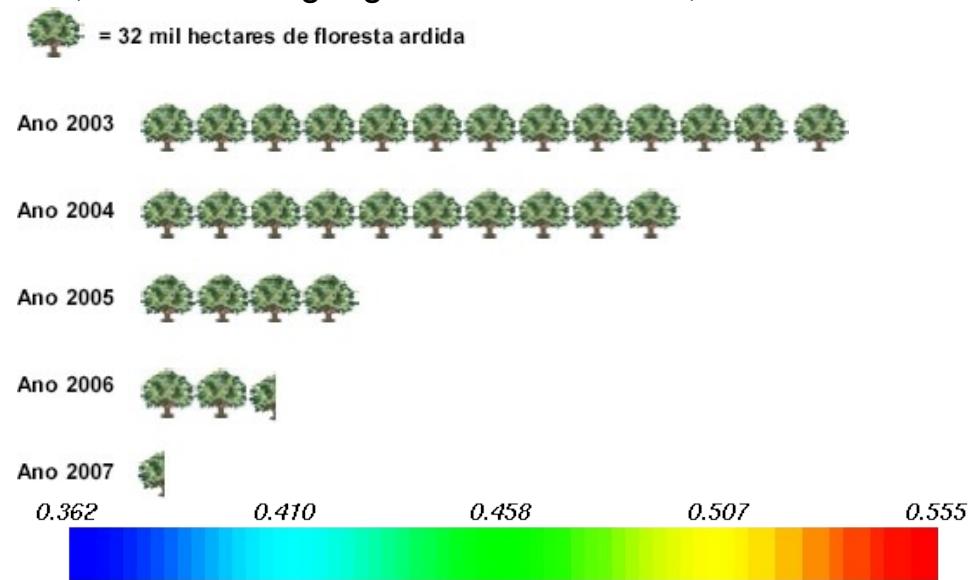
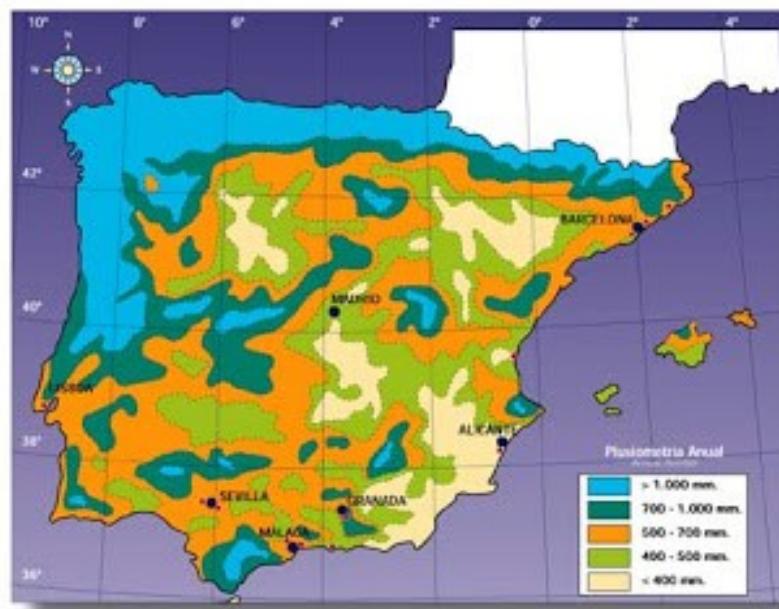
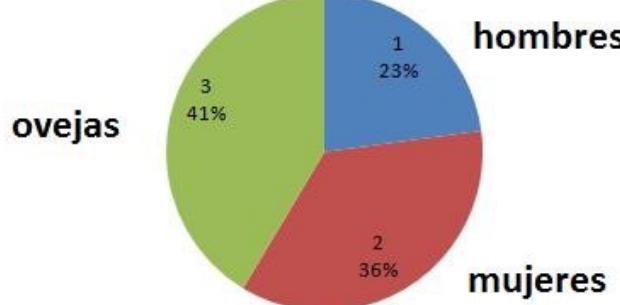
Representación gráfica

- **Diagramas de Pareto:** asocia rectángulos de altura igual a la frecuencia relativa a las clases o eventos ordenados según su frecuencia, en orden descendiente.
- **Polygonos o curvas de frecuencia:** se trazan líneas rectas o curvas que unan los extremos superiores de las barras del diagrama de barras de frecuencias o del histograma correspondiente.
- **Series temporales o de evolución:** corresponde a la representación de una variable aleatoria con respecto a una variable que evolucione, normalmente el tiempo de la medida.



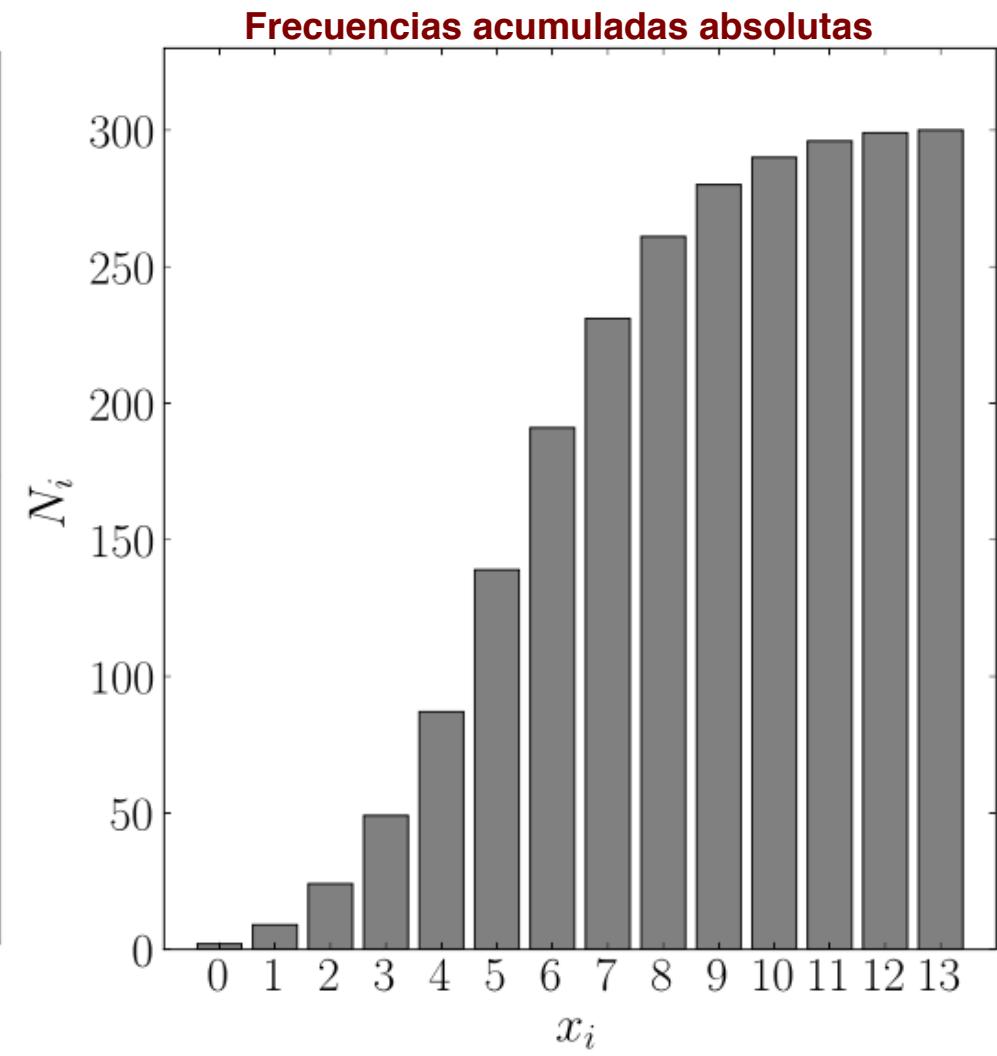
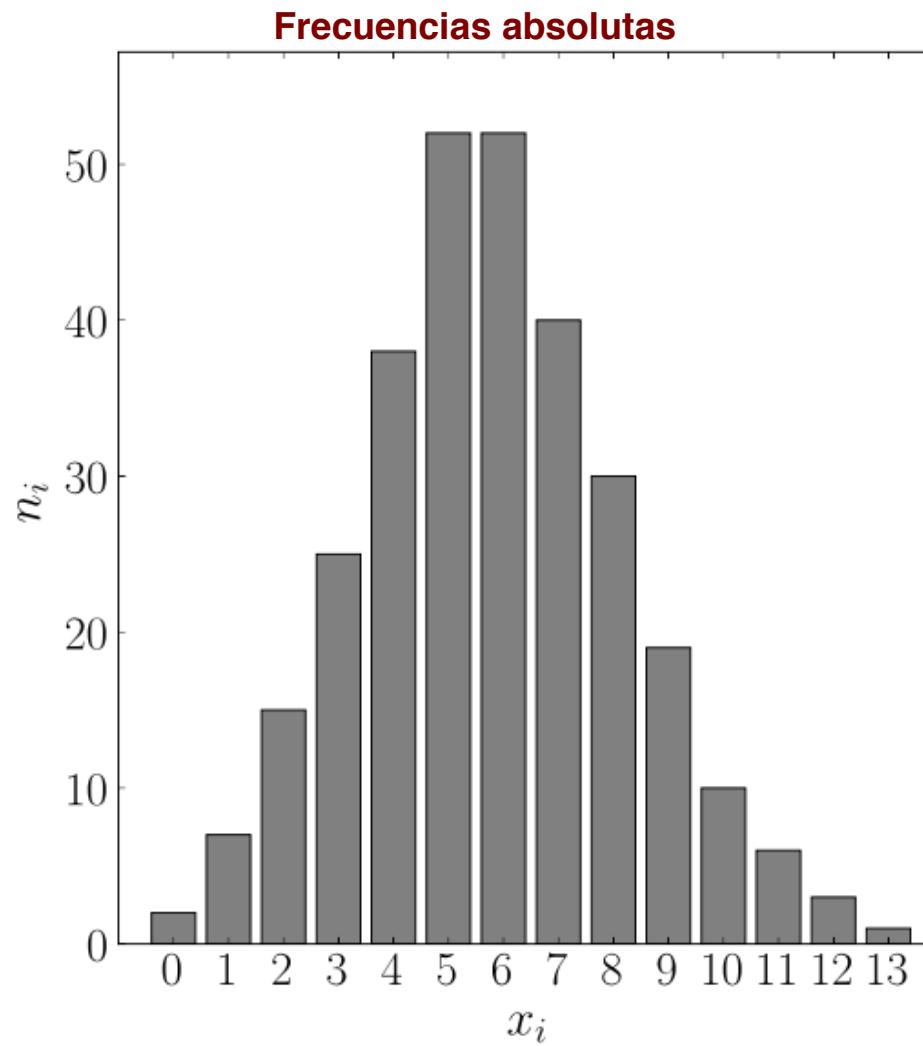
Representación gráfica

- **Diagramas de sectores o circular, pictogramas, cartogramas, ...** : otras representaciones estadísticas variadas que permiten representar los datos mediante la equivalencia de áreas, identificación ideográfica con los posibles resultados, localización geográfica de los datos, ...



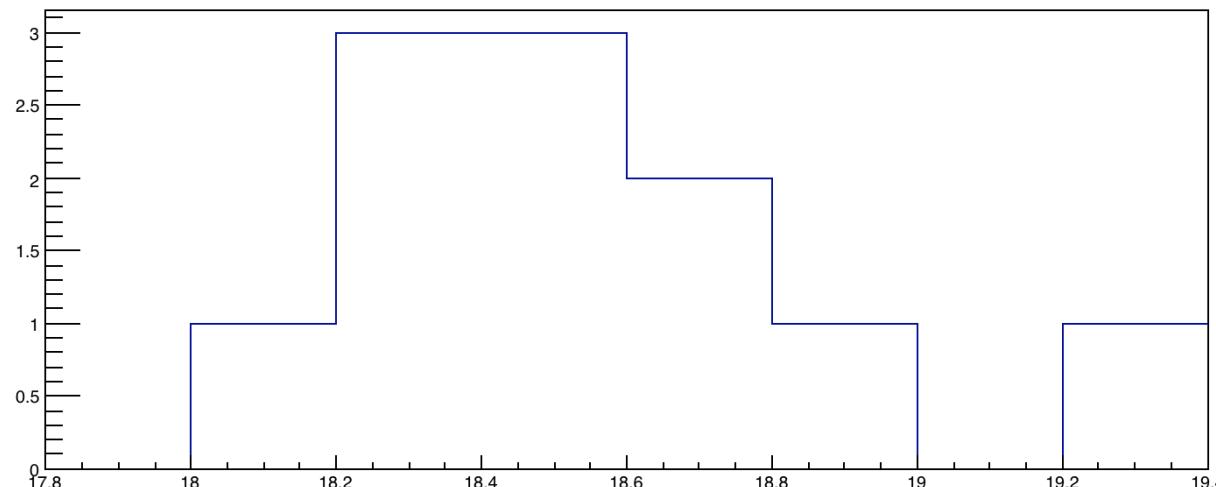
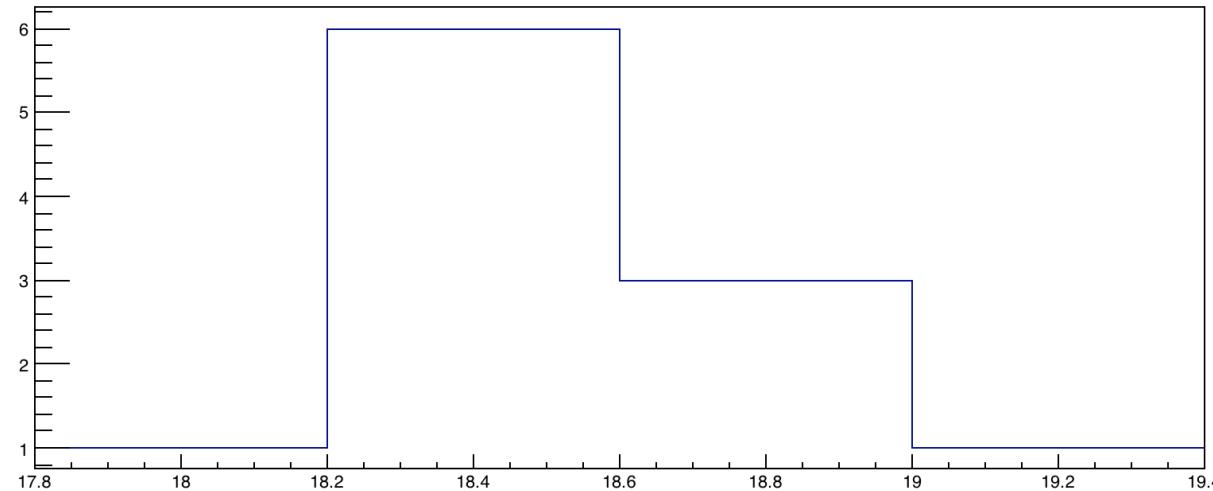
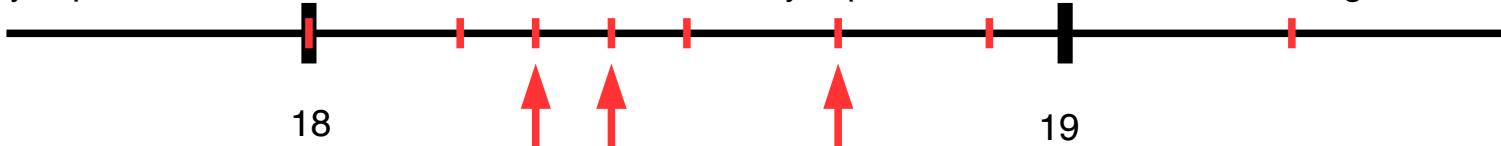
Representación gráfica

Ejemplo: (1.8 en [VAR10]) Datos de una desintegración radiactiva medidas por un contador Geiger en el que se han efectuado medidas cada 30 segundos.



Representación gráfica

Ejemplos de división en marcas de clase en el ejemplo anterior de la muestra criogénica:



Momentos de una distribución de frecuencias

Se define como momento de orden j de una distribución de frecuencias f_i con respecto al punto c como:

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

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.

Veremos a continuación que muchas de las medidas características muestrales se construyen utilizando los momentos de distinto orden de la distribución de frecuencias. En particular, la **media será el momento de primer orden con respecto al origen** y la **varianza el momento central de segundo orden**:

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

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

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 como 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 mas 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 simetría alrededor de las medidas de posición centrales, esto es, si los valores se distribuyen homogeneamente 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 muestrales

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 como 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 mas 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 simetría alrededor de las medidas de posición centrales, esto es, si los valores se distribuyen homogeneamente 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: corresponde al promedio de los valores que constituyen la medida.

En el caso de **N** medidas con resultados { $x_1, x_2, \dots, x_{n-1}, x_n$ } la media será: $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

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

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

Es muy importante señalar que **la media aritmética de una medida aleatoria es a su vez una variable aleatoria**. Además, la media aritmética de una función $g(x)$ definida sobre los datos obtenidos para la variable aleatoria X será a su vez una variable aleatoria:

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

La media corresponde el momento de primer orden de la distribución de frecuencias con respecto al origen, $m_1(0)$.

Medidas características de centralización

Media geométrica: corresponde a la raíz N-esima del productorio de los valores que constituyen la medida. Al ser un productorio pierde su valor (se anula) si un dato es cero.

En el caso de **N** medidas con resultados $\{x_1, x_2, \dots, x_{n-1}, x_n\}$ la media geométrica será:

$$\bar{x}_g = \left(\prod_{i=1}^N x_i \right)^{\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^N x_i}$$

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

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

O tomando logaritmos:

$$\ln \bar{x}_g = \frac{1}{N} \sum_{i=1}^k n_i \ln x_i = \sum_{i=1} f_i \ln x_i$$

Esto establece que el logaritmo neperiano de la media geométrica es la media aritmética de una muestra cuyos datos fueran los logaritmos neperianos de valores distribuidos de la misma forma.

Medidas características de centralización

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

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

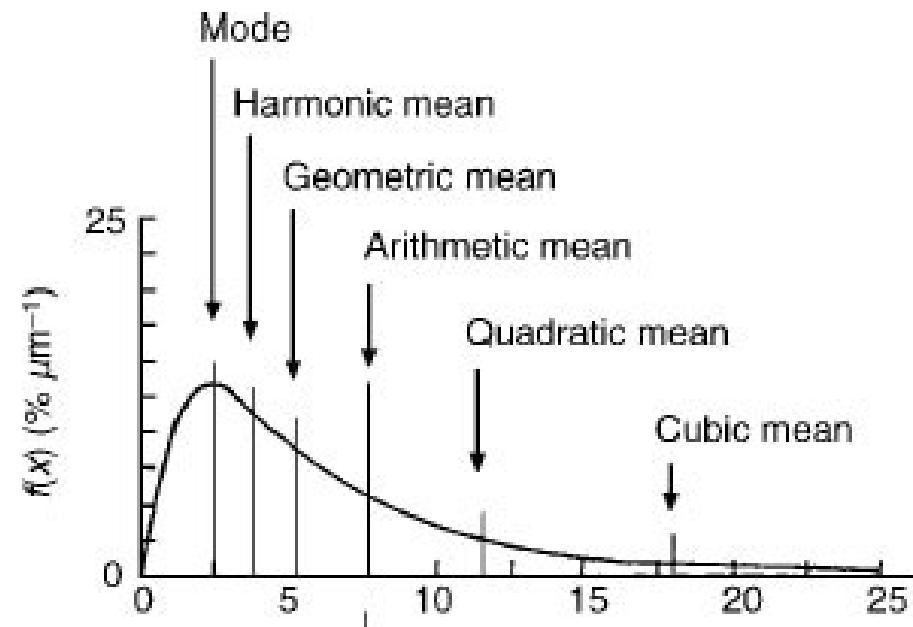
$$\bar{x}_q = \sqrt{\sum_{i=1}^k f_i x_i^2}$$

Media armónica: corresponde al inverso de la suma de los inversos de los valores que constituyen la medida. En el caso de **N** resultados en **k** marcas de clase, cada una con frecuencias relativas f_i , la media armónica es:

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

Se puede demostrar que

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

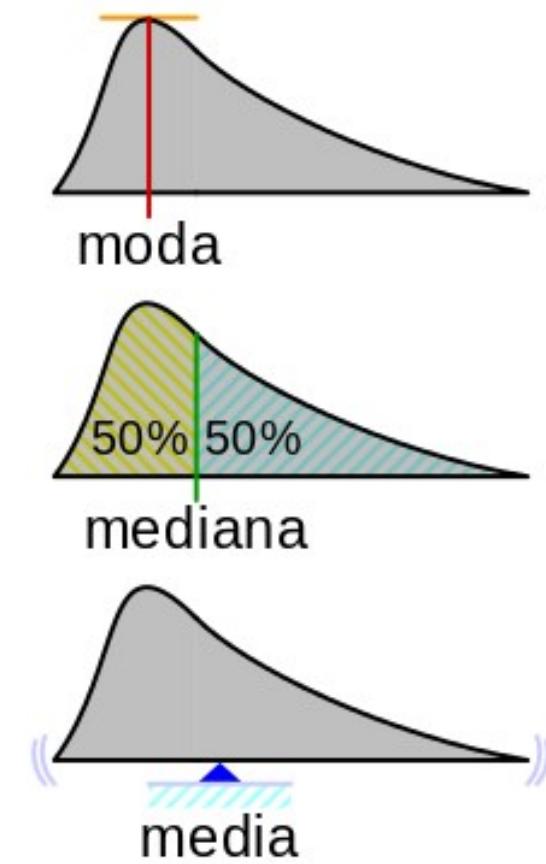


Medidas características de centralización

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

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 veremos a continuación.



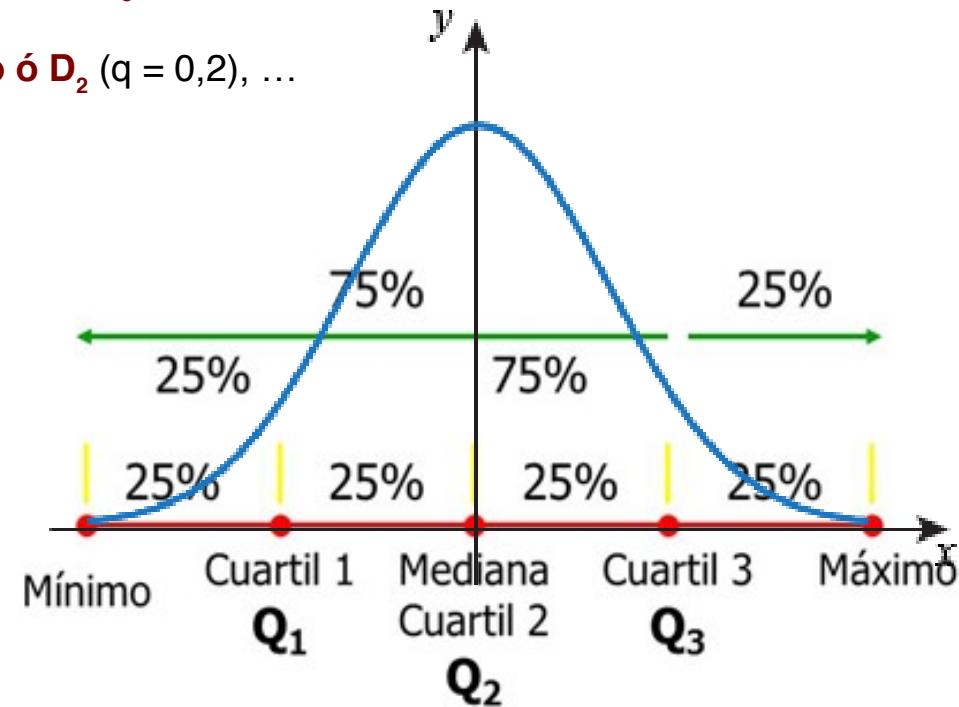
Medidas características de centralización

Percentiles q-ésimos: corresponde al valor de la variable aleatoria escogido tal que la frecuencia relativa de todos los valores menores o iguales sean menores que q ($0 \leq q \leq 1$).

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.

Algunos percentiles tienen nombre propio:

- **Mediana** ($q = 0,5$).
- **Cuartiles primero ó Q_1** ($q = 0,25$) y **tercero ó Q_3** ($q = 0,75$).
- **Deciles primero ó D_1** ($q = 0,1$), **segundo ó D_2** ($q = 0,2$), ...



