**Temas de Estadística Práctica**
**Antonio Roldán Martínez**

**Proyecto http://www.hojamat.es/**

**Tema 2**

**Medidas de tipo paramétrico**



## Medidas de tendencia central

Se llaman medidas de tendencia central, de posición o *promedios* de una distribución de datos a aquellos valores que indican el centro de la distribución, y pretenden representar todos los datos en un solo punto.

### Media

Llamaremos *media aritmética* o simplemente *media* al valor resultante de sumar todos los datos y después dividir el resultado entre el número de ellos. Es, pues, el resultado de un reparto igualitario de los valores. También se puede interpretar como el *centro de gravedad* de los datos.

Su fórmula más simple es

$$\bar{x} = \frac{\sum x}{N}$$

(suprimimos los subíndices en los sumatorios cuando no exista peligro de confusión)

Si los datos están agrupados en una tabla de frecuencias, la fórmula anterior se convertiría en

$$\bar{x} = \frac{\sum x \cdot n}{N} = \sum x \cdot f$$

que es la más usada en la práctica. Obsérvese que el uso de frecuencias relativas simplifica bastante su cálculo.

La suma de desviaciones de todos los datos respecto de la media es cero

$$\sum (x - \bar{x}) = 0$$

y la suma de los cuadrados de las desviaciones es la **mínima** respecto a la que resultaría al elegir otro valor cualquiera en la resta

$$\sum (x - \bar{x})^2 \leq \sum (x - k)^2$$

para cualquier valor de **k**.

La anterior suma se puede expresar también como

$$\sum (x - \bar{x})^2 = \sum x^2 - N \cdot \bar{x}^2$$

La media es propia sólo de datos cuantitativos. Si estos están agrupados, se establece la hipótesis de que están todos situados en la **marca de la clase** o punto medio.

La media es sensible a todos los datos. Si uno de ellos cambia, la media también se ve alterada. Por esta razón, influyen mucho en ella los valores extremos, que suelen ser poco fiables, por lo que a veces se desechan.

Su importancia radica en que es base de muchas técnicas estadísticas, pero no nos vale cuando existen intervalos abiertos o la medida solo tiene carácter ordinal.

## Otras medias

También podemos usar

## Media geométrica

Es la raíz enésima del producto de los datos. Se usa cuando el producto es más representativo que la suma, como ocurre cuando se promedian cocientes o razones.

$$mg = \sqrt[N]{\prod x} = e^{medLog(x)}$$

## Media armónica

Es la media diseñada para promediar cantidades inversamente proporcionales y equivale al inverso de la media de los inversos de x

$$xa = \frac{1}{\sum \left(1/x\right) / N}$$

## Media cuadrática

Es muy usada en la teoría de errores y en estudios sobre ajustes de datos. Es la raíz cuadrada de la media de los cuadrados de los datos.

$$mc = \sqrt{\frac{\sum x^2}{N}}$$

## Media ponderada

Es muy interesante en la enseñanza, para promediar calificaciones con distintos pesos, además de múltiples utilidades en problemas de mezclas, centros de gravedad, cestas de inversiones, etc.

Se calcula asignando a cada dato un *peso*, cuya significación depende de cada problema, y suponer que cada dato se repite tantas veces como indica su peso (aunque este no sea entero. Esa es la diferencia entre peso y frecuencia). Se procede pues a multiplicar cada dato por su peso, para dividir después entre la suma de los pesos.

$$mp = \frac{\sum x \cdot p}{\sum p}$$

Mediana

Llamaremos *mediana* de un conjunto de datos de tipo ordinal (o de intervalo o razón) al dato que ocupa el punto medio de la distribución ordenada de datos. Es decir, es el punto que divide a la distribución en dos partes iguales: el total de frecuencias de los datos inferiores a la mediana es igual al de las frecuencias de los datos mayores. Si los datos son continuos o muy numerosos, se puede afirmar con cierta aproximación que ese total de frecuencias es el 50%.