Medidas características de centralización

La obtención de los percentiles se realiza ordenando los valores en orden creciente y construyendo su tabla de frecuencias absolutas acumuladas. A continuación se busca en que conjunto o valor se encuentra qN con q percentil buscado y N el número de medidas:

- **En el caso de variable discreta**, si qN es entero, y coincide con una de las frecuencias absolutas acumuladas N_i , entonces el valor del percentil será

$$P_q = \frac{x_{i+1} + x_i}{2}$$

En caso de que no sea entero o no corresponda a ningún 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 $P_q = x_{i+1}$

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

De un ejemplo anterior: los cuartiles primero y tercero Q_1 y Q_3 se obtendrán observando los valores de qN en cada caso (75 y 225, respectivamente). Entonces:

$$Q_1 = 4 \text{ (comprende hasta } N_i = 87)$$

$$Q_3 = 7 \text{ (comprende hasta } N_i = 231)$$

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

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

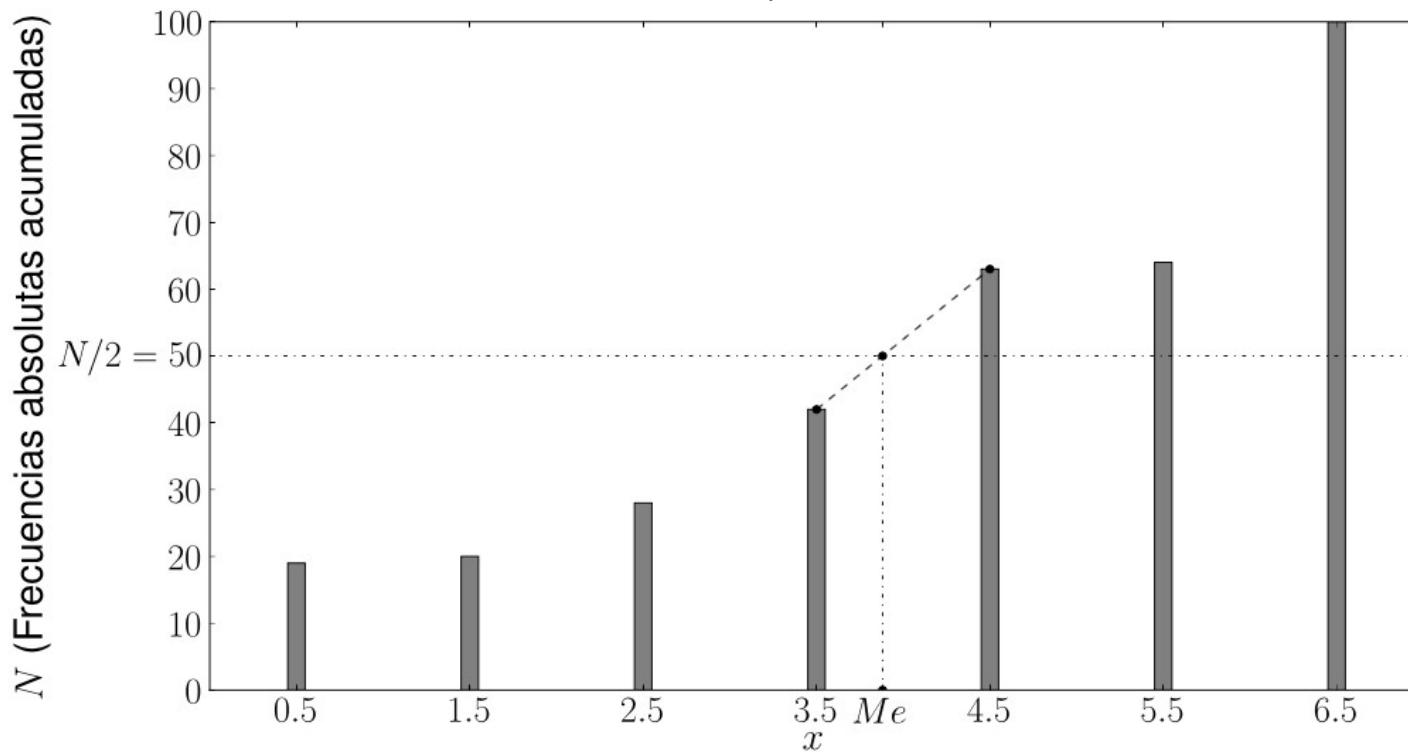
$$\sum_{i=1}^k n_i = 300$$

Medidas características de centralización

- **En el caso de variable continua**, si el valor qN corresponde a una de las frecuencias absolutas acumuladas N_i , entonces el valor del percentil será el extremo superior de la clase correspondiente.

En caso de que no, se interpolará entre las dos marcas de clase x_i y x_{i+1} cuyos N_i dejan en medio el valor buscado:

$$\frac{N_{i+1} - N_i}{x_{i+1} - x_i} = \frac{qN - N_i}{P_q - x_i} \quad 1 \leq i \leq k$$



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

Medidas características muestrales

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 como 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 mas 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 simetría alrededor de las medidas de posición centrales, esto es, si los valores se distribuyen homogeneamente 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 dispersión

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

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** marcas de clase, cada una con frecuencias relativas f_i , (con i moviéndose en este caso de 1 a **k**), entonces:

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

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

$$\begin{aligned} s^2 &= \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \\ &= \sum_{i=1}^k f_i x_i^2 + \sum_{i=1}^k f_i \bar{x}^2 - 2 \sum_{i=1}^k f_i x_i \bar{x} = \bar{x}^2 + \bar{x}^2 - 2\bar{x}\bar{x} = \bar{x}^2 - \bar{x}^2 \end{aligned}$$

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

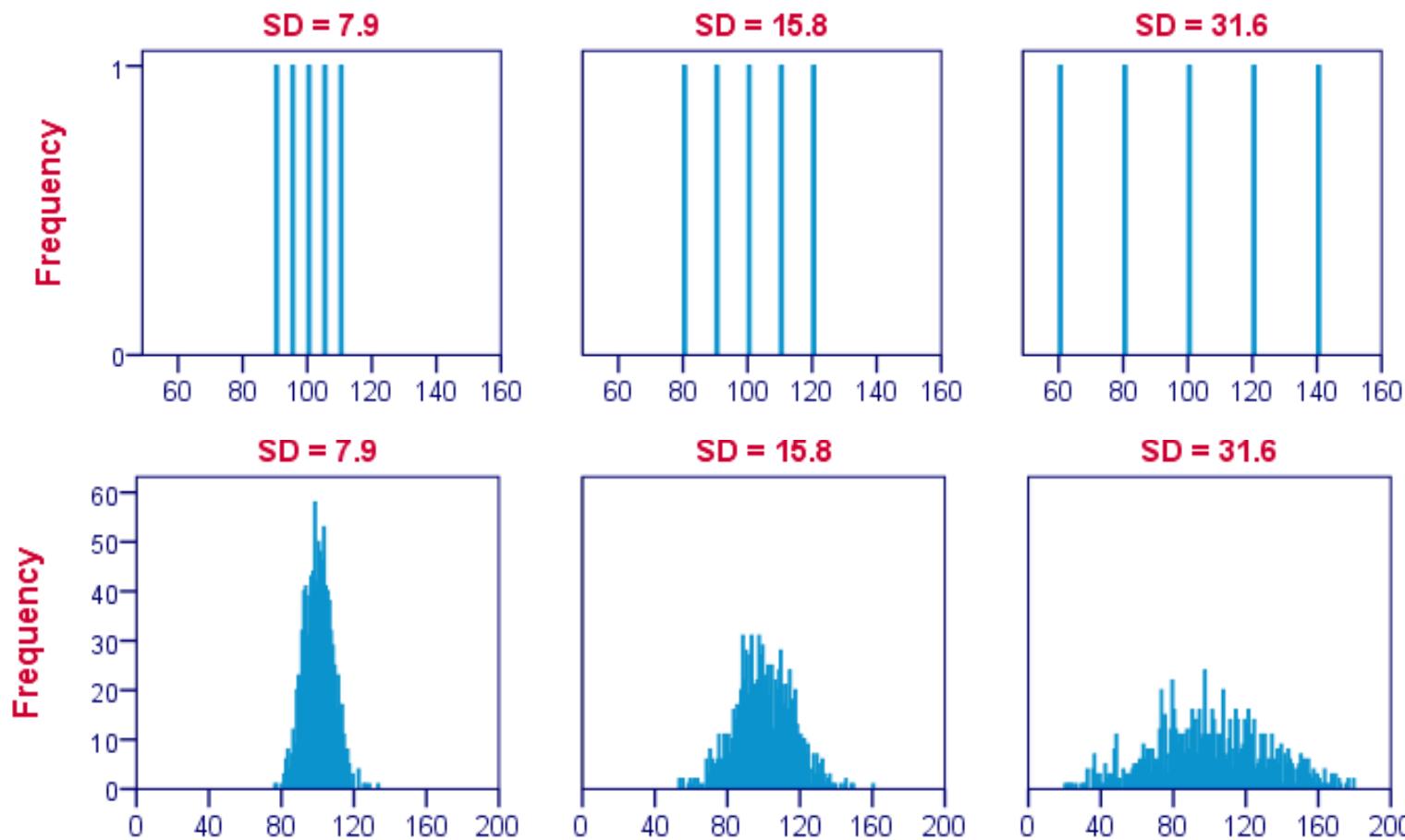
La varianza corresponde al momento central de segundo orden de la distribución de frecuencias.

Medidas características de dispersión

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

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

La varianza es una cantidad definida positiva y dimensionalmente idéntica a la variable aleatoria.



Medidas características de dispersión

Desigualdad de Tchebychev para la desviación típica*.

Establece que la frecuencia de los datos que distan de la media muestral más de $\alpha > 0$ veces la desviación típica está acotada, de forma que

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

Es importante darse cuenta de que **esta desigualdad es aplicable a cualquier distribución de datos de varianza finita y que se puede aplicar aunque desconozcamos la distribución de los datos.**

A partir de la desigualdad, sabremos que:

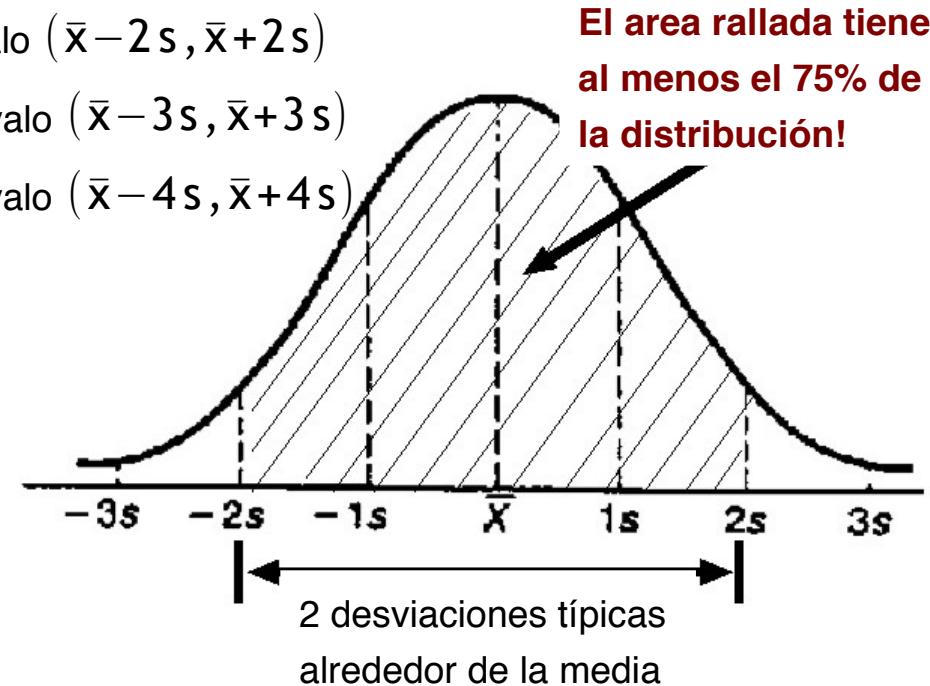
El 75% de los datos estará comprendido en el intervalo $(\bar{x} - 2s, \bar{x} + 2s)$

El ~89% de los datos estará comprendido en el intervalo $(\bar{x} - 3s, \bar{x} + 3s)$

El ~94% de los datos estará comprendido en el intervalo $(\bar{x} - 4s, \bar{x} + 4s)$

Y así sucesivamente...

* La desigualdad de Tchebychev es un resultado algo mas general... Aquí estamos mostrando un resultado particular para el momento central de orden 2 de la distribución de frecuencia (varianza).



Medidas características de dispersión

Demostración de la desigualdad de Tchebychev para la desviación típica.

La desigualdad dice que la frecuencia de los datos que distan de la media muestral más de $\alpha > 0$ veces la desviación típica está acotada, de forma que $f(x_i \mid |x_i - \bar{x}| > \alpha s) < \frac{1}{\alpha^2}$

Denominaremos:

- A1 a los datos muestrales a distancia menor o igual a αs : $A1 = \{x_i \mid |x_i - \bar{x}| \leq \alpha s\}$
- A2 a los datos muestrales a distancia mayor a αs : $A2 = \{x_i \mid |x_i - \bar{x}| > \alpha s\}$

Con esto, por la definición de varianza:

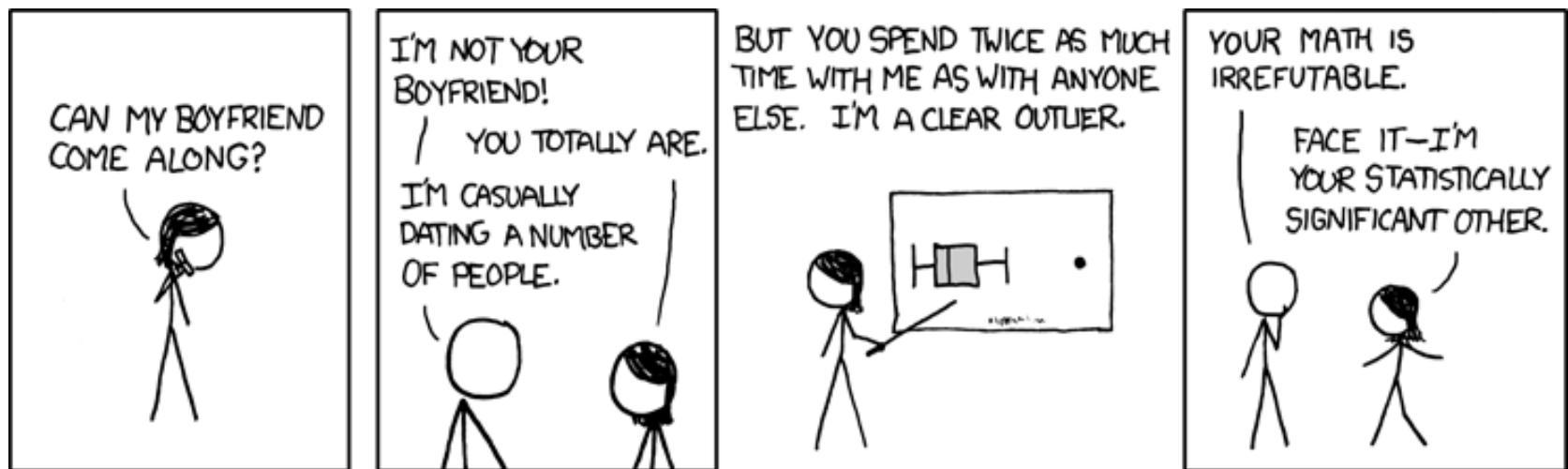
$$\begin{aligned} s^2 &= \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \sum_{x_i \in A1} f_i (x_i - \bar{x})^2 + \sum_{x_i \in A2} f_i (x_i - \bar{x})^2 \\ &\geq \sum_{x_i \in A2} f_i (x_i - \bar{x})^2 > \sum_{x_i \in A2} f_i (\alpha s)^2 = (\alpha s)^2 f(x_i \mid |x_i - \bar{x}| > \alpha s) \end{aligned}$$

Con lo que $s^2 \frac{1}{(\alpha s)^2} > f(x_i \mid |x_i - \bar{x}| > \alpha s)$ y la desigualdad queda demostrada.

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 intervalo $RI = P_{0.75} - P_{0.25}$.
- Se considera un **dato atípico leve** el que aparece a más de $1.5 \cdot RI$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$ y se considera un **dato atípico extremo** el que aparece a más de $3 \cdot RI$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$.

¡Los datos atípicos 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.



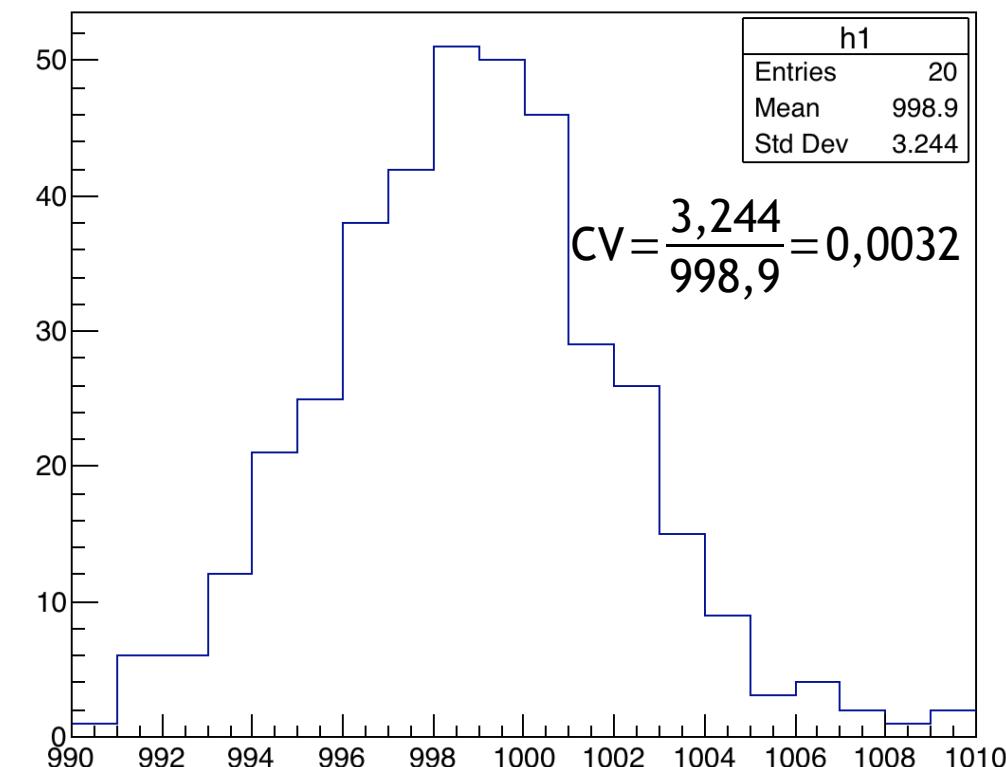
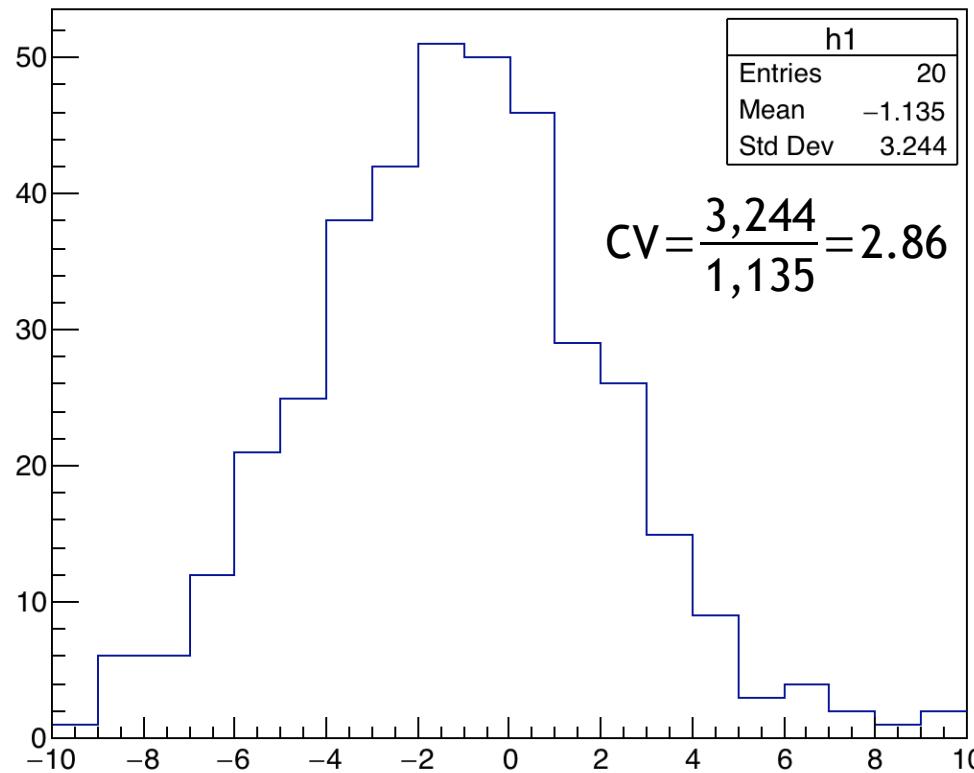
<https://xkcd.com/>

Medidas características de dispersión

Coeficiente de variación de Pearson (CV). Corresponde a una medida relativa de la desviación de la distribución frente a su valor medio, esto es, de la variabilidad 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 introduce la novedad de observar la desviación de la muestra pesada por su valor central, esto es, da una visión de la desviación relativa de la muestra, relacionada con la incertidumbre relativa de un conjunto de medidas.



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): $CV_{MMe} = \frac{DM_{Me}}{|Me|}$

Recorrido de la variable: $R = \max(x) - \min(x)$

Recorrido semi-intercuartílico: $R_{SI} = R_I/2 = (P_{3/4} - P_{1/4})/2$

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 como se distribuyen los resultados de la variable aleatoria. Fundamentalmente son:

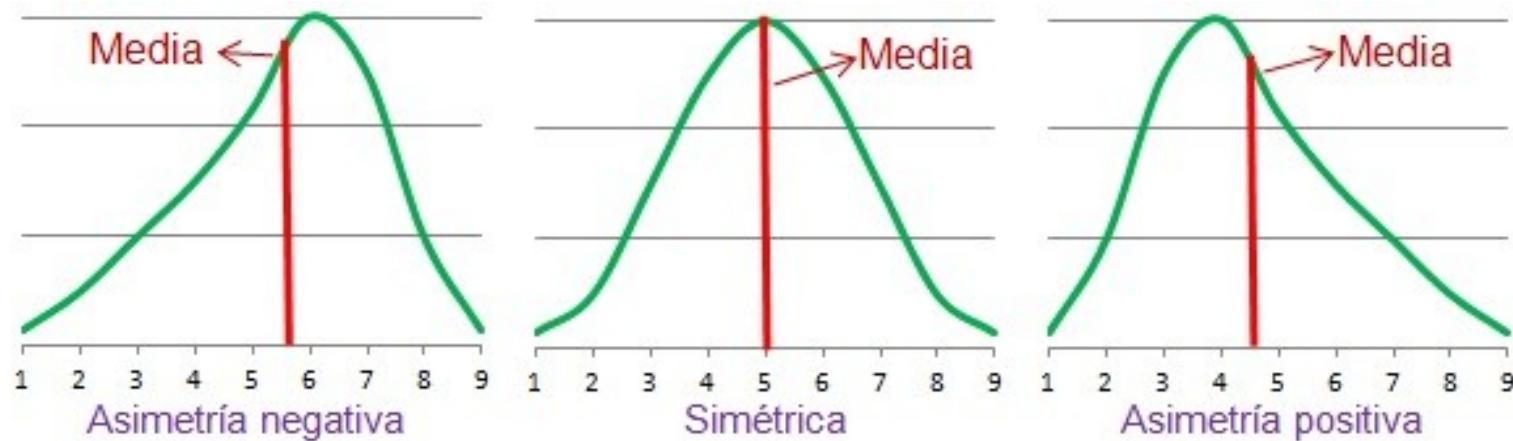
- **Medidas de la posición central de la distribución:** medidas que muestran el valor alrededor del que se centran los resultados, los valores mas probables, los valores centrales de la distribución ... Veremos como ejemplos la media, moda, mediana y los percentiles.
- **Medidas de la dispersión de 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 simetría alrededor de las medidas de posición centrales, esto es, si los valores se distribuyen homogeneamente 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 asimetría

Coeficiente de asimetría de Fisher: definido como $A_F = \frac{1}{s^3} \sum_{i=1}^k f_i(x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{s^3}$

y corresponde al momento central de orden 3 de la distribución de datos.

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: si $A_F = 0$ entonces la distribución puede o no ser simétrica.

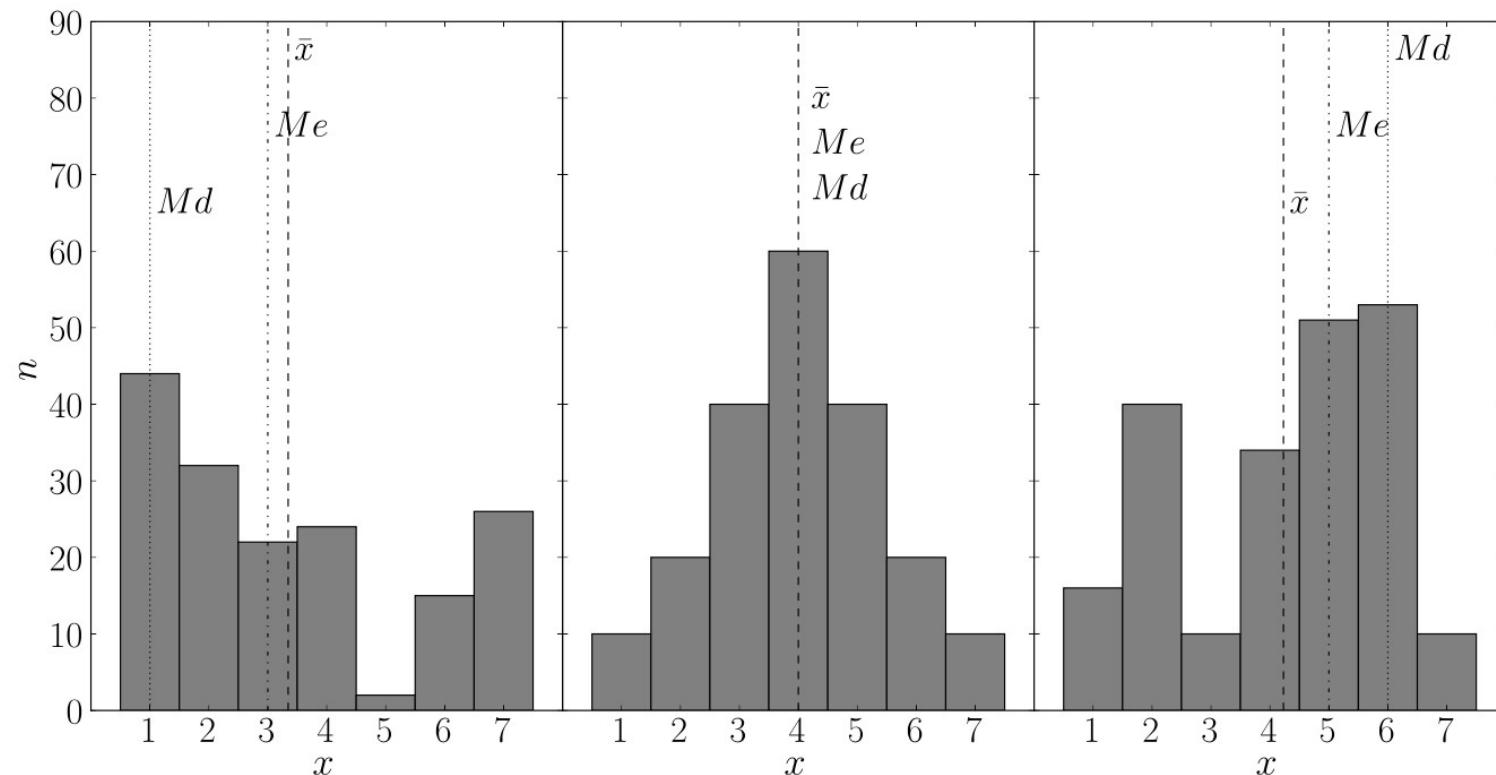
En ocasiones se denomina al coeficiente de asimetría de Fisher como γ_1 .

Medidas características de asimetría

Coeficiente de asimetría de Pearson: definido como $A_p = \frac{\bar{x} - Md}{s}$

Como antes, 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) la moda es mayor que la media y la asimetría se denomina negativa.

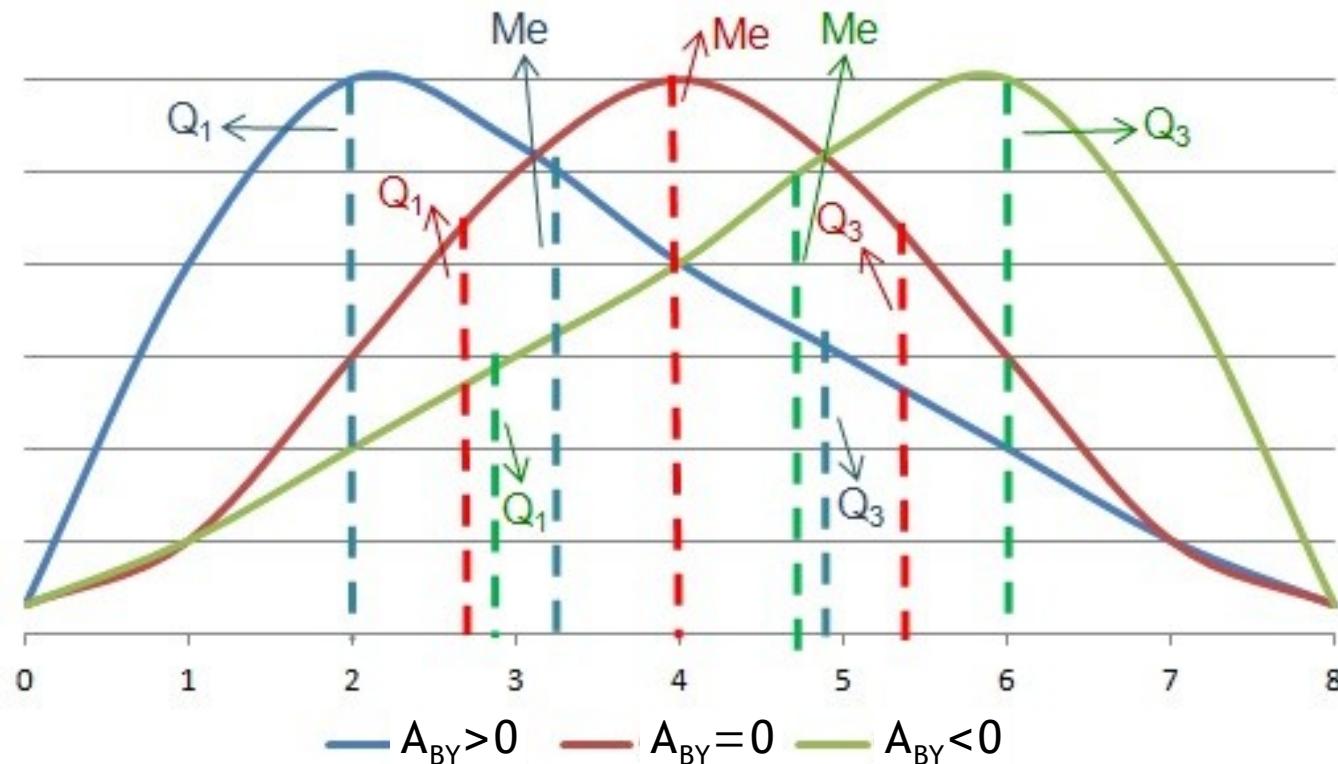
No está definido en caso de distribuciones multimodales.



Medidas características de asimetría

Coeficiente de asimetría de Bowley-Yule: definido como

$$A_{BY} = \frac{(P_{3/4} - Me) - (Me - P_{1/4})}{P_{3/4} - P_{1/4}} = \frac{P_{3/4} + P_{1/4} - 2Me}{P_{3/4} - P_{1/4}} = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$



Medidas características de apuntamiento

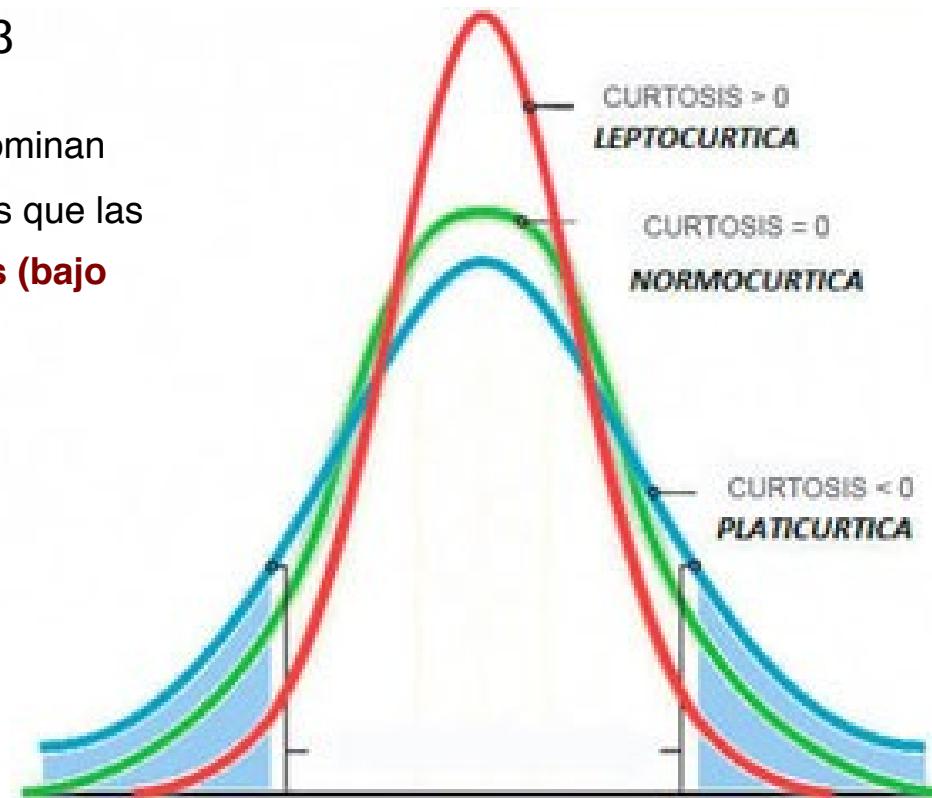
Coeficiente de apuntamiento o de curtosis: definido como

$$g = \beta_2 = \frac{1}{s^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{s^4}$$

Debido a que el coeficiente de curtosis de la distribución normal es $g = 3$, a veces se define como

$$g_2 = \frac{m_4(\bar{x})}{s^4} - 3$$

Las distribuciones con $g > 3$ ($g_2 > 0$) se denominan **leptocúrticas (alto apuntamiento)**, mientras que las de $g < 3$ ($g_2 < 0$) se denominan **platicúrticas (bajo apuntamiento)**.



Transformaciones de variable aleatoria

A partir de las distribuciones de frecuencia de una variable aleatoria se pueden obtener distribuciones de frecuencia derivadas **mediante transformaciones lineales o no lineales**, obteniendo nuevas distribuciones de variables aleatorias. Encontraremos transformaciones cuando:

- Operamos sobre el espacio de resultados, con transformaciones tanto lineales (cambios de escala, transformación de unidades, ...) como no lineales (transformaciones logarítmicas).
- Obtenemos observables que combinan diferentes medidas de variables aleatorias (como en el cálculo de la velocidad a través de las medidas de tiempo y distancia).
- Cuando cambiamos el sistema de coordenadas en el que se obtienen los resultados.
- En general, cuando realizamos operaciones sobre las variables resultantes de una medición para obtener cantidades derivadas.

Transformaciones lineales de variable aleatoria

Transformaciones lineales: dada una variable aleatoria X , son transformaciones del tipo $Y = a + bX$, de forma que, para una muestra $\{x_1, x_2, \dots, x_n\}$, se obtienen los valores tras la transformación lineal $\{y_1, y_2, \dots, y_n\} = \{a + bx_1, a + bx_2, \dots, a + bx_n\}$ en la nueva variable aleatoria Y . Las medidas características tras una transformación lineal son:

- La **media aritmética** de la nueva variable aleatoria Y tras una transformación lineal será:

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

- La **varianza** de la nueva variable aleatoria Y tras una transformación lineal será:

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

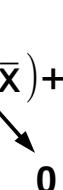
- Los **percentiles** se conservan, esto es, se transforman de la misma forma que la media, en las transformaciones lineales: por ejemplo, $Q_1(Y) = a + b Q_1(X)$

Transformaciones no lineales: son transformaciones que modifican la distribución de frecuencias a través de una función no lineal, como puede ser la función potencial $Y = X^p$, logarítmica $Y = \ln X$, ... o cualquier combinación de estas. En general no se pueden deducir las medidas características tras una transformación no lineal, aunque hay algunas propiedades que se pueden estudiar:

- La **media aritmética** de la nueva variable aleatoria Y no verifica en general $\bar{y} = h(\bar{x})$ como en el caso lineal. En general:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N h(x_i) = \sum_{i=1}^k f_i h(x_i)$$

Utilizando un desarrollo de Taylor alrededor de la media se podría expandir:

$$\begin{aligned} \bar{y} &\approx \frac{1}{N} \sum_{i=1}^N [h(\bar{x}) + h'(\bar{x})(x_i - \bar{x}) + \frac{1}{2!} h''(\bar{x})(x_i - \bar{x})^2] \\ &= h(\bar{x}) + \frac{h'(\bar{x})}{N} \sum_{i=1}^N (x_i - \bar{x}) + \frac{h''(\bar{x})}{2N} \sum_{i=1}^N (x_i - \bar{x})^2 \approx h(\bar{x}) + \frac{1}{2} h''(\bar{x}) s_x^2 \end{aligned}$$


Lo que nos permite calcular la media con una aproximación a primer orden. Si $h''(\bar{x}) s_x^2 \rightarrow 0$ la media aritmética de la variable transformada es aproximadamente la transformación aplicada a la media de X .

- La **varianza** en una transformación no lineal se puede aproximar utilizando el desarrollo de Taylor alrededor de la media:

$$\begin{aligned}s_y^2 &= \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \approx \frac{1}{N} \sum_{i=1}^N \left[h(\bar{x}) + h'(\bar{x})(x_i - \bar{x}) + \frac{1}{2!} h''(\bar{x})(x_i - \bar{x})^2 - \left[h(\bar{x}) + \frac{1}{2!} h''(\bar{x})s_x^2 \right] \right]^2 \\ &\approx \frac{1}{N} \sum_{i=1}^N [h'(\bar{x})(x_i - \bar{x})]^2 = [h'(\bar{x})]^2 s_x^2\end{aligned}$$

En este caso la varianza tras la transformación no lineal presenta un factor $h'(\bar{x})$ que cambia su valor.

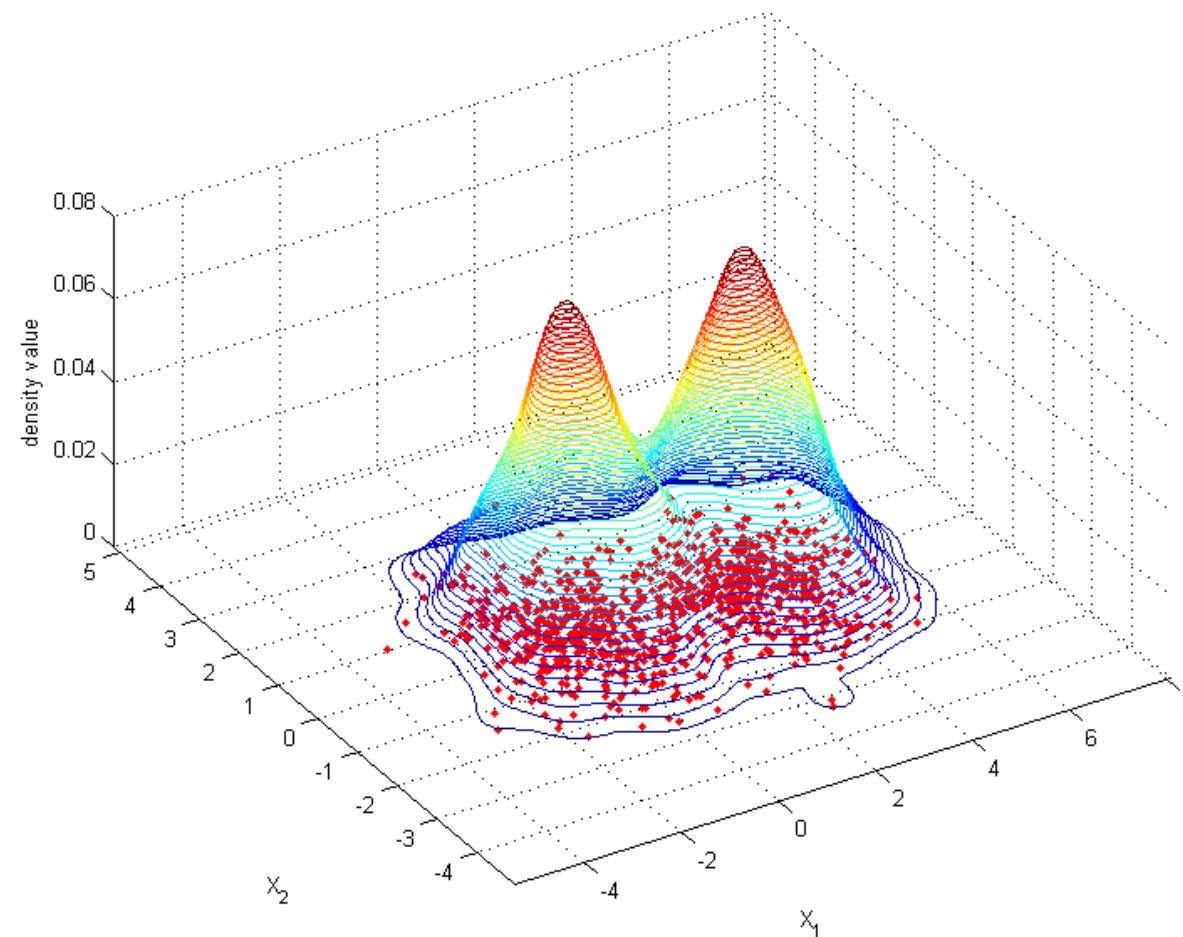
- Los **percentiles**, al tener solo en cuenta el orden de los datos, mantendrá sus valores para ciertas funciones. En particular, las transformaciones monótonas crecientes cumplirán:

$$x_1 < x_2 \Leftrightarrow h(x_1) < h(x_2)$$

Lo que transforma sus percentiles para un valor q dado de acuerdo a: $P_q(y) = h[P_q(x)]$

Distribuciones de frecuencia multivariantes

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



Distribuciones de frecuencia multivariantes

Si consideramos dos variables estadísticas **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 proporcionará 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:

$Y \ X$	y_1	y_2	\dots	y_j	\dots	y_l	
x_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1l}	$f_{ij} = \frac{n_{ij}}{N}$
x_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2l}	$\sum_{i=1}^k \sum_{j=1}^l n_{ij} = N$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{il}	$\sum_{i=1}^k \sum_{j=1}^l f_{ij} = 1$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	
x_k	n_{k1}	n_{k2}	\dots	n_{kj}	\dots	n_{kl}	

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

Distribuciones marginales de frecuencia multivariantes

Las **distribución de frecuencia marginal** para una variable aleatoria **X** dentro de una distribución multivariante se corresponde con la distribución de frecuencias asociada a los posibles valores de la variable **X** ($\{x_1, x_2, \dots, x_k\}$) independientemente del valor que tomen el resto de variables aleatorias.

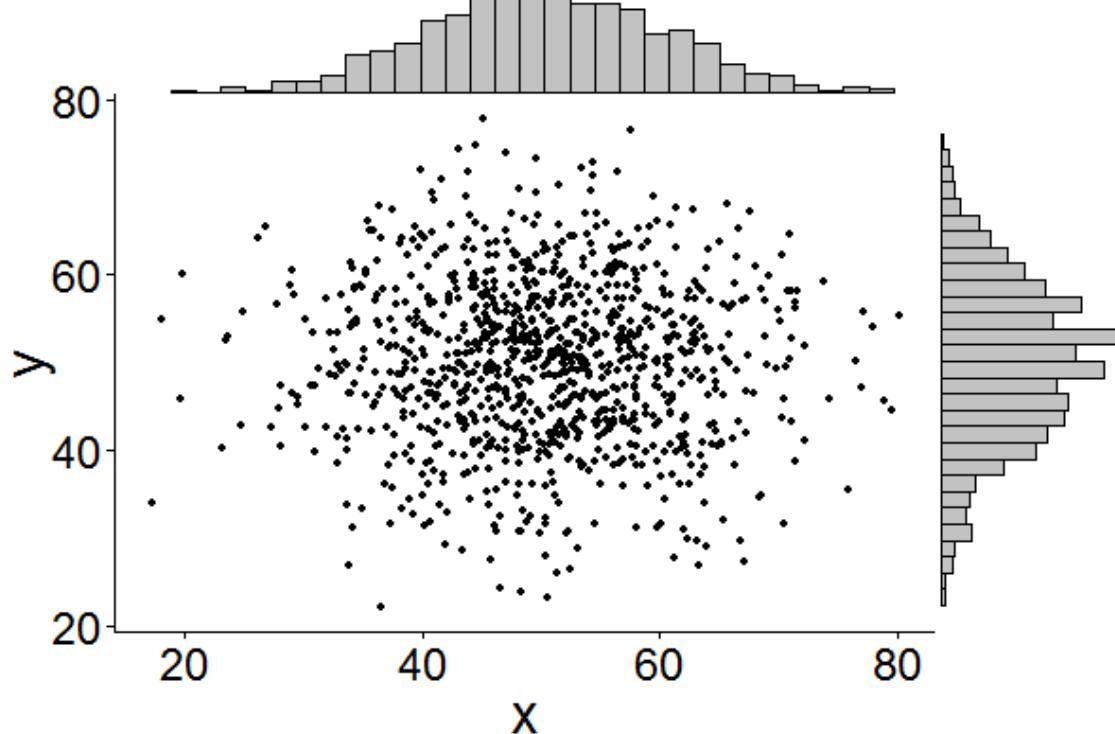
Así, la **distribución de frecuencia absoluta marginal** para la variable **X** de la distribución bivariante **X, Y**, con valores posibles $\{x_1, x_2, \dots, x_k\}$ y $\{y_1, y_2, \dots, y_l\}$, será la serie de valores $\{n_{x_i}\}_{i=1}^k$ definidos como:

$$n_{x_i} = \sum_{j=1}^l n_{ij} = n_{i1} + n_{i2} + \dots + n_{il}$$

Y análogamente para la variable **Y**,

la serie $\{n_{y_j}\}_{j=1}^l$ donde:

$$n_{y_j} = \sum_{i=1}^k n_{ij} = n_{1j} + n_{2j} + \dots + n_{kj}$$



Las **distribuciones de frecuencias relativas marginales** se definirán como las series $\{f_{x_i}\}_{i=1}^k$ y

$\{f_{y_j}\}_{j=1}^l$ donde:

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

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

con las propiedades:

$$\sum_{i=1}^k n_{x_i} = N$$

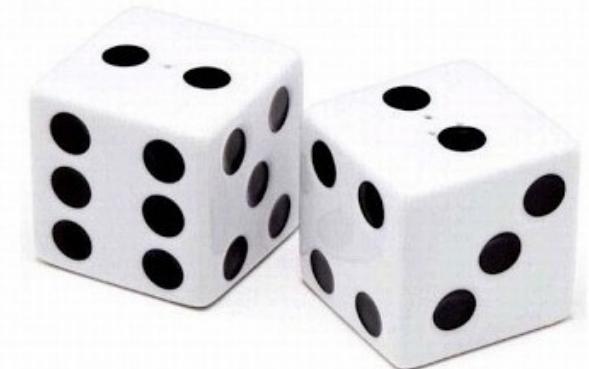
$$\sum_{j=1}^l n_{y_j} = N$$

$$\sum_{i=1}^k f_{x_i} = 1$$

$$\sum_{j=1}^l f_{y_j} = 1$$

Distribuciones de frecuencia multivariantes

Ejemplo de distribución multivariante (lanzamiento de dos dados):