En el cálculo de la mediana sólo se utiliza el **orden** de los datos y no su magnitud. Por eso es la medida adecuada para escalas de tipo ordinal, como las que se usan en Psicología y Ciencias de la Educación.

La mediana no es una medida bien establecida, pues contiene elementos convencionales. Tanto es así que en algunos casos su definición cambia de unos manuales a otros.

Podemos establecer algunas convenciones para su cálculo:

**Datos aislados**

Si el número de datos es **impar**, se toma como mediana el dato central, que deja (n-1)/2 datos menores que él y otros tantos mayores: La mediana de 2 2 2 3 3 3 4 5 6 7 7 8 9 es **4**.

Si el número de datos es **par**, se toma como mediana el promedio de los dos datos centrales: la mediana de 2 3 4 5 6 8 9 10 es **5,5**.

Algunos autores dan reglas más precisas para el caso en el que los valores centrales están repetidos. En caso de duda es preferible agrupar los datos por frecuencias y usar la teoría correspondiente.

## Datos agrupados

Si los datos están agrupados por intervalos, se usa la siguiente fórmula (que es fácil de justificar mediante proporciones si se supone que todos los datos están distribuidos de manera uniforme en el intervalo)

$$Me = L_i + \frac{(N/2 - n_{ant})}{n_{interv}} \cdot Ampl.$$

En la fórmula los símbolos se interpretan así:

*N/2* es la mitad del número de datos, con decimales si los hubiere.

*Ampl.* es la amplitud del intervalo mediano, el que contiene la frecuencia acumulada 0,5.

$n_{ant}$ es la frecuencia acumulada anterior al intervalo mediano.

$n_{interv.}$ es la frecuencia absoluta de dicho intervalo.

$L_i$ representa el límite inferior verdadero del intervalo mediano. Este concepto es muy importante, pues en caso de datos agrupados por frecuencias, pero no por intervalos, los límites verdaderos de un valor, por ejemplo 23, serían 22,5 y 23,5.

## Propiedades de la mediana

La mediana es menos sensible a datos extremos que la media, por lo que se pueden considerar estos con menos peligro de sesgo en los resultados. También es muy estable para pequeñas variaciones en los datos.

Es muy útil si los datos son ordinales, o excesivamente asimétricos, o si en los intervalos existe alguno abierto, del tipo *60 o más*.

Cumple que la suma de los valores absolutos de las desviaciones respecto a ella es mínima

$$\sum |x - me|$$

es mínimo

Moda

Es la más pobre de las tres medidas y la más convencional. Llamaremos **Moda** al valor de la distribución de datos que presente una frecuencia mayor. Así, en el conjunto 2 2 3 3 3 4 4 5 6 6 la moda es 3, valor más repetido.

La definición se vuelve ambigua si existen varios intervalos con la misma frecuencia. En estos casos se distingue:

Valores separados: se consideran **moda** todos los valores de frecuencia máxima, y la distribución recibe el nombre de **bimodal, trimodal,** etc. Por ejemplo en el conjunto 3 3 4 4 4 5 6 7 7 7 8 8 9 las dos modas son 4 y 7 y la distribución será bimodal.

Valores consecutivos: Se usa, por convención la siguiente fórmula:

$$Mo = L_i + \frac{d_1}{d_1 + d_2} \cdot Ampl.$$

en la que $L_i$ representa el límite inferior verdadero del intervalo, $d_1$ la diferencia de frecuencia con el intervalo anterior, la $d_2$ la diferencia con el siguiente y *Ampl.* la amplitud del intervalo.

La **moda** es una medida sencilla pero poco representativa del conjunto de datos. Es la única posible en datos cualitativos, pero se puede usar en intervalos abiertos. Su mayor inconveniente es la falta de unicidad en su definición y su sesgo en distribuciones muy asimétricas. Por el contrario, es bastante estable frente a variaciones de los datos.