D1\D2	1	2	3	4	5	6
1	25	19	33	32	35	27
2	22	30	30	19	27	22
3	30	27	23	37	23	31
4	22	24	39	27	24	49
5	32	28	25	15	14	30
6	42	43	12	34	28	20

$$n_{D1=1} = 171$$

$$n_{D1=2} = 150$$

$$n_{D1=3} = 171$$

$$n_{D1=4} = 185$$

$$n_{D1=5} = 144$$

$$n_{D1=6} = 179$$

$$n_{D2=1} = 173$$

$$n_{D2=3} = 162$$

$$n_{D2=5} = 151$$

$$n_{D2=2} = 171$$

$$n_{D2=4} = 164$$

$$n_{D2=6} = 179$$

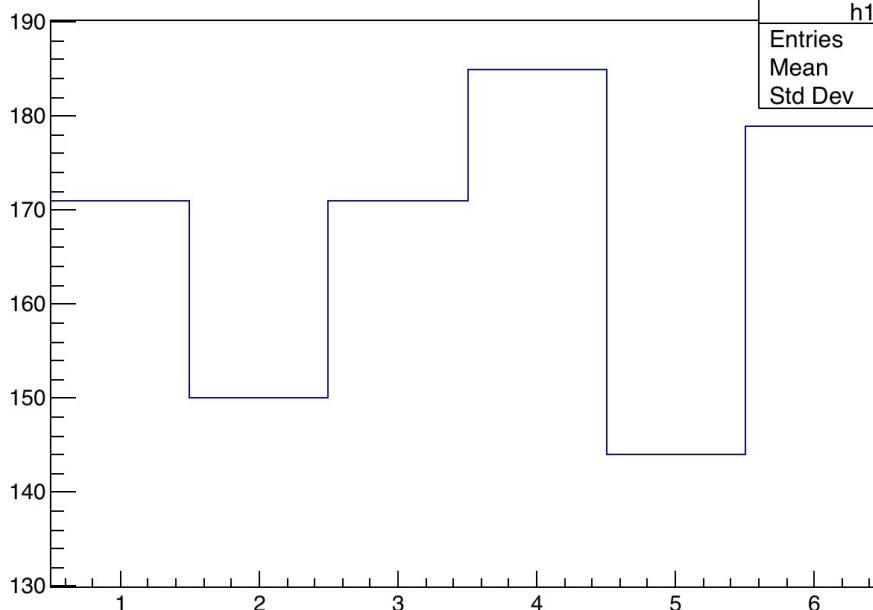
Distribuciones de frecuencia multivariantes

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

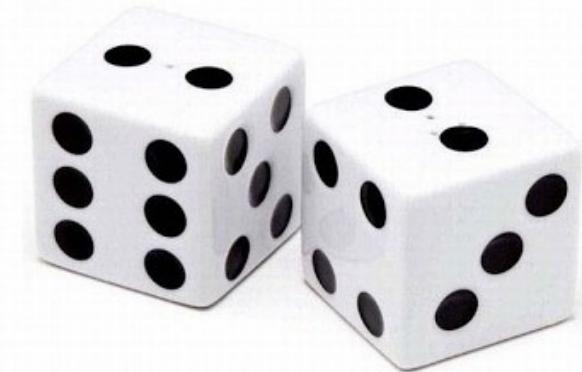
$$n_{D2=2} = 171 \quad n_{D2=4} = 164 \quad n_{D2=6} = 179$$

1\2	1	2	3	4	5	6
1	25	19	33	32	35	27
2	22	30	30	19	27	22
3	30	27	23	37	23	31
4	22	24	39	27	24	49
5	32	28	25	15	14	30
6	42	43	12	34	28	20

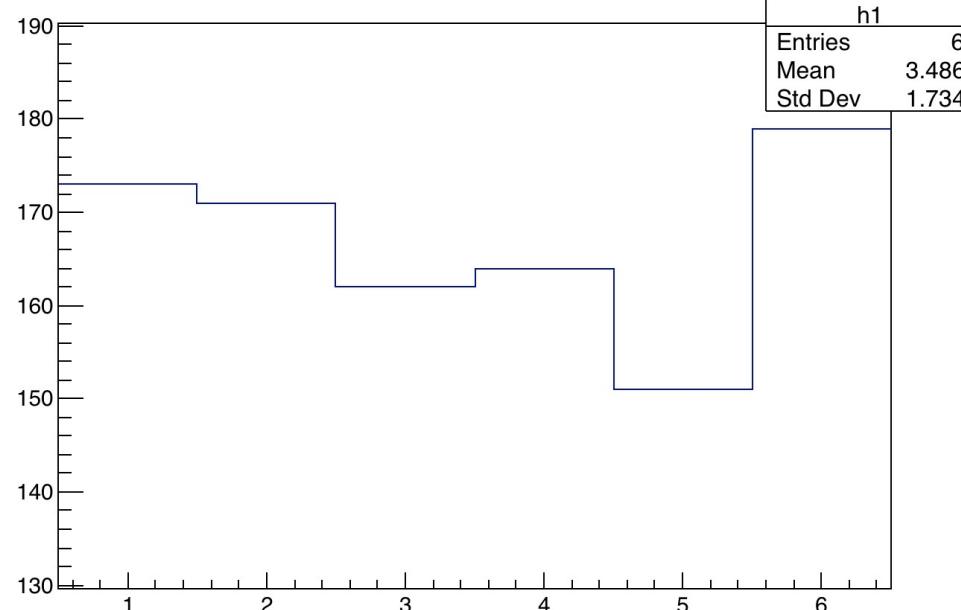
Distribución marginal para el dado 1



$$\begin{aligned} n_{D1=1} &= 171 \\ n_{D1=2} &= 150 \\ n_{D1=3} &= 171 \\ n_{D1=4} &= 185 \\ n_{D1=5} &= 144 \\ n_{D1=6} &= 179 \end{aligned}$$



Distribución marginal para el dado 2



Distribuciones de frecuencia condicionadas

Para una distribución multivariante, se define la **distribuciones de frecuencia absoluta condicionada a $Y = y_j$ para la variable X** como la serie de valores $\{n(x_i \mid Y=y_j)\}_{i=1}^k$ obtenidos observando las frecuencias de cada valor x_i en la variable X condicionadas a que aparezca el valor y_j en la variable Y , o sea, tal que:

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

De igual forma, la frecuencia absoluta con la que aparece el valor y_j de la variable Y condicionada a observar simultáneamente el valor x_i en la variable X será $n(y_j \mid X=x_i) = n(y_j \mid x_i) = n_{ij}$

Lo que resulta importante es que las **distribuciones de frecuencia relativa condicionada**, definidas como:

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

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

corresponden a las frecuencias relativas de aparición de $X = x_i$ **restringidas únicamente al conjunto donde $Y = y_j$** , y no respecto al total de las observaciones, lo que hace que $f(x_i \mid y_j) \neq f(y_j \mid x_i)$

Distribuciones de frecuencia condicionadas

Distribuciones de frecuencia

absoluta condicionada

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

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

$X \setminus Y$	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}

Distribución de frecuencia
relativa marginal para y_j :

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

Distribuciones de frecuencia relativa condicionada

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

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

Distribución de frecuencia
relativa marginal para x_i :

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

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

Distribuciones de frecuencia multivariantes

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

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

1\2	1	2	3	4	5	6
1	25	19	33	32	35	27
2	22	30	30	19	27	22
3	30	27	23	37	23	31
4	22	24	39	27	24	49
5	32	28	25	15	14	30
6	42	43	12	34	28	20

$$n_{D1=1} = 171$$

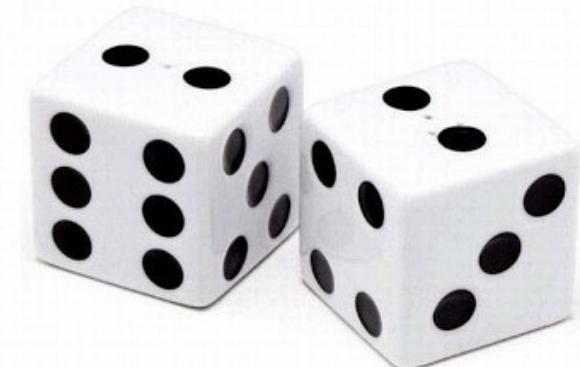
$$n_{D1=2} = 150$$

$$n_{D1=3} = 171$$

$$n_{D1=4} = 185$$

$$n_{D1=5} = 144$$

$$n_{D1=6} = 179$$



La distribución de frecuencia relativa condicionada de obtener D1 = 2, condicionada a que D2 = 5 es:

$$f(D1=2 \mid D2=5) = \frac{n(D1=2 \mid D2=5)}{n_{D2=5}} = \frac{n_{25}}{n_{D2=5}} = \frac{f_{25}}{f_{D2=5}} = \frac{27}{151} = 0,1788$$

Distribuciones de frecuencia condicionadas

Se pueden verificar las siguientes propiedades:

$$\sum_{i=1}^k f(x_i \mid y_j) = \frac{\sum_{i=1}^k n(x_i \mid Y=y_j)}{n_{y_j}} = 1$$

$$\sum_{j=1}^l f(y_j \mid x_i) = \frac{\sum_{j=1}^l n(y_j \mid X=x_i)}{n_{x_i}} = 1$$

Y también las siguientes relaciones que pueden ser útiles en desarrollos posteriores:

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

Con esto, las frecuencias relativas condicionadas se pueden escribir como:

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

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

Que resulta una expresión útil para caracterizar los subconjuntos de la muestra.

Independencia estadística para variables aleatorias

Dos variables aleatorias X e Y se consideran **estadísticamente independientes entre si** cuando la distribución de frecuencias de cada uno de los posibles valores x_i de la variable X condicionado a que Y tome el valor y_j es igual a la distribución marginal para x_i , esto es: $f(x_i | Y=y_j) = f_{x_i}$

Se deduce un resultado muy importante de la definición: si X e Y son variables independientes, entonces $f(x_i | Y=y_j) = f_{x_i}$ y $f(y_j | X=x_i) = f_{y_j}$ con lo que $f_{ij} = f_{x_i}f_{y_j}$

Para variables 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 .**

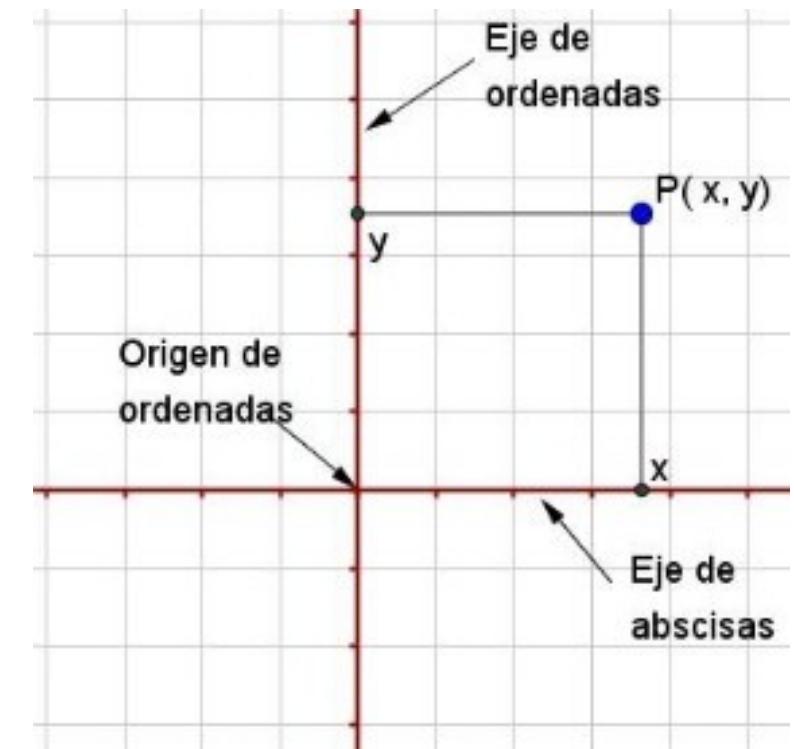
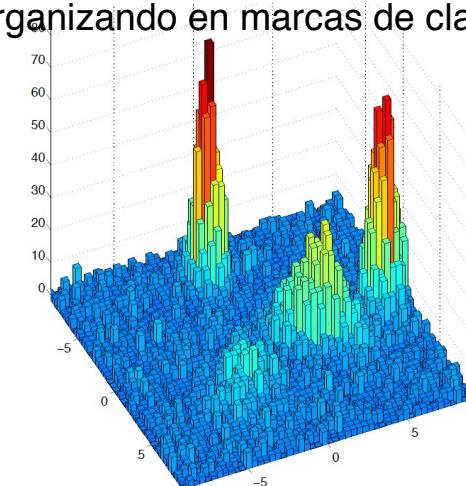
Demostrarlo es simple a partir de los resultados (ver página anterior):

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

Representación gráfica multivariantes

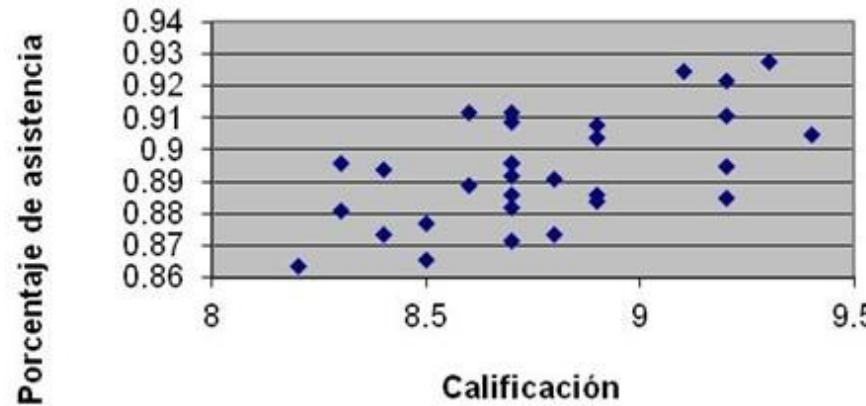
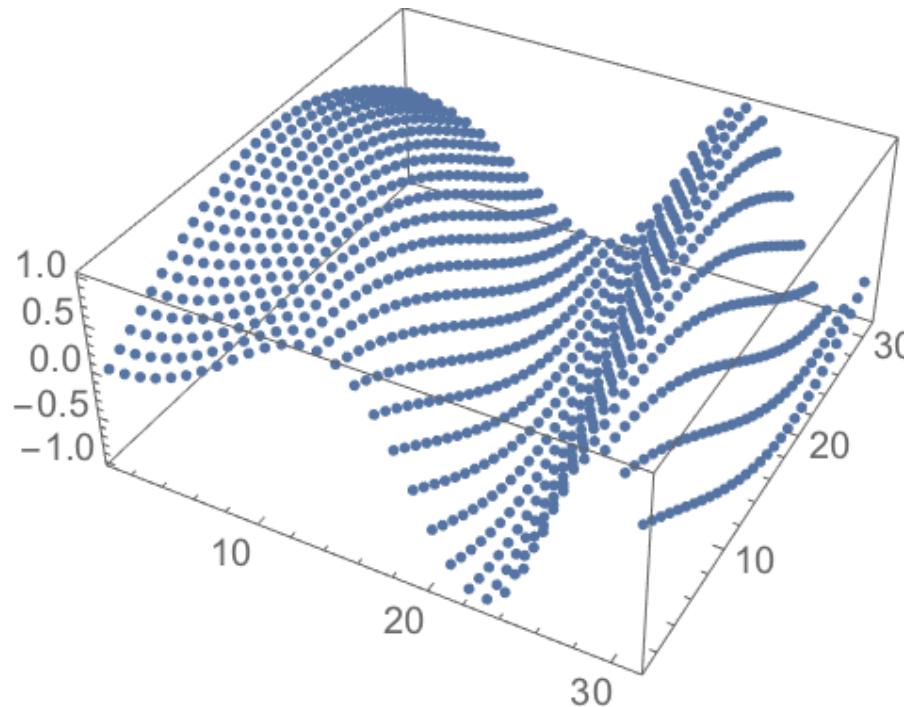
En el caso multivariante, las representaciones gráficas de las distribuciones de frecuencia más importantes son:

- **Diagramas de dispersión (scatter plot)**: representación sobre ejes coordinados de un conjunto de datos $\{(x_i, y_i)\}_{i=1}^N$. Se denomina abcisa a los valores $\{x_i\}_{i=1}^k$ y ordenada a $\{y_i\}_{i=1}^k$.
- **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_j) de las variables aleatorias **X** e **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 o organizando en marcas de clase las variables continuas, ...

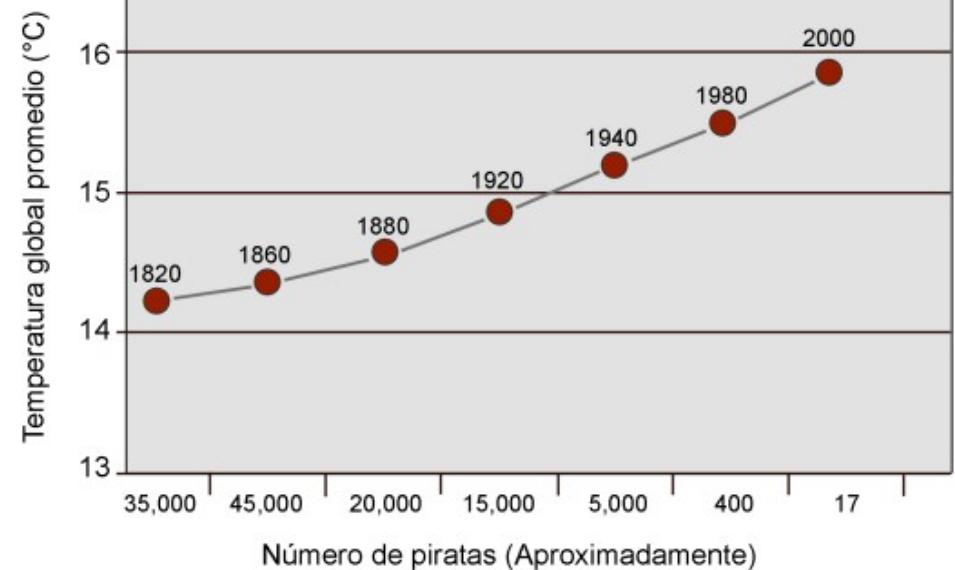
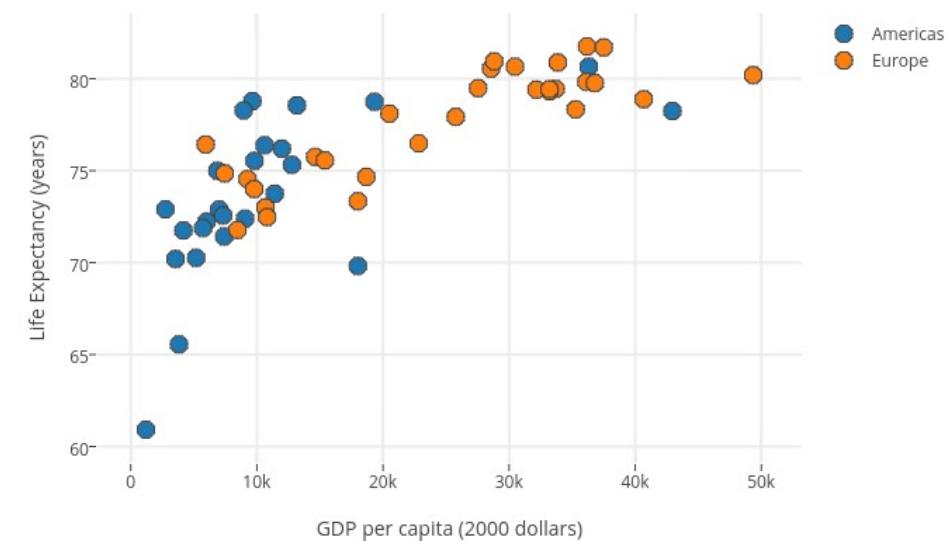


Representación gráfica multivariantes

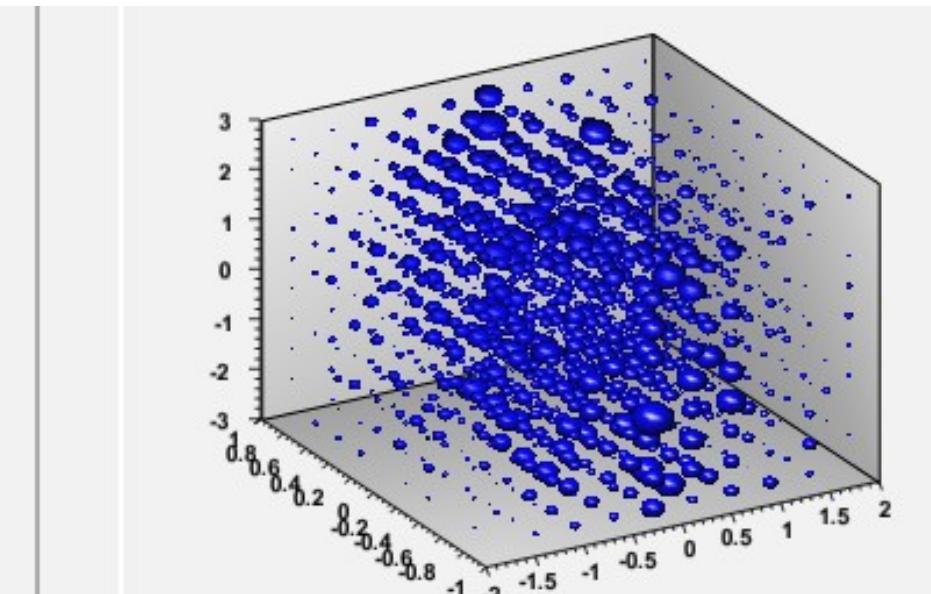
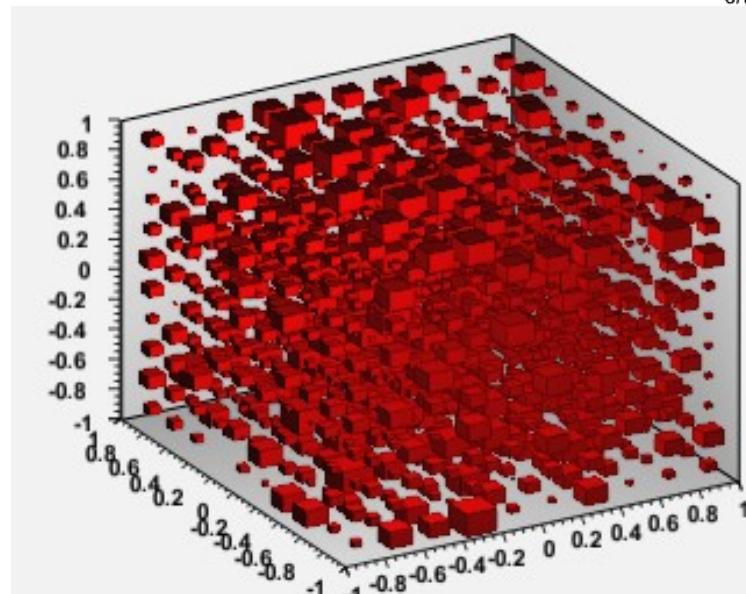
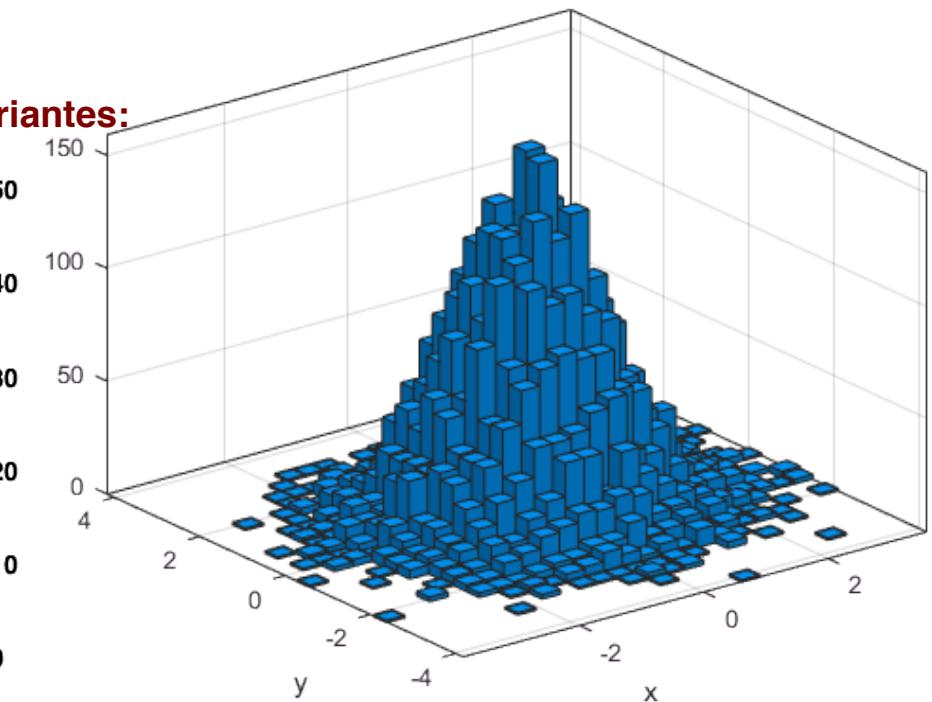
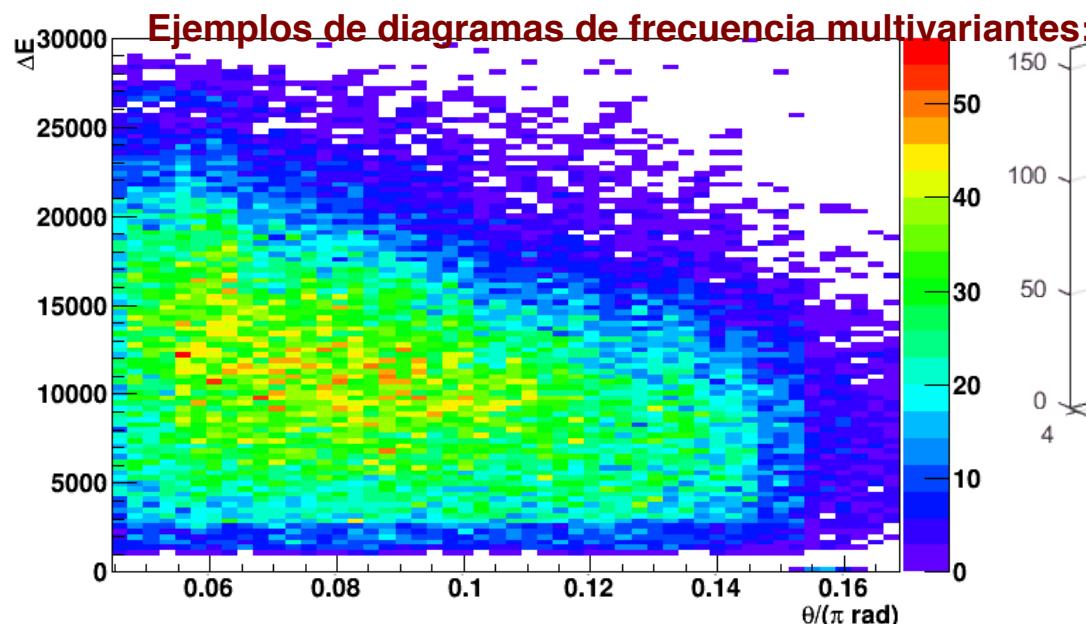
Ejemplos de diagramas de dispersión:



Life Expectancy v. Per Capita GDP, 2007



Representación gráfica multivariantes



Para una muestra $\{(x_i, y_i)\}_{i=1}^N$ discreta o agrupada en clases, asociada a una variable estadística bidimensional (X, Y) , con posibles valores $\{x_1, x_2, \dots, x_k\}$ y $\{y_1, y_2, \dots, y_l\}$, se define como **momento de ordenes r y s respecto al punto (c,d) como:**

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

Donde f_{ij} es la frecuencia del par (x_i, y_j) .

Los momentos respecto al punto $(c, d) = (0,0)$ se denominan **momentos respecto al origen**, $m_{r,s}(0,0)$, y cuando $(c, d) = (\bar{x}, \bar{y})$ **momentos centrales o momentos respecto a la media**.

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

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

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

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

Donde f_{ij} es la frecuencia del par (x_i, y_j) .

Se puede observar también, utilizando de las distribuciones marginales de frecuencia relativa, que:

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

(medias a partir de las distribuciones marginales, o en alguna bibliografía **medias marginales**).

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

**Distribución de frecuencia
relativa marginal para x_i :**

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

La **varianza de cada variable estadística** que compone la **variable bidimensional (X,Y)**, se define como el **momento central de segundo orden en cada variable**:

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

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

Donde f_{ij} es la frecuencia del par (x_i, y_j) .

Se puede observar también, utilizando de las distribuciones marginales de frecuencia relativa, que:

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

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

(varianzas a partir de las distribuciones marginales, o en alguna bibliografía **varianzas marginales**).

La **covarianza del par de variables estadísticas (X, Y)**, se define como el **momento central de primer orden en las dos variables de la distribución conjunta de frecuencias**:

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

Donde f_{ij} es la frecuencia del par (x_i, y_j) .

Desarrollando el producto llegamos a una formulación diferente y muy usada de la covarianza:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^l f_{ij}(x_i - \bar{x})(y_j - \bar{y}) &= \sum_{i=1}^k \sum_{j=1}^l f_{ij} x_i y_j - \bar{x} \sum_{i=1}^k \sum_{j=1}^l f_{ij} y_j - \bar{y} \sum_{i=1}^k \sum_{j=1}^l f_{ij} x_i + \bar{x} \bar{y} \sum_{i=1}^k \sum_{j=1}^l f_{ij} \\ &= \sum_{i=1}^k \sum_{j=1}^l f_{ij} x_i y_j - \bar{x} \bar{y} = \bar{xy} - \bar{x} \bar{y} \end{aligned}$$

Donde $\bar{xy} = \sum_{i=1}^k \sum_{j=1}^l f_{ij} x_i y_j$ es el valor medio del productorio, de igual forma que, de modo general:

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

Es el valor medio de $g(x, y)$ sobre la distribución de frecuencias de las variables estadísticas (X, Y) .

Propiedades de la covarianza del par de variables estadísticas (X, Y):

- Simetría: $\text{cov}(x, y) = \text{cov}(y, x)$
- Si las variables estadísticas del par (X, Y) son independientes, $\text{cov}(x, y) = 0$

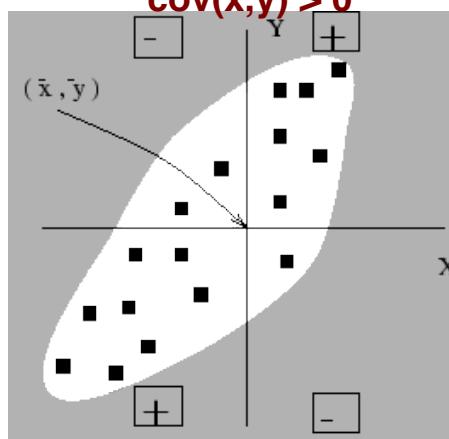
En este caso $\text{cov}(x, y) = \sum_{i=1}^k \sum_{j=1}^l f_{ij} x_i y_j - \bar{x}\bar{y} = \sum_{i=1}^k \sum_{j=1}^l f_{x_i} f_{y_j} x_i y_j - \bar{x}\bar{y} = \sum_{i=1}^k f_{x_i} x_i \sum_{j=1}^l f_{y_j} y_j - \bar{x}\bar{y} = 0$

¡Esto no significa que si $\text{cov}(x, y) = 0$ las variables son independientes!

- La covarianza de una variable consigo misma es igual a la varianza de la variable: $\text{cov}(x, x) = s_x^2$
- En transformaciones lineales se tiene: $\text{cov}(a+bx, c+dy) = bd \text{cov}(x, y)$
- La covarianza entre funciones informa de la correlación entre variables estadísticas:

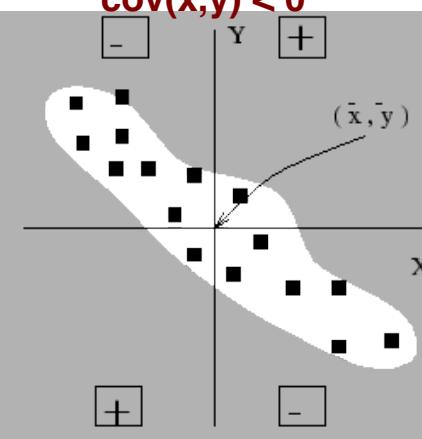
Correlación positiva

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



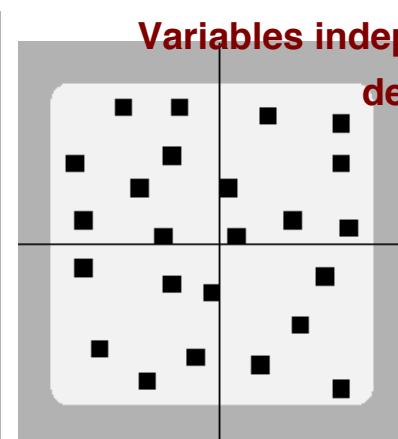
Correlación negativa

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

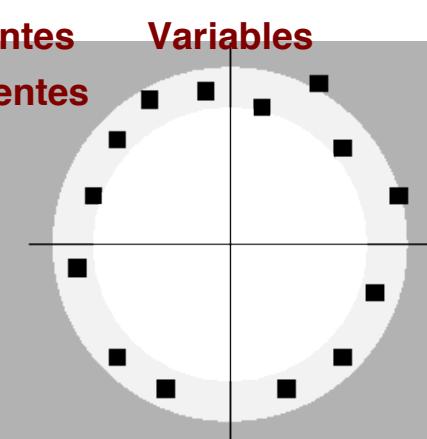


Ausencia de correlación $\text{cov}(x, y) = 0$

Variables independientes

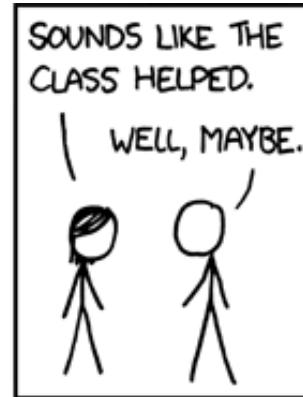
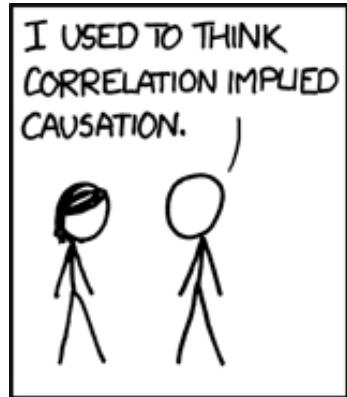


Variables dependientes

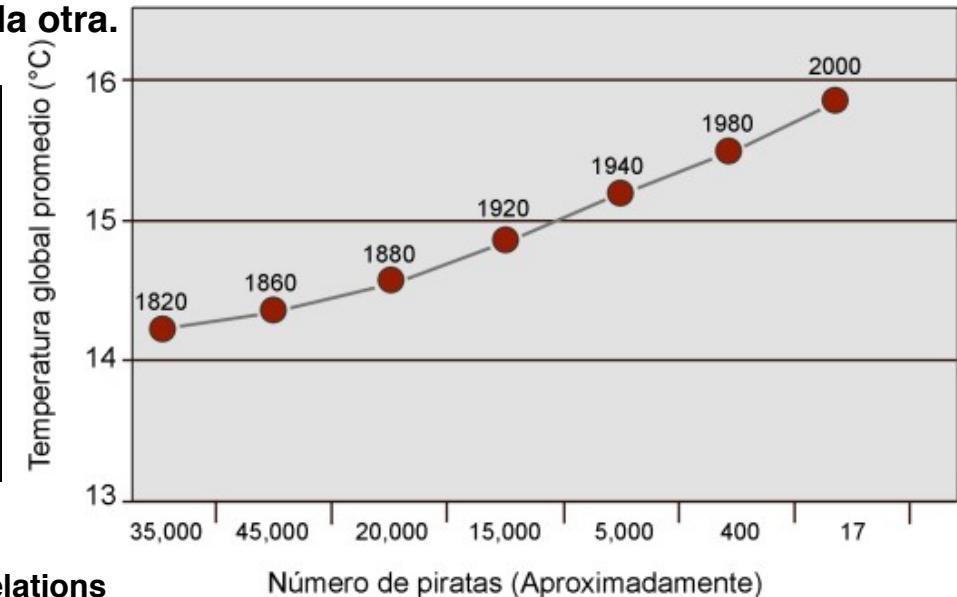


Medidas características muestrales multivariantes

Atención: correlación no implica causalidad. El hecho de que exista una correlación entre variables no significa que una sea causa de la otra.



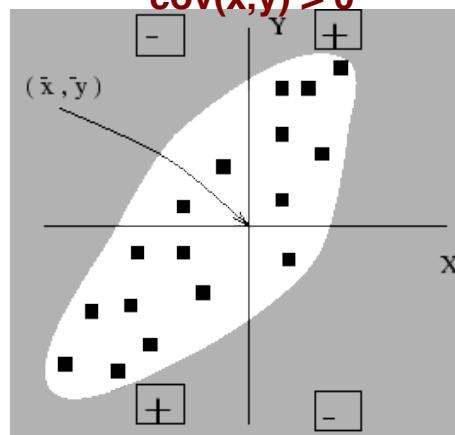
<https://xkcd.com/>



Mas ejemplos en <http://www.tylervigen.com/spurious-correlations>

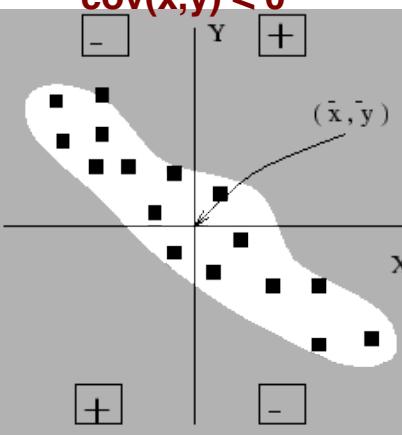
Correlación positiva

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



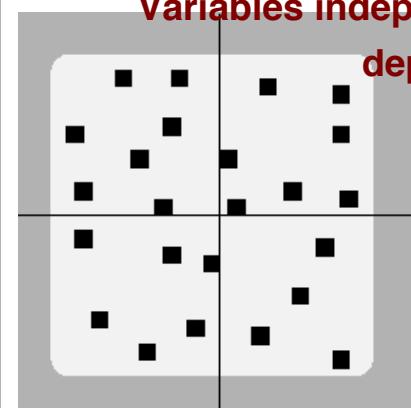
Correlación negativa

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

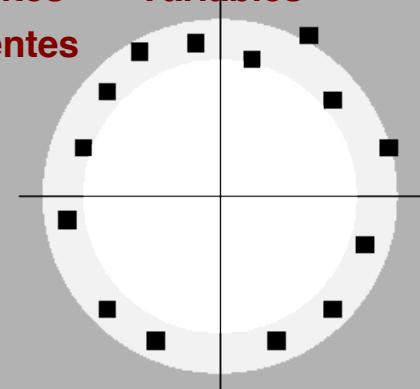


Ausencia de correlación $\text{cov}(x,y) = 0$

**Variables independientes
Variables dependientes**



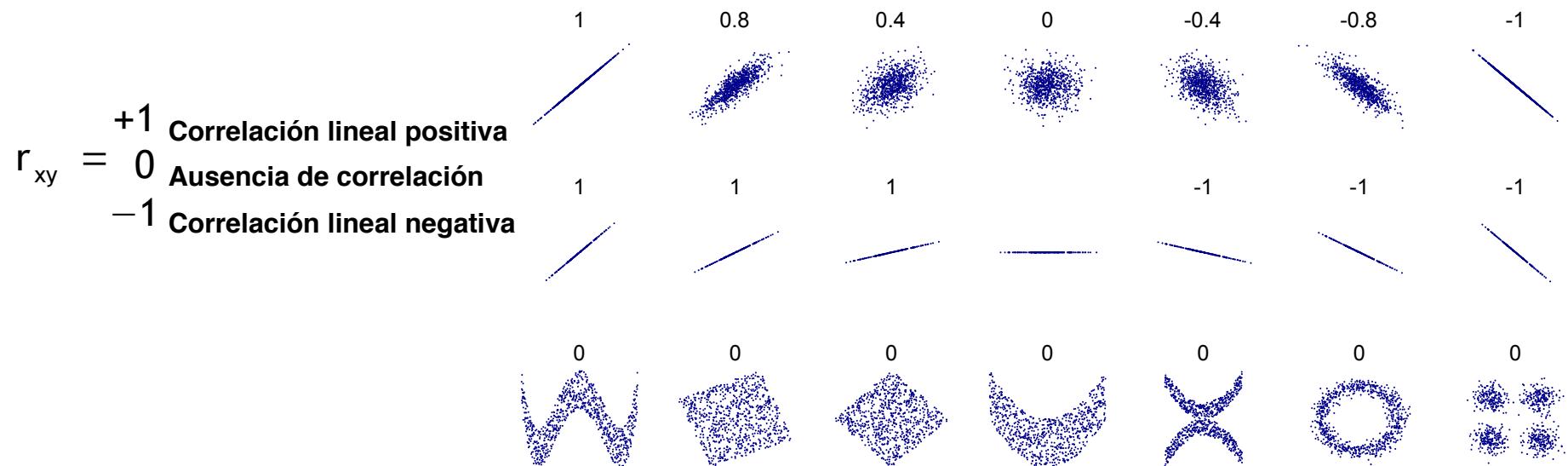
**Variables
independientes
Variables
dependientes**



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

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

- El coeficiente de correlación lineal es **adimensional e invariante ante cambios de escala**.
- Su dominio está **acotado** en el intervalo $[-1,1]$ (como se probará en la siguiente transparencia).



Vamos a demostrar que **el dominio del coeficiente de correlación lineal está acotado en el intervalo [-1,1]**. Para ello escogeremos un valor de λ en la expresión (definida positiva para cualquier valor de λ):

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

Desarrollando la expresión:

$$\sum_{i=1}^k \sum_{j=1}^l f_{ij} [(x_i - \bar{x}) - \lambda(y_j - \bar{y})]^2 =$$

$$\sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})^2 + \lambda^2 \sum_{i=1}^k \sum_{j=1}^l f_{ij} (y_j - \bar{y})^2 - 2\lambda \sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})(y_j - \bar{y}) = s_x^2 + \lambda^2 s_y^2 - 2\lambda \text{cov}(x, y)$$

Si se elige el coeficiente $\lambda = \frac{\text{cov}(x, y)}{s_y^2}$, tendremos que:

$$\begin{aligned} s_x^2 + \lambda^2 s_y^2 - 2\lambda \text{cov}(x, y) \\ = s_x^2 + \frac{[\text{cov}(x, y)]^2}{(s_y^2)^2} s_y^2 - 2 \frac{\text{cov}(x, y)}{s_y^2} \text{cov}(x, y) \\ = s_x^2 - \frac{[\text{cov}(x, y)]^2}{s_y^2} \geq 0 \end{aligned}$$

Con lo que $|\text{cov}(x, y)| \leq s_x s_y$ y $|r_{xy}| = \frac{\text{cov}(x, y)}{s_x s_y} \leq 1$

Medidas características muestrales multivariantes

Vamos a demostrar que **los valores -1 y 1 corresponden a una correlación lineal perfecta entre las variables**. Para ello escogeremos un valor de λ en la expresión:

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

Para $r_{xy} = 1$ tendremos, por definición, que $\text{cov}(x, y) = s_x s_y$, con lo que:

$$s^2(x + \lambda y) = s_x^2 + \lambda^2 s_y^2 + 2\lambda s_x s_y = [s_x + \lambda s_y]^2$$

Con lo que tomando $\lambda = \frac{-s_x}{s_y}$ se llega a:

$$s^2\left(x - \frac{s_x}{s_y}y\right) = \left[s_x - \frac{s_x}{s_y}s_y\right]^2 = 0$$

Donde se ve que existe una coeficiente de la relación lineal que anula la varianza.



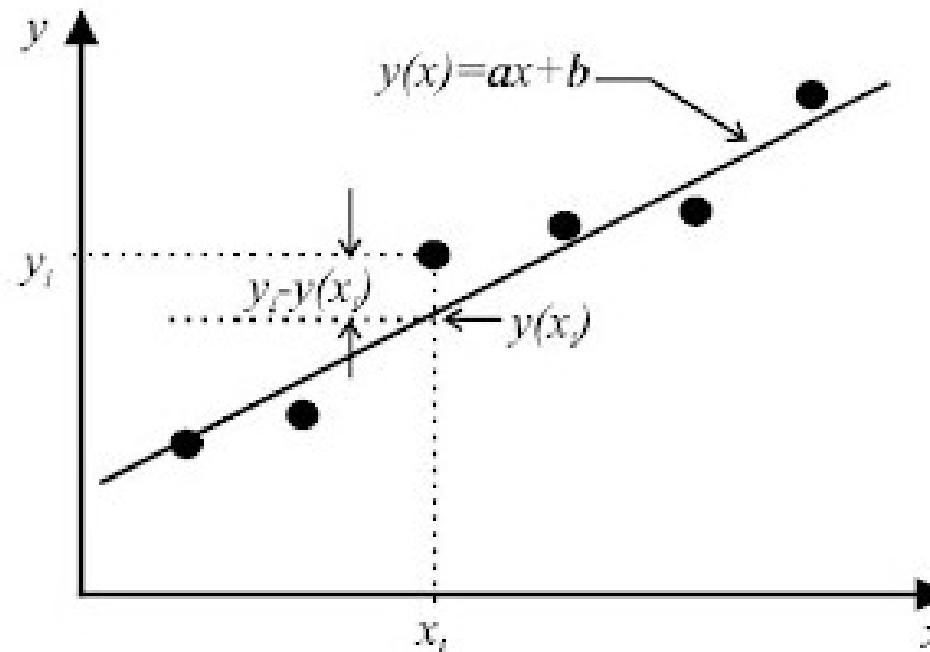
Regresión lineal y covarianza

Sobre una muestra bidimensional $\{(x_i, y_i)\}_{i=1}^N$, asociada a una variable estadística bidimensional (**X,Y**), emitiremos la hipótesis de que **X e Y se relacionan linealmente** según:

$$Y = a + bX$$

Los resultados del experimento aleatorio, sujetos a la variabilidad propia del experimento, de sus variables características o de las distribuciones individuales de las variables, hacen necesario que los parámetros **a** y **b** se tengan que deducir del mejor ajuste posible a los datos.

Para obtener este ajuste, intentaremos reducir la distancia vertical entre los puntos (x_i, y_i) y la recta definida como $y = a + bx$, esto es, reducir las distancias definidas como $d_i = y_i - a - b x_i$



Regresión lineal y covarianza

Para reducir estas distancias $d_i = y_i - a - b x_i$, tomaremos la suma de sus cuadrados:

$$G(a, b) = (y_1 - a - b x_1)^2 + (y_2 - a - b x_2)^2 + \dots + (y_n - a - b x_n)^2 = \sum_{i=1}^N (y_i - a - b x_i)^2 = \sum_{i=1}^N d_i^2$$

Y buscaremos su mínimo en función de a y b , esto es, los parámetros (a_0, b_0) de la recta que reducen la suma de las distancias (al cuadrado) a los puntos (x_i, y_i) .

$$\left(\frac{\partial G(a, b)}{\partial a} \right)_{(a_0, b_0)} = 0 \quad \left(\frac{\partial G(a, b)}{\partial b} \right)_{(a_0, b_0)} = 0$$

Este método se denomina **método de ajuste por mínimos cuadrados**. Resolviendo las ecuaciones algebraicas lineales para los parámetros (a_0, b_0) de la recta, se obtiene:

$$\begin{aligned} \frac{\partial G(a, b)}{\partial a} &= -2 \sum_{i=1}^N (y_i - a_0 - b_0 x_i) = 0 \quad \rightarrow \quad \sum_{i=1}^N y_i = \sum_{i=1}^N a_0 + b_0 \sum_{i=1}^N x_i \\ \frac{\partial G(a, b)}{\partial b} &= -2 \sum_{i=1}^N x_i (y_i - a_0 - b_0 x_i) = 0 \quad \rightarrow \quad \sum_{i=1}^N x_i y_i = \sum_{i=1}^N a_0 x_i + b_0 \sum_{i=1}^N x_i^2 \end{aligned}$$

Regresión lineal y covarianza

Multiplicando por $1/N$ y utilizando la definición de los promedios de las variables estadísticas, se tiene:

$$\frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N a_0 + \frac{b_0}{N} \sum_{i=1}^N x_i \rightarrow \bar{y} = a_0 + b_0 \bar{x}$$

$$\frac{1}{N} \sum_{i=1}^N x_i y_i = \frac{1}{N} \sum_{i=1}^N a_0 x_i + \frac{b_0}{N} \sum_{i=1}^N x_i^2 \rightarrow \bar{xy} = a_0 \bar{x} + b_0 \bar{x^2}$$

podemos resolver el sistema

$$\bar{xy} = (\bar{y} - b_0 \bar{x}) \bar{x} + b_0 \bar{x^2} \rightarrow \bar{xy} - \bar{x} \bar{y} = b_0 (\bar{x^2} - (\bar{x})^2)$$

y obtener los parámetros:

$$b_0 = \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x^2} - (\bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2}$$

$$a_0 = \bar{y} - \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x^2} - (\bar{x})^2} \bar{x} = \bar{y} - \frac{\text{cov}(x, y)}{s_x^2} \bar{x}$$

Que serán los valores de los parámetros de la recta $y = a_0 + b_0 x$ que mejor ajusta al conjunto de los datos.