## Medidas de variabilidad

Las medidas de variabilidad nos informan sobre el grado de concentración o dispersión que presentan los datos respecto a su promedio. Llamaremos ***homogénea***, concentrada o poco dispersa a aquella distribución en la que todos los datos están cercanos al centro, como 4 4 5 5 5 5 6 6 6 6 7, y ***heterogénea*** o dispersa a la distribución con datos más separados del centro, como 1 3 5 8 10 16 20.

Existen muchas formas de medir la variabilidad. Destacaremos las más importantes.

Rango

También llamado ***Recorrido*** o ***Amplitud total***, es la diferencia entre el máximo valor del conjunto de datos y el mínimo de ellos. A mayor rango, mayor dispersión.

El rango del conjunto 4 6 4 7 8 6 5 3 4 7 7 9 6 5 es 6, la diferencia entre el máximo 9 y el mínimo 3.

A veces se usa el ***Rango verdadero*** que consiste en considerar cada dato rodeado de una unidad, por efecto de los redondeos, con lo que en el ejemplo anterior el mínimo sería 2,5 y el máximo 9,5. Con ello el rango se convertiría en 7.

No es una medida buena, pues ignora todo lo que ocurre dentro de ese rango.

Desviación media

Es una medida de la dispersión consistente en la media aritmética de las desviaciones individuales respecto a la media, tomadas en valor absoluto. También se usan desviaciones respecto a la mediana.

Su fórmula, pues, será:

$$DM = \frac{\sum n_i \cdot |x_i - \bar{x}|}{N}$$

No es muy útil, pues carece de propiedades importantes. Además, tiene el inconveniente de que no es derivable.

Varianza

Si en la fórmula anterior sustituimos los valores absolutos por cuadrados (es otra forma de convertirlos en positivos), obtendremos la ***Varianza $s^2$***. Su fórmula será:

$$s^2 = \frac{\sum n_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum x_i^2 \cdot n_i}{N} - \bar{x}^2$$

Es fácil demostrar la equivalencia de las dos fórmulas.

Si los datos están aislados basta suprimir las frecuencias $n_i$ de las fórmulas.

Es una medida muy sensible de la variabilidad y base de muchas técnicas estadísticas. Junto con la media forma el conjunto más importante de medidas.

Es propia de las medidas de intervalo o razón. Su inconveniente es que no usa la misma unidad que los datos, sino su cuadrado.

No se deben comparar varianzas en conjuntos de unidades muy distintas, como estatura e inteligencia.

En teoría del muestreo se sustituye por la ***cuasivarianza***, de idéntica fórmula, pero con cociente N-1 en lugar de N. En este caso no sería válida la segunda fórmula. En otro momento trataremos

de ella.

Desviación típica

Es la raíz cuadrada de la anterior. Su objeto es conseguir medir la variabilidad en las mismas unidades que los datos. Así, un conjunto medido en metros, tendrá la varianza medida en metros cuadrados, pero la desviación típica en metros. Tiene como fórmula

$$s = \sqrt{\frac{(x_i - \bar{x})^2 \cdot n_i}{N}} = \sqrt{\frac{\sum x_i^2 \cdot n_i}{N} - \bar{x}^2}$$

Como en la varianza, para datos aislados basta con suprimir las frecuencias $n_i$.

La desviación típica ***s*** es base de muchas técnicas, al igual que la media y la varianza. Su gran ventaja es estar medida en las mismas unidades que los datos y la media, lo que permite establecer razones y proporciones entre ellas.

La desviación típica cumple la llamada ***desigualdad de Tchebychev***:

$$Fr\left(|x_i - \bar{x}| \leq ks\right) \geq 1 - \frac{1}{k^2}$$

según la cual, los datos que se alejan de la media una distancia igual o menor que **s** multiplicado por un coeficiente **k** suponen más de la proporción $1\text{-}1/k^2$. Así, el 75% de los datos al menos, se encuentra a menos de dos desviaciones típicas y el 89% a menos de tres.