Extensión multivariante: matriz de covarianza

Generalizando el concepto a distribuciones de frecuencia multivariantes del tipo $\vec{X} = (X_1, X_2, \dots, X_p)$ de la que se extrae una muestra $\{(x_1^i, x_2^i, \dots, x_p^i)\}_{i=1}^N$ definiremos el vector de las medias como

$$\bar{X}_i = \frac{1}{N} \sum_{k=1}^N x_i^k$$

Y la matriz de covarianza correspondiente tendrá elementos del tipo:

$$m_{ij} = \text{cov}(x_i, x_j) = \frac{1}{N} \sum_{k=1}^N (x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j)$$

De forma que la matriz M de covarianza será:

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

Su determinante $s_g^2 = \det(M)$ es una medida de la variabilidad global de la muestra, denominada **varianza generalizada**, y es no negativa por definición.

Propiedades de la matriz de covarianza:

- Esta matriz es cuadrada y simétrica de orden p , donde los términos diagonales son las varianzas y los no diagonales, las covarianzas entre las variables.
- La matriz de covarianza es simétrica respecto a su diagonal principal.
- La matriz de covarianza es semidefinida positiva (esto es, para cualquier vector \mathbf{y} se tiene $\mathbf{y}^* \mathbf{S} \mathbf{y} \geq 0$, siendo \mathbf{y}^* el vector transpuesto). También implica que los autovalores son todos positivos.
- El determinante de la matriz de covarianza es siempre no negativo.
- En el caso bidimensional tendremos: $\det \mathbf{M} = S_x^2 S_y^2 - (S_{xy})^2$

Extensión multivariante: matriz de correlación

Llamaremos matriz de correlación a la matriz cuadrada y simétrica que tiene unos en la diagonal y fuera de ella los coeficientes de correlación entre las variables, esto es, términos del tipo:

$$r_{ij} = r(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{s_{x_i} s_{x_j}} = \frac{\frac{1}{N} \sum_{k=1}^N (x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j)}{\sqrt{\frac{1}{N} \sum_{k=1}^N (x_i^k - \bar{x}_i)^2} \sqrt{\frac{1}{N} \sum_{k=1}^N (x_j^k - \bar{x}_j)^2}}$$

De forma que la matriz R de correlación será:

$$R = \begin{pmatrix} 1 & r(x_1, x_2) & r(x_1, x_3) & \dots & r(x_1, x_p) \\ r(x_2, x_1) & 1 & r(x_2, x_3) & \dots & r(x_2, x_p) \\ r(x_3, x_1) & r(x_3, x_2) & 1 & \dots & r(x_3, x_p) \\ \dots & \dots & \dots & \dots & \dots \\ r(x_p, x_1) & r(x_p, x_2) & r(x_p, x_3) & \dots & 1 \end{pmatrix}$$

Al igual que en el caso del coeficiente de correlación lineal entre dos variables, los elementos de la matriz R son **adimensional e invariantes ante cambios de escala**, con su dominio **acotado** en el intervalo [-1,1].

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños:	2	6	10	5	10	3	2	2
Número de metros:	1	2	3	4	5	6	7	8

Se pide:

- a) Tabla de frecuencias. Diagrama de barras para frecuencias absolutas y relativas de la variable $X = \text{"número de metros que el niño anda"}$. Diagrama de frecuencias absolutas y relativas acumuladas.
- b) Moda, mediana, cuartiles y deciles.
- c) Rango intercuartílico. ¿Existe algun dato atípico leve? ¿Y algun dato atípico extremo? ¿Que significado tendrían en el caso de que apareciesen?
- d) Media geométrica, aritmética, cuadrática y armónica.
- e) Analícese la dispersión de la distribución a través la varianza, la desviación típica y el coeficiente de variación de Pearson.
- f) Momentos respecto al origen de primer, segundo y tercer orden.
- g) Momentos centrales (respecto a la media) de orden primero y tercero.
- h) Estúdiese la asimetría y la curtosis de la distribución.

Problemas

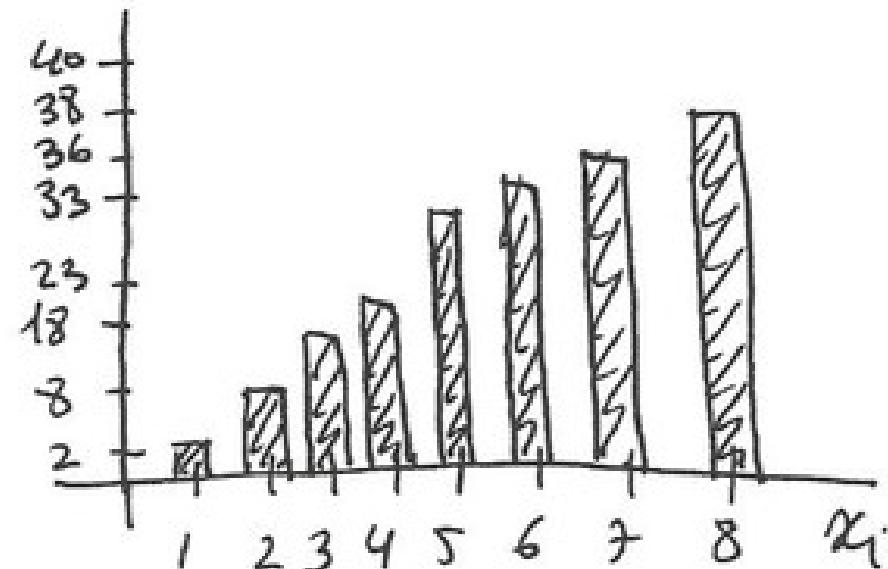
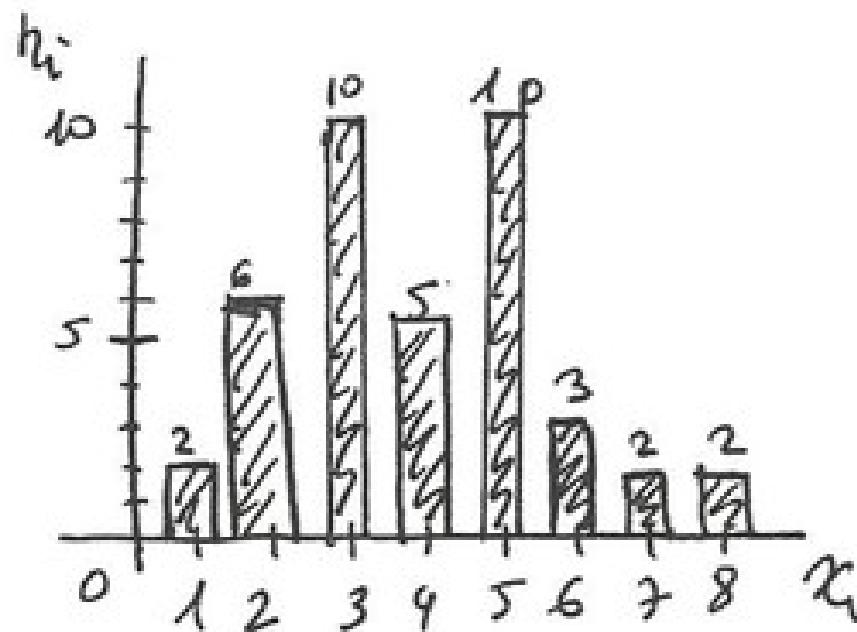
Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2

Número de metros: 1 2 3 4 5 6 7 8

- a) Tabla de frecuencias. Diagrama de barras para frecuencias absolutas y relativas de la variable X = “número de metros que el niño anda”. Diagrama de frecuencias absolutas y relativas acumuladas.

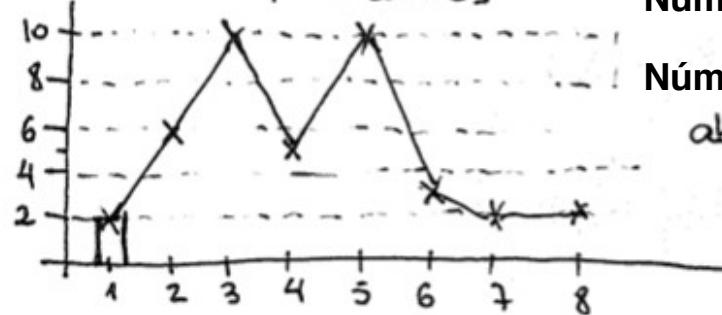
El número total de niños es $N = 40$



Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Diagramas de frecuencias



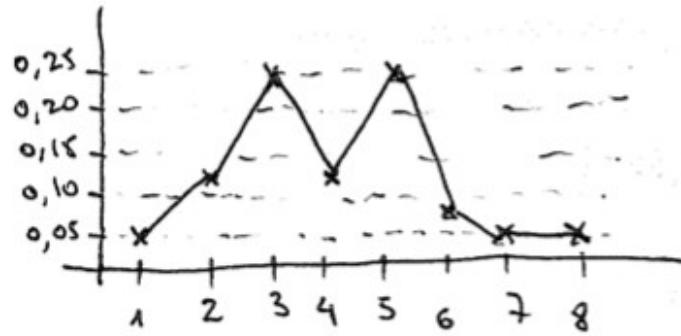
Número de niños:

2 6 10 5 10 3 2 2

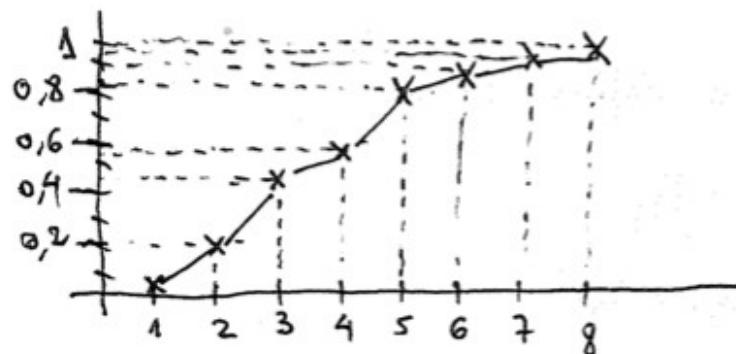
Número de metros:

1 2 3 4 5 6 7 8

absolutas



relativas



relativas acumuladas

x_i	n_i	f_i	N_i	F_i
1	2	0.05	2	0.05
2	6	0.15	8	0.15
3	10	0.25	18	0.4
4	5	0.125	23	0.525
5	9	0.225	33	0.75
6	4	0.1	36	0.86
7	2	0.05	38	0.91
8	2	0.05	40	1.0

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 N = 40

Número de metros: 1 2 3 4 5 6 7 8

b) Moda, mediana, cuartiles y deciles.

Moda: La distribución es bimodal, con valores **Md = {3, 5}**.

Mediana: Si los datos se consideran valores discretos, tendremos que **N/2 = 20**. Como en la tabla no hay ningún **N_i** con el valor 20, se tomaría el inmediatamente superior, **N₄ = 23**, que corresponde a **Md = 4**.

En el caso de considerar los datos como agrupados sobre una variable continua (la longitud) entonces interpolariamos entre los datos agrupados en **x_i = 3** y **x_{i+1} = 4** de forma que

$$\frac{N_{i+1} - N_i}{x_{i+1} - x_i} = \frac{qN - N_i}{P_q - x_i} \rightarrow \frac{23 - 18}{4 - 3} = \frac{20 - 18}{P_{0.5} - 3} \rightarrow P_{0.5} = M_d = 3 + \frac{2}{5} = 3,4$$

Cuartiles: Como distribuciones discretas, se puede ver que **P_{0.25} = 3** (N/4=10) y **P_{0.75} = 5** (3N/4=30).

Como datos agrupados, entonces **P_{0.25} = 2,2** (N/4=10) y **P_{0.75} = 4,7** (3N/4=30).

Deciles: Como distribuciones discretas, se puede ver que **P_{0.1} = 2** (N/10=4), **P_{0.2} = 2,5** (2N/10=8), ...

Como datos agrupados, entonces **P_{0.25} = 1,33** (N/10=4), ...

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 **N = 40**

Número de metros: 1 2 3 4 5 6 7 8

c) Rango intercuartílico. ¿Existe algún dato atípico leve? ¿Y algún dato atípico extremo? ¿Qué significado tendrían en el caso de que apareciesen?

El rango intercuartílico es $RI = P_{0.75} - P_{0.25}$. En el caso discreto $RI = 2$. En el caso agrupado $RI = 2.5$.

Se considera un **dato atípico leve** el que aparece a más de $1.5RI$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$. Se considera un **dato atípico extremo** el que aparece a más de $3RI$ por encima de $P_{0.75}$ o por debajo de $P_{0.25}$.

En este caso no se encuentran datos atípicos de ninguna de las dos clases.

¡Los datos atípicos 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.

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 **N = 40**

Número de metros: 1 2 3 4 5 6 7 8

d) Media geométrica, aritmética, cuadrática y armónica.

$$\text{La media aritmética es } \bar{x} = \sum_{i=1}^k f_i x_i = \frac{2}{40}1 + \frac{6}{40}2 + \dots + \frac{2}{40}8 = 4,05$$

$$\text{La media geométrica es } \bar{x}_g = \sqrt[N]{\prod_{i=1}^k x_i^{n_i}} = \sqrt[40]{1^2 \times 2^6 \times \dots \times 8^2} = 3,6325$$

$$\text{La media cuadrática es } \bar{x}_q = \sqrt{\sum_{i=1}^k f_i x_i^2} = \sqrt{\frac{2}{40} \times 1^2 + \frac{6}{40} \times 2^2 + \dots + \frac{2}{40} \times 8^2} = 4,416$$

$$\text{La media armónica es } \frac{1}{\bar{x}_a} = \sum_{i=1}^k \frac{f_i}{x_i} \rightarrow \bar{x}_a = \frac{1}{\frac{2}{40 \cdot 1} + \frac{6}{40 \cdot 2} + \dots + \frac{2}{40 \cdot 8}} = 3,17$$

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

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 **N = 40**

Número de metros: 1 2 3 4 5 6 7 8

e) Analícese la dispersión de la distribución a través la varianza, la desviación típica y el coeficiente de variación de Pearson.

La varianza es

$$s^2 = \sum_{i=1}^k f_i(x_i - \bar{x})^2 = \frac{2}{40}(1-4,05)^2 + \frac{6}{40}(2-4,05)^2 + \dots + \frac{2}{40}(8-4,05)^2 = 3,0975$$

La desviación típica es

$$s = \sqrt{s^2} = 1,76$$

El coeficiente de variación de Pearson es $CV = \frac{s}{|\bar{x}|} = \frac{1,76}{4,05} = 0,43$

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 **N = 40**

Número de metros: 1 2 3 4 5 6 7 8

f) Momentos respecto al origen de primer, segundo y tercer orden.

El momento de primer orden respecto al origen es la media

$$m_1(0) = \sum_{i=1}^k f_i(x_i - 0)^1 = \overline{(x_i - 0)^1} = \sum_{i=1}^k f_i x_i = \bar{x} = 4,05$$

El momento de segundo orden respecto al origen es

$$m_2(0) = \sum_{i=1}^k f_i(x_i)^2 = 19,5$$

El momento de tercer orden respecto al origen es

$$m_3(0) = \sum_{i=1}^k f_i(x_i)^3 = 106,2$$

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 **N = 40**

Número de metros: 1 2 3 4 5 6 7 8

g) Momentos centrales (respecto a la media) de orden primero y tercero.

El momento central de primer orden es nulo por definición

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

El momento central de segundo orden es la varianza

$$m_2(\bar{x}) = \sum_{i=1}^k f_i(x_i - \bar{x})^2 = 3,0975$$

El momento central de tercer orden es

$$m_3(\bar{x}) = \sum_{i=1}^k f_i(x_i - \bar{x})^3 = 2,1353$$

Problemas

Problema 1.1: En una clínica infantil se han ido anotando durante un mes, el número de metros que andan sus niños seguidos y sin caerse, el primer día que comienzan a caminar. La tabla obtenida fue:

Número de niños: 2 6 10 5 10 3 2 2 N = 40

Número de metros: 1 2 3 4 5 6 7 8

h) Estúdiese la asimetría y la curtosis de la distribución.

La asimetría es

$$A_F = \frac{1}{S^3} \sum_{i=1}^k f_i (x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{S^3} = 0,3917$$

La curtosis es

$$g = \beta_2 = \frac{1}{S^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{S^4} = 2,5563$$

Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

Calcúlense: a) Histograma y curva de frecuencias de la muestra de datos.

b) Medidas características de la muestra: medias (aritmética y geométrica), varianza y desviación típica, coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento. ¿Cuál es la concentración a la que podemos decir que comienza la formación de los agregados objeto de estudio? ¿Con qué incertidumbre?

c) Analizar la existencia de datos atípicos. ¿Qué medida de centralización es más resistente a este tipo de datos?

Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

a) Histograma y curva de frecuencias de la muestra de datos.

$$[38, 44) \rightarrow \{38, 41\}$$

$$[44, 50) \rightarrow \{49, 47, 46, 46\}$$

$$[50, 56) \rightarrow \{52, 51, 50, 52, 55, 54\}$$

$$[56, 62) \rightarrow \{61, 59, 57, 56, 57\}$$

$$[62, 68) \rightarrow \{63, 67, 65\}$$

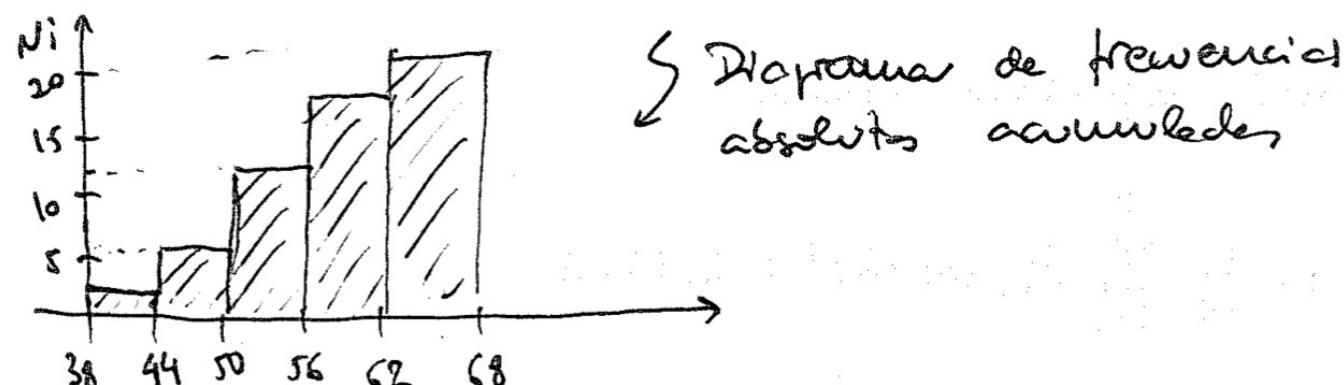
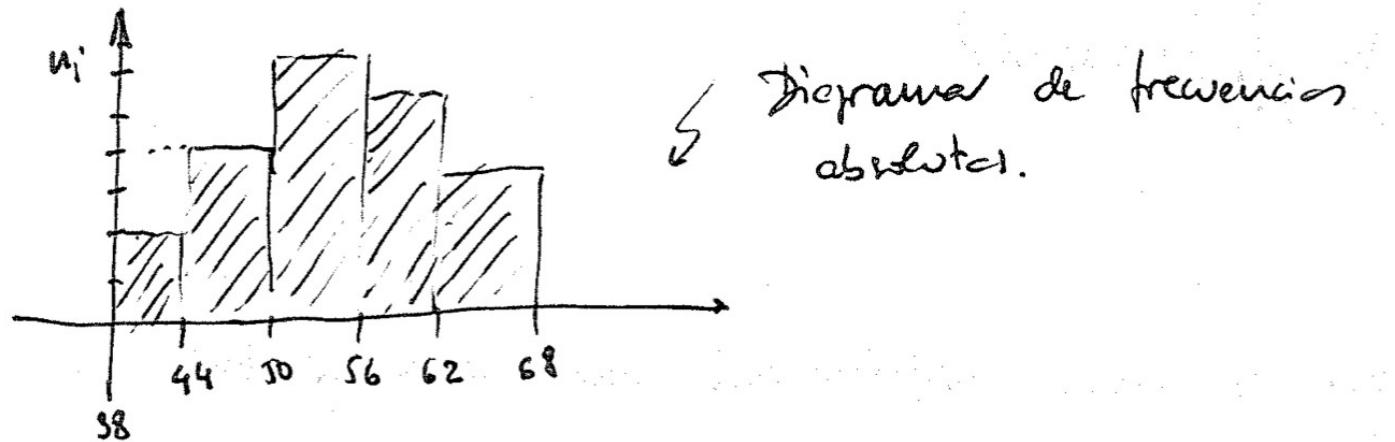
Intervalo	x	n _i	f _i	N _i	F _i
[38, 44)	41	2	0.1	2	0.1
[44, 50)	47	4	0.2	6	0.3
[50, 56)	53	6	0.3	12	0.6
[56, 62)	59	5	0.25	17	0.85
[62, 68)	65	3	0.15	20	1

Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

a) Histoarama v curva de frecuencias de la muestra de datos.



Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

b) Medidas características de la muestra: medias (aritmética y geométrica), varianza y desviación típica, coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento.

La media aritmética es $\bar{x} = \sum_{i=1}^k f_i x_i = 53,09$

La media geométrica es $\bar{x}_g = \sqrt[N]{\prod_{i=1}^k x_i^{n_i}} = 53,411$

La varianza es $s^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2 = 51,39$

La desviación típica es $s = \sqrt{s^2} = 7,17$

Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

b) ... el coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento. ¿Cuál es la concentración a la que podemos decir que comienza la formación de los agregados objeto de estudio? ¿Con qué incertidumbre?

El coeficiente de variación de Pearson es $CV = \frac{s}{|\bar{x}|} = 0,133$

El coeficiente de asimetría de Pearson-Fisher es $A_F = \frac{1}{S^3} \sum_{i=1}^k f_i (x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{S^3} = -0,1148$

mostrando asimetría negativa.

El coeficiente de apuntamiento es $g = \beta_2 = \frac{1}{S^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{S^4} = 2,1466$

con los que, al ser menor que 3, la distribución es platicúrtica.

La concentración resultado del experimento es $\bar{x} = 53,09$ $s_A(\bar{x}) = \frac{7,17}{\sqrt{20}} = 1,6$

Nótese que si no se hubieran agrupado los datos, tendríamos

$$\bar{x} = 53,3 \quad s_A(\bar{x}) = 1,67$$

Problemas

Problema 1.2: En un determinado experimento de medición de la concentración de formación de agregados en una disolución, se obtuvieron los siguientes resultados (en mmol/l):

52	61	49	46	51	50	59	63	57	67
56	41	52	47	46	55	38	65	57	54

c) Analizar la existencia de datos atípicos. ¿Qué medida de centralización es más resistente a este tipo de datos?

Cuartiles: Como datos agrupados, $P_{0.25} = 45,5$ ($N/4=5$) y $P_{0.75} = 56,6$ ($3N/4=15$). El rango intercuartilico es $RI = P_{0.75} - P_{0.25} = 11,1$.

Se considera un dato atípico leve el que aparece a más de 1.5 RI por encima de $P_{0.75}$ (73,25) o por debajo de $P_{0.25}$ (28,85). En este caso no se encuentran datos atípicos .

Tipicamente, la mediana $Md = P_{0.5} = 51$ (53 si consideramos los datos individualmente, sin agruparlos en clases) es la medida de centralización mas resistente frente a los datos atípicos.

RECUERDA: ¡Los datos atípicos 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.

Problemas

Problema 1.3: En un determinado experimento de medición de la densidad de una disolución acuosa diluida a 25 C de temperatura mediante medidas de masa (m) y volumen (V), se obtuvieron los siguientes resultados:

m (g)	1,1	1,2	1,1	0,9	1,0	1,0	0,9	1,1	1,0	0,9
V (cm ³)	1,1	1,1	1,1	0,9	1,0	1,0	0,8	1,1	1,0	0,9

Se pide:

- Histograma de la densidad.
- Medidas características de la muestra de densidades: medias (aritmética y geométrica), varianza y desviación típica, coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento. ¿Cuál es la densidad que podemos atribuir a la disolución problema? ¿Con qué incertidumbre?

Problemas

Problema 1.3: En un determinado experimento de medición de la densidad de una disolución acuosa diluida a 25 C de temperatura mediante medidas de masa (m) y volumen (V), se obtuvieron los siguientes resultados:

m (g)	1,1	1,2	1,1	0,9	1,0	1,0	0,9	1,1	1,0	0,9
V (cm ³)	1,1	1,1	1,1	0,9	1,0	1,0	0,8	1,1	1,0	0,9

Se pide:

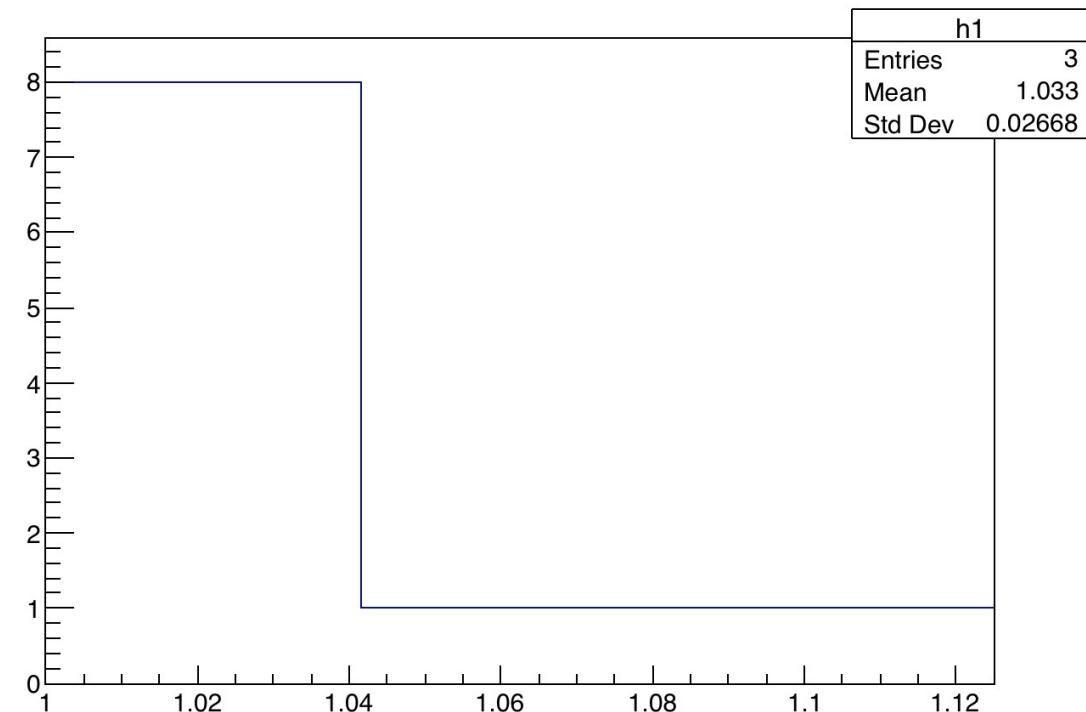
- a) Histograma de la densidad.

$$d \text{ (g/cm}^3\text{)} = \{ 1,0, 1,09, 1,0, 1,0, 1,0, 1,0, 1,125, 1,0, 1,0, 1,0 \}$$

$$n(d=1,0) = 8; f(d=1,0) = 8/10;$$

$$n(d=1,09) = 1; f(d=1,09) = 1/10;$$

$$n(d=1,125) = 1; f(d=1,125) = 1/10;$$

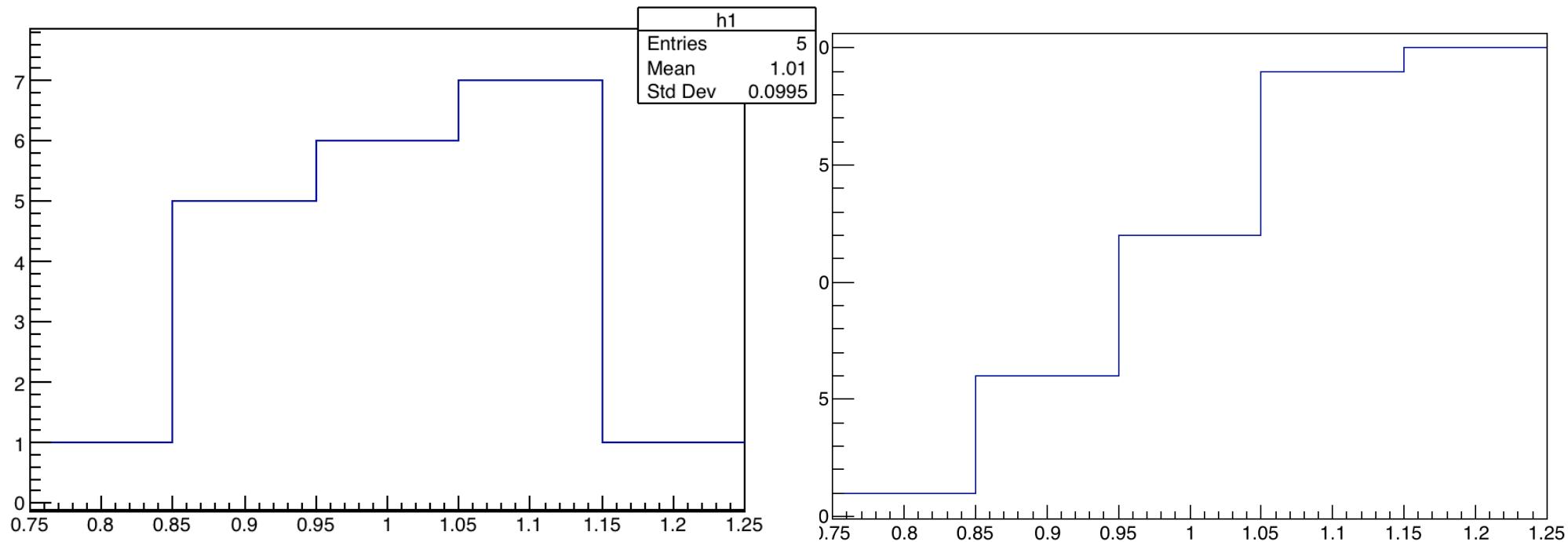


Problemas

Problema 1.3: En un determinado experimento de medición de la densidad de una disolución acuosa diluida a 25 C de temperatura mediante medidas de masa (m) y volumen (V), se obtuvieron los siguientes resultados:

m (g)	1,1	1,2	1,1	0,9	1,0	1,0	0,9	1,1	1,0	0,9
V (cm ³)	1,1	1,1	1,1	0,9	1,0	1,0	0,8	1,1	1,0	0,9

a) Histograma de la densidad.



Problemas

Problema 1.3: En un determinado experimento de medición de la densidad de una disolución acuosa diluida a 25 C de temperatura mediante medidas de masa (m) y volumen (V), se obtuvieron los siguientes resultados:

m (g)	1,1	1,2	1,1	0,9	1,0	1,0	0,9	1,1	1,0	0,9
V (cm ³)	1,1	1,1	1,1	0,9	1,0	1,0	0,8	1,1	1,0	0,9

b) Medidas características de la muestra de densidades: medias (aritmética y geométrica), varianza y desviación típica, coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento. ¿Cuál es la densidad que podemos atribuir a la disolución problema? ¿Con qué incertidumbre?

$$\text{La media aritmética es } \bar{x} = \sum_{i=1}^k f_i x_i = \frac{1}{10}(8*1+1,09+1,125) = 1,0215$$

$$\text{La media geométrica es } \bar{x}_g = \sqrt[N]{\prod_{i=1}^k x_i^{n_i}} = \sqrt[10]{1^{10}*1,09*1,125} = 1,0206$$

$$\begin{aligned} \text{La varianza es } s^2 &= \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \\ &\frac{1}{10}(8(1-1,0215)^2 + (1,09-1,0215)^2 + (1,125-1,0215)^2) = 0,00191 \end{aligned}$$

$$\text{La desviación típica es } s = \sqrt{s^2} = 0,044$$

Problemas

Problema 1.3: En un determinado experimento de medición de la densidad de una disolución acuosa diluida a 25 C de temperatura mediante medidas de masa (m) y volumen (V), se obtuvieron los siguientes resultados:

m (g)	1,1	1,2	1,1	0,9	1,0	1,0	0,9	1,1	1,0	0,9
V (cm ³)	1,1	1,1	1,1	0,9	1,0	1,0	0,8	1,1	1,0	0,9

b) Medidas características de la muestra de densidades: medias (aritmética y geométrica), varianza y desviación típica, coeficiente de variación de Pearson, coeficientes de asimetría de Pearson y Fisher, y coeficiente de apuntamiento. ¿Cuál es la densidad que podemos atribuir a la disolución problema? ¿Con qué incertidumbre?

El coeficiente de variación de Pearson es $CV = \frac{s}{|\bar{x}|} = \frac{0,044}{1,0215} = 0,043$

El coeficiente de asimetría de Pearson-Fisher es $A_F = \frac{1}{s^3} \sum_{i=1}^k f_i (x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{s^3} = \frac{0,000135}{(0,044)^3} = 1,585$
mostrando asimetría positiva.

El coeficiente de apuntamiento $g = \beta_2 = \frac{1}{s^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{s^4} = \frac{0,00001385}{(0,044)^4} = 3,695$
con los que, al ser mayor que 3, la distribución es leptocurtica.

La concentración densidad del experimento es $\bar{x} = 1,0215$ $s_A(\bar{x}) = \frac{0,044}{\sqrt{10}} = 0,014$

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

- a) ¿Qué tipo de variable aleatoria estamos analizando? ¿Cuál es el número óptimo de clases?
Agrúpense los datos en intervalos de amplitud constante.
- b) Media aritmética, media geométrica y armónica.
- c) Mediana y desviación media respecto a la mediana.
- d) Coeficientes de variación media.
- e) Coeficiente de variación de Pearson, de asimetría de Pearson-Fisher y coeficiente de apuntamiento.

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

a) ¿Qué tipo de variable aleatoria estamos analizando? ¿Cuál es el número óptimo de clases?
Agrúpense los datos en intervalos de amplitud constante.

Es una variable en principio continua (altura de los reclutas) medida con una aparato que parece presentar una resolución limitada por la distribución de los valores. Al ser 16 valores podemos agruparla en 4 diferentes clases, aunque un número mayor de intervalos es posible.

Como el valor mínimo es 160,0 y el máximo 198,0, podremos dividir el recorrido de la variable (38) en 4 clases: de 160,0 a 169,5, ... O por simplicidad podemos hacer entre 160 y 200 →

[160,170) → $f_1 = 5$ con marca de clase 165

[170,180) → $f_2 = 6$ con marca de clase 175

[180,190) → $f_3 = 3$ con marca de clase 185

[190,200) → $f_4 = 2$ con marca de clase 195

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

b) Media aritmética, media geométrica y armónica.

$$\text{La media aritmética es } \bar{x} = \sum_{i=1}^k f_i x_i = (5*165+6*175+3*185+2*195)/16 = 176,25$$

$$\bar{x} = \sum_{i=1}^N x_i = (160.0+172.4+168.0+\dots+173.5)/16 = 175,775$$

La media geométrica es

$$\bar{x}_g = \sqrt[N]{\prod_{i=1}^k x_i^{n_i}} = 175,975$$

La media armónica es

$$\frac{1}{\bar{x}_a} = \sum_{i=1}^k \frac{f_i}{x_i} \rightarrow \bar{x}_a = 175,706$$

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

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

c) Mediana y desviación media respecto a la mediana.

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

Vamos a tratarlo como un problema de variable continua para ver este caso:

Por tanto: $N_i = 5$, $N_{i+1} = 11$, $x_i = 165$, $x_{i+1} = 175$, con lo que $P_{0.5} = 170$.

Si se consideraran los valores individualmente sin agrupar en marcas de clase, $Me = P_{0.5} = 175$.

La desviación media respecto a la mediana es:

$$DM_{Me} = \sum_{i=1}^k f_i |x_i - Me| = 9,375$$

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

d) Coeficientes de variación media.

Coeficiente de variación media respecto a la mediana

$$CV_{MMe} = \frac{DM_{Me}}{|Me|} = \frac{9,375}{170} = 0,055$$

Problemas

Problema 1.4: En un reclutamiento militar se ha tomado una muestra de 16 jóvenes teniéndose las siguientes estaturas (en cm): 160,0 172,4 168,0 167,0 175,0 179,0 180,0 198,0
164,0 166,0 174,0 177,0 182,5 185,0 191,0 173,5

e) Coeficiente de variación de Pearson, de asimetría de Pearson-Fisher y coeficiente de apuntamiento.

$$\text{El coeficiente de variación de Pearson es } CV = \frac{s}{|\bar{x}|} = \frac{10,25}{176,25} = 0,058$$

$$\text{El coeficiente de asimetría de Pearson-Fisher es } A_F = \frac{1}{s^3} \sum_{i=1}^k f_i (x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{s^3} = 0,516$$

mostrando asimetría negativa.

$$\text{El coeficiente de apuntamiento es } g = \beta_2 = \frac{1}{s^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{s^4} = 2,224$$

con los que, al ser menor que 3, la distribución es platicúrtica.

Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

- a) Histogramense los datos de frecuencia absoluta y relativa de la variable aleatoria **X= “masa atómica del elemento”**. ¿Cuántas y qué clases se han elegido?
- b) ¿Qué valor podemos asignar a la masa atómica del elemento?
- c) Si tomamos como medida de la incertidumbre experimental la desviación típica de la distribución (incertidumbre estadística de tipo A). ¿qué valor toma dicho incertidumbre? Usualmente, sin embargo, se prefiere como medida de esta incertidumbre estadística experimental la denominada desviación típica de la medida cuyo valor es

$$s_A(\bar{x}) = \frac{s_x}{\sqrt{N}}$$

Dónde **N** es el número de datos. Calcúlese dicho valor y coméntese brevemente la definición anterior.

- d) Analíicense la existencia de datos atípicos leves y extremos y las posibles causas de los mismos.
- e) Obténganse los momentos de primer, segundo, tercero y cuarto orden respecto a la media y al origen.
- f) Analíicense la asimetría y el apuntamiento de la distribución.

Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
 134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
 134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
 134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

a) Histogramense los datos de frecuencia absoluta y relativa de la variable aleatoria X = “masa atómica del elemento”. ¿Cuántas y qué clases se han elegido?

Hay un total de 32 medidas, con lo que una buena partición será utilizar ~ 6 clases.

$$[132, 133) \quad - \quad x_1 = 132,5$$

$$[133, 134) \quad - \quad x_2 = 133,5$$

$$[134, 135) \quad - \quad x_3 = 134,5$$

$$[135, 136) \quad - \quad x_4 = 135,5$$

$$[136, 137) \quad - \quad x_5 = 136,5$$

$$[137, 138] \quad - \quad x_6 = 137,5$$

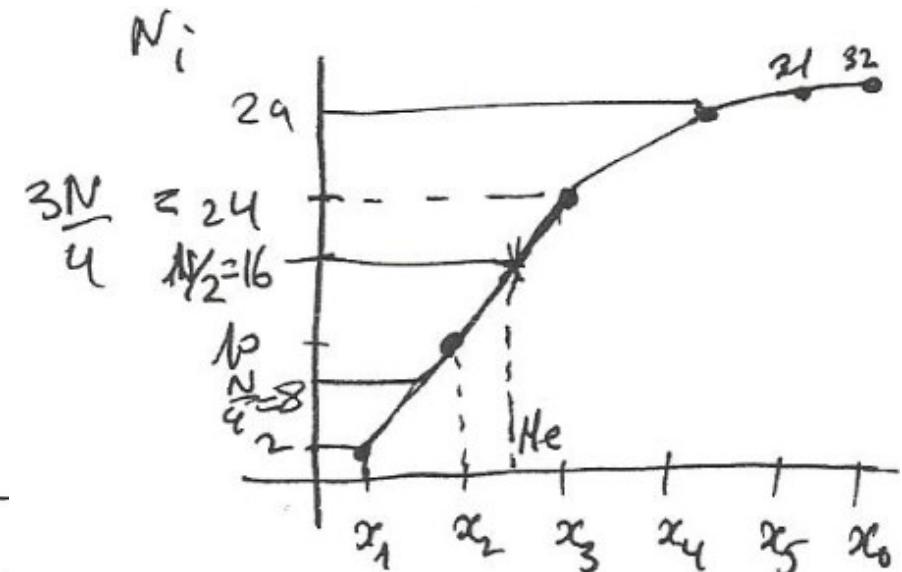
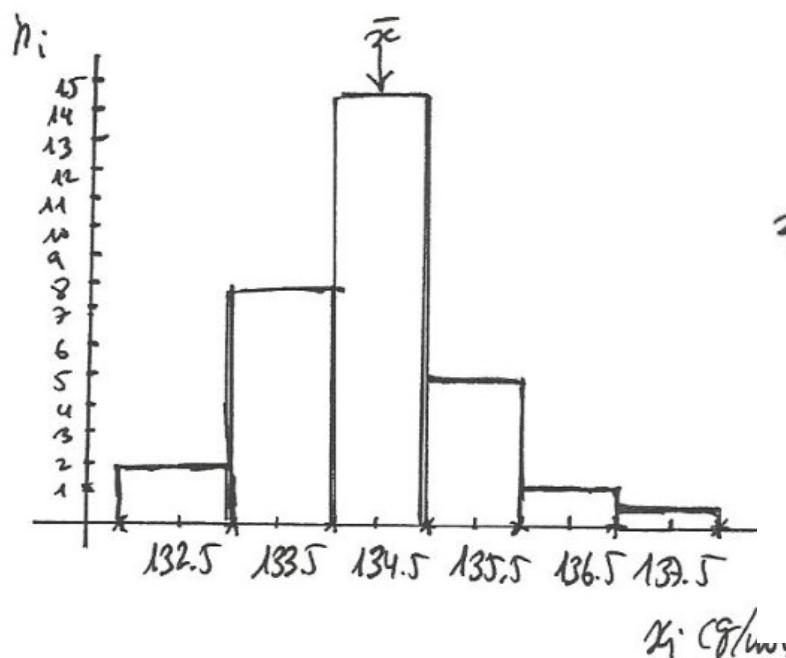
x_i	n_i	f_i	$x_i n_i$	$n_i(x_i - \bar{x})^2$
x_1	2	$2/32$	265	8
x_2	8	$8/32$	1068	8
x_3	14	$14/32$	1883	0
x_4	5	$5/32$	677,5	5
x_5	2	$2/32$	273	8
x_6	1	$1/32$	137,5	9
	32		4304	38

Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
 134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
 134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
 134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

a) Histogramense los datos de frecuencia absoluta y relativa de la variable aleatoria X = “masa atómica del elemento”. ¿Cuántas y qué clases se han elegido?



Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20

134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65

134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85

134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

b) ¿Qué valor podemos asignar a la masa atómica del elemento?

La media es $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{4304}{32} = 134,5$

Que es nuestro mejor estimador del valor de la masa atómica del elemento.

Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
 134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
 134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
 134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

c) Si tomamos como medida de la incertidumbre experimental la desviación típica de la distribución (incertidumbre estadística de tipo A). ¿qué valor toma dicho incertidumbre? Usualmente, sin embargo, se prefiere como medida de esta incertidumbre estadística experimental la denominada desviación típica de la media cuyo valor es

$$s_A(\bar{x}) = \frac{s_x}{\sqrt{N}}$$

Dónde **N** es el número de datos. Calcúlese dicho valor y coméntese brevemente la definición anterior.

La desviación típica de la distribución de frecuencias es: $s_x = \sqrt{s_x^2} = \sqrt{\sum_{i=1}^k f_i(x_i - \bar{x})^2} = 1,1$

La desviación típica de la media es $s_A(\bar{x}) = \frac{s_x}{\sqrt{N}} = \frac{1,1}{\sqrt{32}} = 0,2$

Por tanto el resultado será $\bar{x} = 134,5$ $s_A(\bar{x}) = 0,2$

Problemas

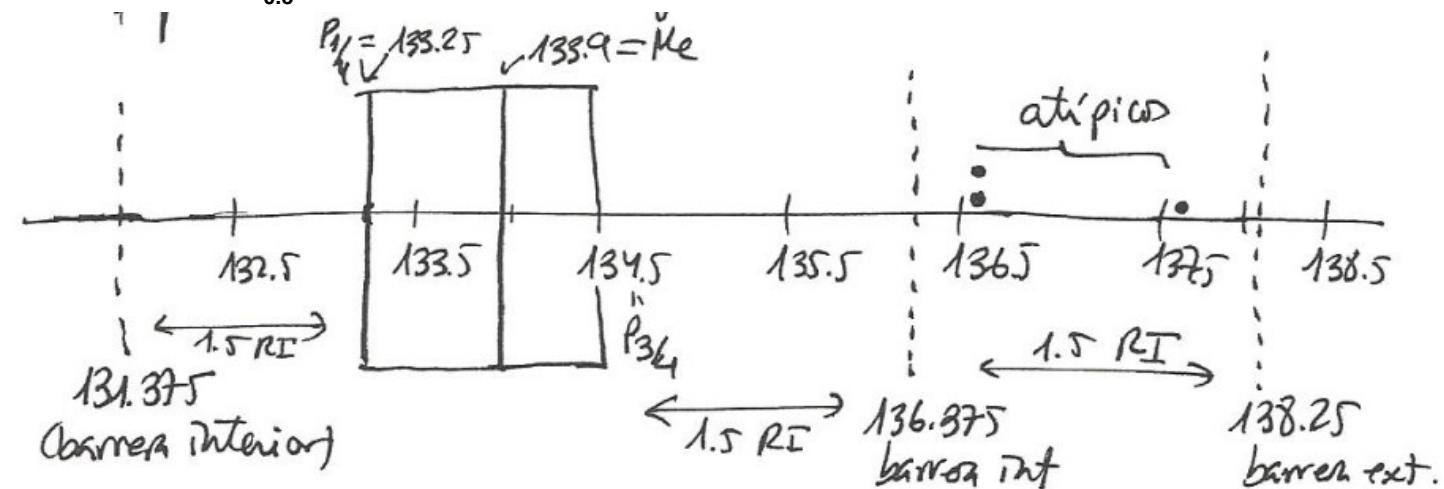
Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
 134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
 134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
 134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

d) Analíicense la existencia de datos atípicos leves y extremos y las posibles causas de los mismos.

Cuartiles: Como datos agrupados, $P_{0.25} = 133,25$ ($N/4=8$) y $P_{0.75} = 134,5$ ($3N/4=24$). El rango intercuartilico es $RI = P_{0.75} - P_{0.25} = 1,75$. Se considera un dato atípico leve el que aparece a más de $1.5 RI$ por encima de $P_{0.75}$ (136,375) o por debajo de $P_{0.25}$ (131,375). ¡Hay datos atípicos leves!

Tipicamente, la mediana $Md = P_{0.5} = 133,9$ es la medida de centralización mas resistente frente a los datos atípicos.



Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20

134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65

134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85

134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

e) Obténganse los momentos de primer, segundo, tercero y cuarto orden respecto a la media y al origen.

El momento de primer orden respecto al origen es la media $m_1(0) = \sum_{i=1}^k f_i x_i = \bar{x} = 134,5$

El momento de segundo orden respecto al origen es $m_2(0) = \sum_{i=1}^k f_i (x_i)^2 = 18091,44$

El momento de tercero orden respecto al origen es $m_3(0) = \sum_{i=1}^k f_i (x_i)^3 = 2,44 \cdot 10^6$

El momento central de primer orden es nulo por definición.

El momento central de segundo orden es la varianza $m_2(\bar{x}) = \sum_{i=1}^k f_i (x_i - \bar{x})^2 = 1,19$

El momento central de tercero orden es $m_3(\bar{x}) = \sum_{i=1}^k f_i (x_i - \bar{x})^3 = 0,75$

Problemas

Problema 1.5: En un experimento de medida de la masa atómica de un determinado elemento químico se han obtenido los siguientes valores (g/mol):

134,56 134,89 133,99 133,56 135,03 134,65 135,10 137,20
134,25 134,78 134,29 133,62 135,23 134,99 135,56 136,65
134,34 134,32 133,05 134,78 133,25 133,01 132,76 132,85
134,66 134,89 133,20 133,82 135,67 136,02 133,78 133,98

f) Analíicense la asimetría y el apuntamiento de la distribución.

La asimetría es positiva

$$A_F = \frac{1}{S^3} \sum_{i=1}^k f_i (x_i - \bar{x})^3 = \frac{m_3(\bar{x})}{S^3} = 0,58 (>0)$$

La curtosis es mayor que 3 (leptocúrtica)

$$g = \beta_2 = \frac{1}{S^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{S^4} = 3,48 (>3)$$

Problemas

Problema 1.6: Demuestrese que si construimos una variable **Z** mezclando n_1 valores de la variable **X** y n_2 valores de la variable **Y** la media de **Z** es: $\bar{z} = \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y}$

Problemas

Problema 1.6: Demuestrese que si construimos una variable **Z** mezclando n_1 valores de la variable **X** y n_2 valores de la variable **Y** la media de **Z** es: $\bar{z} = \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y}$

$$\begin{aligned}\bar{z} &= \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} z_i = \frac{1}{n_1+n_2} \left[\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i \right] = \frac{1}{n_1+n_2} \left[\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i \right] \\ &= \frac{1}{n_1+n_2} [n_1\bar{x} + n_2\bar{y}] = \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y}\end{aligned}$$

En el caso de la varianza:

$$\begin{aligned}s_z^2 &= \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} (z_i - \bar{z})^2 = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} \left(z_i - \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y} \right)^2 \\ &= \frac{1}{n_1+n_2} \left(\sum_{i=1}^{n_1} \left(x_i - \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y} \right)^2 + \sum_{i=1}^{n_2} \left(y_i - \frac{n_1}{n_1+n_2}\bar{x} + \frac{n_2}{n_1+n_2}\bar{y} \right)^2 \right)\end{aligned}$$

Problemas

Problema 1.7: Consideremos la distribución de datos:

x	1	2	3	4	5
y	0.34	0.70	1.08	1.43	1.70

Obtégase la covarianza de las variables aleatorias anteriores.