En el estudio de las distribuciones teóricas podremos precisar más estas acotaciones.

Coeficiente de variación

Cuando se comparan conjuntos de medias muy distintas (como podrían ser los diámetros de los planetas y la altura de mis alumnos) no sirve de nada comparar las distintas variabilidades. Entre las dos desviaciones típicas existiría una diferencia enorme en magnitud. Por ello, se suele corregir la desviación típica dividiéndola entre su media. De esta forma obtenemos una medida *relativa* de la variablidad, que permite las comparaciones.

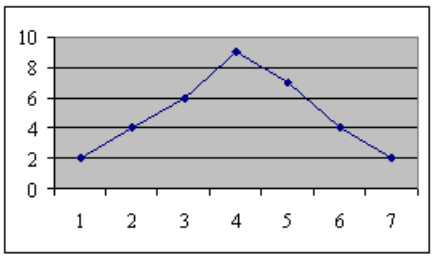
$$CV = \frac{\bar{x}}{s}$$

## Medidas de asimetría

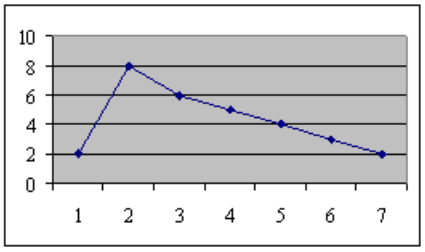
La ***asimetría*** o ***sesgo*** de una distribución es la característica por la que los datos pierden su simetría respecto a la media. Expresado de otra forma, es el mayor o menor grado de desviación que existe entre la media (reparto equitativo) y la mediana (punto medio de la distribución).

Si en un conjunto coinciden media y mediana, se presenta una ***simetría*** y cuanto más se separen, mayor será la asimetría de la distribución.

Será simétrica (aproximadamente) la distribución




y asimétrica esta otra




La distribución anterior, en la que existen muchas medidas bajas y pocas altas, diremos que presenta una ***asimetría positiva***. Si ocurriera lo contrario, diríamos que era ***asimétrica negativa***.

## Índices de asimetría

En las distribuciones asimétricas positivas la media es mayor que la moda (punto más alto de la gráfica), y lo contrario ocurre en las negativas. Por eso Pearson sugirió medir la asimetría mediante su diferencia dividida entre la desviación típica.

$$A_{p1} = \frac{\bar{x} - Mo}{s}$$

pero como existe una ley empírica por la cual la diferencia entre la media y la moda es el triple de la existente entre media y mediana, propuso también

$$A_{p2} = \frac{3(\bar{x} - Me)}{s}$$

Más precisa es la medida de Fisher, porque usa ***momentos de tercer orden***, es decir, los cubos de las desviaciones respecto a la media. El índice de Fisher no tiene unidades, es una razón o comparación, y su fórmula es

$$g_1 = \frac{\sum (x - \bar{x})^3 \cdot n_i}{s^3}$$

Será positivo o negativo cuando la asimetría también lo sea.

Existe otro índice, el de Bowley, que estudiaremos en otro momento.

## Medidas de aplastamiento o curtosis

Independientemente de su asimetría, una distribución puede presentar los datos con un reparto más uniforme, en el que las frecuencias sean muy parecidas. El gráfico aparecerá como aplastado y diremos que la distribución es ***platicúrtica*** o de ***poca curtosis***. En el otro extremo, si las frecuencias cercanas al centro son mayores (con diferencia) que las alejadas, diremos que es ***leptocúrtica*** o con ***gran curtosis***. Al caso intermedio lo denominaremos como distribución ***mesocúrtica***.


Para la medida del aplastamiento se usa otro índice de Fisher que usa el ***momento de cuarto orden***.

$$g_2 = \frac{\sum (x_i - \bar{x})^4 \cdot n_i}{s^4} - 3$$

En las leptocúrticas este índice es positivo, y negativo en las platicúrticas. Como veremos más adelante, en la distribución normal es nulo.