Problemas

Problema 1.7: Consideremos la distribución de datos:

x	1	2	3	4	5
y	0.34	0.70	1.08	1.43	1.70

Obtégase la covarianza de las variables aleatorias anteriores.

$$\bar{x} = \sum_{i=1}^5 f_i x_i = 3 \quad \bar{y} = \sum_{i=1}^5 f_i y_i = 1,05$$

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5} \sum_{i=1}^5 (x_i - 3)(y_i - 1,05) = 0,69$$

Problemas

Problema 1.8: Considérense dos variables aleatorias **X** e **Y** con varianzas iguales. Calcúlese la covarianza siguiente: **cov [(x + y) , (x - y)]**.

Problemas

Problema 1.8: Considérense dos variables aleatorias **X** e **Y** con varianzas iguales. Calcúlese la covarianza siguiente: **cov [(x + y) , (x - y)]**.

$$\begin{aligned}
 \text{cov}(x+y, x-y) &= \frac{1}{n} \sum_{i=1}^n (x_i + y_i) - (\bar{x} + \bar{y})((x_i - y_i) - (\bar{x} - \bar{y})) \\
 &= \frac{1}{n} \sum_{i=1}^n ((x_i - \bar{x}) + (y_i - \bar{y}))((x_i - \bar{x}) - (y_i - \bar{y})) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - (y_i - \bar{y})^2 \\
 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = s_x^2 - s_y^2 = 0 \quad \text{ya que } s_x^2 = s_y^2
 \end{aligned}$$

Problemas

Problema 1.9: Pruébese que si $q = q(x,y)$ es una función cualquiera de las variables X e Y , y las variaciones de la media de ambas variables son pequeñas entonces se verifica

$$s_q^2 = \left(\frac{\partial q}{\partial x} \right)^2 s_x^2 + \left(\frac{\partial q}{\partial y} \right)^2 s_y^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \text{cov}(x, y)$$

Donde las derivadas deben entenderse evaluados en los valores medios de X e Y , y $\text{cov}(x, y)$ es la covarianza de las variables aleatorias X e Y : $\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^k \sum_{j=1}^l f_{ij}(x_i - \bar{x})(y_j - \bar{y})$

Problemas

Problema 1.9: Pruébese que si $q = q(x, y)$ es una función cualquiera de las variables X e Y , y las variaciones de la media de ambas variables son pequeñas entonces se verifica

$$s_q^2 = \left(\frac{\partial q}{\partial x} \right)^2 s_x^2 + \left(\frac{\partial q}{\partial y} \right)^2 s_y^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \text{cov}(x, y)$$

Donde las derivadas deben entenderse evaluados en los valores medios de X e Y , y $\text{cov}(x, y)$ es la covarianza de las variables aleatorias X e Y : $\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^k \sum_{j=1}^l f_{ij}(x_i - \bar{x})(y_j - \bar{y})$

La varianza es $s_q^2 = \sum_{i=1}^k f_i(q(x_i, y_i) - \bar{q})^2$

Donde el valor medio $\bar{q} = \sum_{i=1}^k f_i q(x_i, y_i)$

Desarrollando q a primer orden en Taylor alrededor de la media en x e y :

$$q(x_i, y_i) = q(\bar{x}, \bar{y}) + \left(\frac{\partial q}{\partial x} \right)_{x=\bar{x}, y=\bar{y}} (x_i - \bar{x}) + \left(\frac{\partial q}{\partial y} \right)_{x=\bar{x}, y=\bar{y}} (y_i - \bar{y}) + O_2$$

Se obtiene:

$$\begin{aligned} \bar{q} &= \sum_{i=1}^k f_i q(x_i, y_i) = \sum_{i=1}^k f_i \left[q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x} (\bar{x} - \bar{x}) + \frac{\partial q}{\partial y} (\bar{y} - \bar{y}) + \dots \right] \\ &= q(\bar{x}, \bar{y}) \cancel{\sum_{i=1}^k f_i} + \frac{\partial q}{\partial x} \cancel{\sum_{i=1}^k f_i (\bar{x} - \bar{x})} + \frac{\partial q}{\partial y} \cancel{\sum_{i=1}^k f_i (\bar{y} - \bar{y})} \simeq q(\bar{x}, \bar{y}) \end{aligned}$$

1 0 0

Problemas

Problema 1.9: Pruébese que si $q = q(x, y)$ es una función cualquiera de las variables X e Y , y las variaciones de la media de ambas variables son pequeñas entonces se verifica

$$s_q^2 = \left(\frac{\partial q}{\partial x} \right)^2 s_x^2 + \left(\frac{\partial q}{\partial y} \right)^2 s_y^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \text{cov}(x, y)$$

Donde las derivadas deben entenderse evaluados en los valores medios de X e Y , y $\text{cov}(x, y)$ es la covarianza de las variables aleatorias X e Y : $\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^k \sum_{j=1}^l f_{ij}(x_i - \bar{x})(y_j - \bar{y})$

Por tanto:

$$\begin{aligned} s_q^2 &= \sum_{i=1}^k \sum_{j=1}^l f_{ij}(q(x_i, y_j) - \bar{q})^2 \simeq \sum_{i=1}^k \sum_{j=1}^l f_{ij}(q(x_i, y_j) - q(\bar{x}, \bar{y}))^2 = \\ &\sum_{i=1}^k \sum_{j=1}^l f_{ij} \left(q(\bar{x}, \bar{y}) + \left(\frac{\partial q}{\partial x} \right)_{x=\bar{x}} (x_i - \bar{x}) + \left(\frac{\partial q}{\partial y} \right)_{y=\bar{y}} (y_j - \bar{y}) + O_2 - q(\bar{x}, \bar{y}) \right)^2 = \\ &\sum_{i=1}^k \sum_{j=1}^l f_{ij} \left(\left(\frac{\partial q}{\partial x} \right)_{x=\bar{x}} (x_i - \bar{x}) + \left(\frac{\partial q}{\partial y} \right)_{y=\bar{y}} (y_j - \bar{y}) + O_2 \right)^2 = \\ &\left(\frac{\partial q}{\partial x} \right)^2 \sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y} \right)^2 \sum_{i=1}^k \sum_{j=1}^l f_{ij} (y_j - \bar{y})^2 + 2 \left(\frac{\partial q}{\partial x} \right) \left(\frac{\partial q}{\partial y} \right) \sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \\ &\left(\frac{\partial q}{\partial x} \right)^2 \sum_{i=1}^k f_{x_i} (x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y} \right)^2 \sum_{j=1}^l f_{y_j} (y_j - \bar{y})^2 + 2 \left(\frac{\partial q}{\partial x} \right) \left(\frac{\partial q}{\partial y} \right) \sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})(y_j - \bar{y}) \end{aligned}$$

Problemas

Problema 1.10: Considérese la función $y = kx$ de la variable aleatoria (siendo k un entero positivo). Calcular la varianza y la desviación típica de y en función de la varianza y la desviación típica de x , s_x^2 y s_x respectivamente, así como la covarianza siguiente: $\text{cov}_k[(x + y), (x - y)]$. Consideraremos ahora el caso de la variable aleatoria $y = \sum_{i=1}^k x_i$ donde $\{x_i\}_{i=1}^k$ es un conjunto de variables aleatorias independientes e idénticamente distribuidas, todas ellas con varianza s_x^2 . Repítanse los cálculos anteriores y compárense los resultados obtenidos.

Problemas

Problema 1.10: Considérese la función $y = kx$ de la variable aleatoria (siendo k un entero positivo). Calcular la varianza y la desviación típica de y en función de la varianza y la desviación típica de x , s_x^2 y s_x respectivamente, así como la covarianza siguiente: $\text{cov}[(x + y), (x - y)]$. Consideremos ahora el caso de la variable aleatoria $y = \sum_{i=1}^k x_i$ donde $\{x_i\}_{i=1}^k$ es un conjunto de variables aleatorias independientes e idénticamente distribuidas, todas ellas con varianza s_x^2 . Repítanse los cálculos anteriores y compárense los resultados obtenidos.

$$s_y^2 = \sum_{i=1}^k f_i(y_i - \bar{y})^2 = \sum_{i=1}^k f_i(kx_i - k\bar{x})^2 = k^2 \sum_{i=1}^k f_i(x_i - \bar{x})^2 = k^2 s_x^2$$

$$s_y = \sqrt{s_y^2} = ks_x$$

$$\begin{aligned} \text{cov}(x+y, x-y) &= \frac{1}{n} \sum_{i=1}^n (x_i + kx_i - \bar{x} - k\bar{x})(x_i - kx_i - \bar{x} + k\bar{x}) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i + kx_i - \bar{x} - k\bar{x})(x_i - kx_i - \bar{x} + k\bar{x}) = \frac{1}{n} \sum_{i=1}^n (1+k)(x_i - \bar{x})(1-k)(x_i - \bar{x}) \\ &= (1-k)(1+k) \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = (1-k^2)s_x^2 \end{aligned}$$

O de un ejercicio anterior:

$$\text{cov}(x+y, x-y) = s_x^2 - s_y^2 = s_x^2 - k^2 s_x^2 = (1-k^2)s_x^2$$

Problemas

Problema 1.10: Considérese la función $y = kx$ de la variable aleatoria (siendo k un entero positivo). Calcular la varianza y la desviación típica de y en función de la varianza y la desviación típica de x , s_x^2 y s_x respectivamente, así como la covarianza siguiente: $\text{cov}[(x + y), (x - y)]$. Consideremos ahora el caso de la variable aleatoria $y = \sum_{i=1}^k x_i$ donde $\{x_i\}_{i=1}^k$ es un conjunto de variables aleatorias independientes e idénticamente distribuidas, todas ellas con varianza s_x^2 . Repítanse los cálculos anteriores y compárense los resultados obtenidos.

$$\text{Si } y = \sum_{i=1}^k x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k x_j^i = \sum_{j=1}^k \left[\frac{1}{n} \sum_{i=1}^n x_j^i \right] = k \bar{x}$$

$$y_1 = \sum_{i=1}^k x_i^1$$

$$s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^k x_j^i - k \bar{x} \right]^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^k (x_j^i - \bar{x}) \right)^2$$

$$y_2 = \sum_{i=1}^k x_i^2$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^k (x_j^i - \bar{x})^2 + 2 \sum_{l < m} (x_l^i - \bar{x})(x_m^i - \bar{x}) \right)$$

$$\dots$$

$$= \sum_{j=1}^k \frac{1}{n} \sum_{i=1}^n (x_j^i - \bar{x})^2 + 2 \sum_{l < m} \frac{1}{n} \sum_{i=1}^n (x_l^i - \bar{x})(x_m^i - \bar{x}) \xrightarrow{0}$$

$$y_n = \sum_{i=1}^k x_i^n$$

$$s_y = \sqrt{s_y^2} = \sqrt{k} s_x$$

0

Problemas

Problema 1.10: Considérese la función $y = kx$ de la variable aleatoria (siendo k un entero positivo). Calcular la varianza y la desviación típica de y en función de la varianza y la desviación típica de x , s_x^2 y s_x respectivamente, así como la covarianza siguiente: $\text{cov}[(x + y), (x - y)]$. Consideraremos ahora el caso de la variable aleatoria $y = \sum_{i=1}^k x_i$ donde $\{x_i\}_{i=1}^k$ es un conjunto de variables aleatorias independientes e idénticamente distribuidas, todas ellas con varianza s_x^2 . Repítanse los cálculos anteriores y compárense los resultados obtenidos.

$$\text{Si } y = \sum_{i=1}^k x_i$$

$$y_1 = \sum_{i=1}^k x_i^1$$

$$\text{cov}(x+y, x-y) = s_{x_i}^2 - s_y^2 = s_x^2 - ks_x^2 = (1-k)s_x^2$$

$$y_2 = \sum_{i=1}^k x_i^2$$

...

$$y_n = \sum_{i=1}^k x_i^n$$

Problemas

Problema 1.11: Considérense los resultados del experimento de medición de la masa atómica del elemento del ejercicio 3. Histográmense las distribuciones correspondientes a las siguientes variables aleatorias:

$$y_i = x_i^2 \quad y_i = \ln x_i \quad y_i = \sqrt{x_i}$$

Calcúlense las medidas características de las anteriores distribuciones de manera exacta. Compárense la media y la desviación típica obtenidas con las resultantes de aplicar las aproximaciones de las transformaciones en torno a la media.

Problemas

Problema 1.12: Dada la variable bidimensional $\vec{X} = (X, Y)$ con la tabla de frecuencias:

Se pide:

$$\text{Calcúlese } \sum_i \sum_j x_{ij}$$

x\y	1	2	4	6
1	2	0	0	1
3	3	1	0	1
5	0	1	0	5

Frecuencia relativas f_{ij}

Las distribuciones marginales n_{x_i} n_{y_i} . Obténganse las medias de las variables X e Y.

Obténgase:

a) Los momentos respecto al origen $M_{1,0}(0,0)$ y $M_{0,1}(0,0)$

b) El momento respecto la media (central) $M_{1,1}(\bar{x}, \bar{y})$. ¿A que corresponde este momento?

Problemas

Problema 1.12: Dada la variable bidimensional $\vec{X} = (X, Y)$ con la tabla de frecuencias:

Se pide:

Calcúlese $\sum_i \sum_j x_{ij}$

$x \setminus Y$	1	2	4	6
1	2	0	0	1
3	3	1	0	1
5	0	1	0	5

$$\sum_i \sum_j x_{ij} = 14$$

Frecuencia relativas f_{ij}

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{14}$$

Las distribuciones marginales n_{x_i} n_{y_i} . Obténganse las medias de las variables X e Y.

$x \setminus Y$	1	2	4	6	n_{xi}
1	2	0	0	1	3
3	3	1	0	1	5
5	0	1	0	5	6
n_{yj}	5	2	0	8	14

$$\bar{x} = \frac{48}{14} = 3,4286$$

$$\bar{y} = \frac{51}{14} = 3,6429$$

Problemas

Problema 1.12: Dada la variable bidimensional $\vec{X} = (X, Y)$ con la tabla de frecuencias:

$x \setminus Y$	1	2	4	6
1	2	0	0	1
3	3	1	0	1
5	0	1	0	5

Se pide:

Obtégase:

a) Los momentos respecto al origen $M_{1,0}(0,0)$ y $M_{0,1}(0,0)$

b) El momento respecto la media (central) $M_{1,1}(\bar{x}, \bar{y})$. ¿A que corresponde este momento?

a) Corresponden a la media de cada variable encontrada anteriormente.

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

b) A la covarianza:

$$\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^k \sum_{j=1}^l f_{ij} (x_i - \bar{x})(y_j - \bar{y}) = 1,8673$$

Lo que implica una correlación positiva entre las variables.

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

- a) Tabla bidimensional.
- b) Distribuciones marginales.
- c) Distribución de **X** condicionada a **Y** = 200
- d) Media, mediana, moda y cuartiles de ambas variables aleatorias.
- e) Recorrido, desviación típica, desviación media (respecto a la media) y coeficiente de variación de Pierson de **X** e **Y**.
- f) ¿Es simétrica la variable **X**? ¿Y la **Y**? g) Curtosis de ambas distribuciones.
- h) ¿Que media es menos representativa?
- I) Porcentaje de años en los que el precio del kilogramo de uva fue inferior a 0,27 €.
- j) Recta de regresión y coeficiente de correlación. Matriz de covarianzas.

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

- a) Tabla bidimensional. b) Distribuciones marginales.

	x_i		y_i
(15,25)	20	(100,140)	120
(27,35)	30	(140,180)	160
(35,45)	40	(180,200)	200

	120 - 160	200	n_{x_i}
x	1	1	2
20	1	1	2
30	1	2	5
40	2	0	3
n_y	4	3	10

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

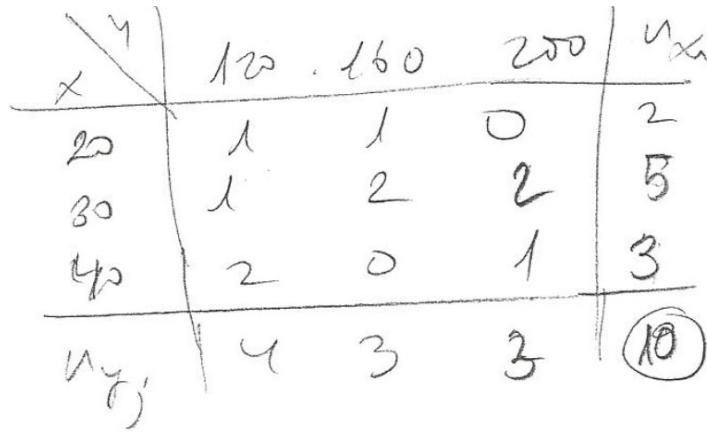
c) Distribución de **X** condicionada a **Y = 200**

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

$$n(X=20 | Y=200) = 0$$

$$n(X=30 | Y=200) = 2$$

$$n(X=40 | Y=200) = 1$$



$$f(x_i | y_j) = \frac{n(x_i | Y=y_j)}{n_{y_j}} = \frac{n_{ij}}{n_{y_j}} = \frac{f_{ij}}{f_{y_j}}$$

$$f(X=20 | Y=200) = \frac{n(X=20 | Y=200)}{n_{Y=200}} = 0$$

$$f(X=30 | Y=200) = \frac{n(X=30 | Y=200)}{n_{Y=200}} = \frac{2}{3}$$

$$f(X=40 | Y=200) = \frac{n(X=40 | Y=200)}{n_{Y=200}} = \frac{1}{3}$$

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

d) Media, mediana, moda y cuartiles de ambas variables aleatorias.

La media es: $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = 31,0$ $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i = 156,0$

La mediana en X será $Me_x = 30$ y la moda también. El primer cuartil será 30 y el tercero 40 en X.

La mediana en Y será $Me_y = 160$ y la moda $Md_y = 120$. El primer cuartil será 120 y el tercero 200 en Y.

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

- e) Recorrido, desviación típica, desviación media (respecto a la media) y coeficiente de variación de Pierson de **X** e **Y**.

El recorrido de la variable **X** es [19,37] y el de la variable **Y** [100,200]

$$s_x^2 = \sum_{i=1}^3 f_{x_i} (x_i - \bar{x})^2 = 49 \quad s_x = 7$$

$$s_y^2 = \sum_{i=1}^3 f_{y_i} (y_i - \bar{y})^2 = 1104 \quad s_y = 33,23$$

$$DM_{\bar{x}} = \sum_{i=1}^3 f_{x_i} |x_i - \bar{x}| = 5,4$$

$$DM_{\bar{y}} = \sum_{i=1}^3 f_{y_i} |y_i - \bar{y}| = 28,8$$

$$CV_x = \frac{s_x}{|\bar{x}|} = \frac{7}{31} = 0,226$$

$$CV_y = \frac{s_y}{|\bar{y}|} = \frac{33,23}{156} = 0,213$$

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

- f) ¿Es simétrica la variable **X**? ¿Y la **Y**? g) Curtosis de ambas distribuciones.

Esta vez (ya que es unimodal y no lo he realizado antes) dare el coef. De asimetría de Pearson:

$$A_{Px} = \frac{\bar{x} - Md_x}{S_x} = \frac{31 - 30}{9} = 0,11$$

$$A_{Py} = \frac{\bar{y} - Md_y}{S_y} = \frac{156 - 120}{33,23} = 1,083$$

$$g_x = \frac{1}{S_x^4} \sum_{i=1}^k f_i (x_i - \bar{x})^4 = \frac{m_4(\bar{x})}{S_x^4} = 2,04$$

$$g_y = \frac{1}{S_y^4} \sum_{i=1}^k f_i (y_i - \bar{y})^4 = \frac{m_4(\bar{y})}{S_y^4} = 1,47$$

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

h) ¿Que media es menos representativa?

Al ser $CV_x > CV_y$, es menos representativa la de X

I) Porcentaje de años en los que el precio del kilogramo de uva fue inferior a 0,27 €.

$f(20) = 2/10 \rightarrow$ el 20% de los años medidos

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

j) Recta de regresión y coeficiente de correlación. Matriz de covarianzas. $Y = a_0 + b_0 X$

De forma práctica es mejor usar la matriz M y resolver:

$$\begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$a = \frac{1}{\det M} \begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i \end{vmatrix} \quad \sum x_i = 301 \quad \sum x_i^2 = 9385 \quad \sum y_i = 1500 \quad \sum x_i y_i = 45510$$

$$b = \frac{1}{\det M} \begin{vmatrix} N & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}$$

$$b_0 = \frac{\bar{xy}}{\bar{x}^2 - (\bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2}$$

$$a_0 = \bar{y} - \frac{\bar{xy}}{\bar{x}^2 - (\bar{x})^2} \bar{x} = \bar{y} - \frac{\text{cov}(x, y)}{s_x^2} \bar{x}$$

$$\det M = \begin{vmatrix} 10 & 301 \\ 301 & 9385 \end{vmatrix} = 3249$$

Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

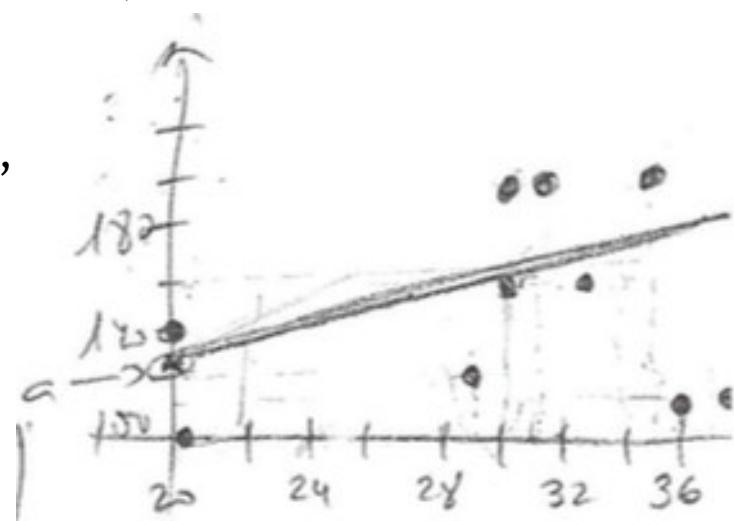
donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

j) Recta de regresión y coeficiente de correlación. Matriz de covarianzas. $Y = a_0 + b_0 x$

De forma práctica es mejor usar la matriz M y resolver:

$$a = \frac{1}{\det M} \begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix} = \frac{1}{3249} \begin{vmatrix} 1500 & 301 \\ 45510 & 9385 \end{vmatrix} = \frac{378990}{3249} = 116,65$$

$$b = \frac{1}{\det M} \begin{vmatrix} N & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix} = \frac{1}{3249} \begin{vmatrix} 10 & 1500 \\ 301 & 45510 \end{vmatrix} = \frac{3600}{3249} = 1,$$



Problemas

Problema 1.13: En una determinada explotación vitivinícola se observaron durante algunos años el precio del kilogramo de uva y la cantidad de producción, obteniéndose la siguiente tabla:

X	21	19	29	36	31	29	37	31	33	35
Y	100	140	120	110	200	200	110	160	160	200

donde **X** es el precio del kilogramo en céntimos de euro e **Y** la cantidad producida en miles de kilogramos. Considerando la variable **X** agrupados en intervalos de amplitud constante y considerando el primero de ellos [15,25] se pide:

j) Recta de regresión y coeficiente de correlación. Matriz de covarianzas. $Y = a_0 + b_0 X$

De forma práctica es mejor usar la matriz M y resolver:

$$\text{cov}(x, y) = m_{1,1}(\bar{x}, \bar{y}) = \sum_{i=1}^k \sum_{j=1}^l f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = 17,6$$

$$M = \begin{pmatrix} s_x^2 & \text{cov}(x, y) \\ \text{cov}(y, x) & s_y^2 \end{pmatrix} = \begin{pmatrix} 49 & 17,6 \\ 17,6 & 33,23 \end{pmatrix}$$

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{17,6}{\sqrt{49} \cdot \sqrt{33,23}} = 0,076$$

Notas varias según voy escribiendo...