TEMA 3: MEDIDAS QUE CARACTERIZAN UNA DISTRIBUCIÓN DE FRECUENCIAS

3.1.- MEDIDAS DE POSICIÓN

3.2.- MEDIDAS DE DISPERSIÓN

3.3.- MEDIDAS DE FORMA

3.4.- MEDIDAS DE CONCENTRACIÓN

La información suministrada por una tabla de distribución de frecuencias (Xi,ni) se puede resumir en un conjunto de medidas que la caractericen.

Fines de dichas medidas:

- Síntesis de la información (proporcionar un resumen cuantitativo de la variable estudiada)
- Facilitar la comparación entre distribuciones diferentes.

Propiedades que deberían cumplir a priori estas medidas:

- Intervención en su determinación de todos y cada uno de los valores de la distribución
- Ser siempre calculable
- Ser única para cada distribución de frecuencias

Clasificación:

Medidas de posición: sitúan una distribución de frecuencias

Medidas de dispersión: miden el grado de esparcimiento de los datos

Medidas de concentración: miden el grado de igualdad en el reparto de los valores de la variable.

Medidas de forma: aspecto de la distribución.

3.1.- MEDIDAS DE POSICIÓN

Serán valores de la variable sobre los cuales tienden a agruparse los datos, según diferentes criterios. Nos dan una idea de la situación de los valores que se han recogido de la variable.

Las medidas de posición serán, pues, valores medidos en las mismas unidades que los datos y nos indicarán la posición en torno a la cual se distribuyen nuestras observaciones (son promedios y pueden ser de tendencia central o no). Según que estos valores nos orienten sobre el lugar central de la distribución o nos indiquen la posición de una parte cualquiera predeterminada de ella, las medidas descriptivas de posición se llamarán de posición central (medias, mediana, moda) o de posición no central (cuartiles, deciles y percentiles).

Todas las medidas de posición deben verificar lógicamente, que su valor esté comprendido entre el menor valor y el mayor valor de los datos de la distribución.

MEDIDAS DE POSICIÓN CENTRAL

1.- MEDIA ARITMÉTICA.

Parece lógico elegir como medida de posición central de una distribución, aquel valor que fuese como su centro de gravedad, es decir, que compensase las desviaciones con respecto de él de los valores de la variable, en un sentido o en otro.

Se define como la suma de todos los valores de la distribución dividida por el número total de observaciones.

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

Ejemplo: Si las puntuaciones de 4 exámenes de un alumno son 5,7,3 y 6 puntos, la media aritmética de estas notas será 5+7+3+6/4 = 21/4 = 5,25 puntos. Como vemos la media aritmética, al igual que las demás medidas de posición que veremos, viene expresada en las mismas unidades que los datos originales. Obsérvese también que la media aritmética puede ser interpretada como aquel valor de la variable, tal que si todos los elementos observados poseyesen su valor en lugar del que realmente tienen, entonces la suma total de los recursos repartidos en la distribución sería la misma. Así, en el ejemplo anterior, si en los 4 exámenes el alumno hubiese tenido 5,25 puntos, en total entre los 4 exámenes habría tenido los mismos 21 puntos que obtuvo. Este sería el caso de una distribución no agrupada en intervalos con frecuencias unitarias.

Ejemplo de distribuciones no agrupadas en intervalos con frecuencias (en clase).

En el caso de una distribución agrupada en intervalos, nos conduce a una pérdida de información (la pérdida del conocimiento exacto e individual de todos y cada uno de los valores de la variable correspondiente a los datos originales). Así, cuando trabajamos con este tipo de distribuciones, no obtendremos, generalmente, el valor exacto de la media aritmética de la distribución, sino un valor aproximado, pero que, en cualquier caso, suele ser bastante aceptable y, por supuesto, desde el punto de vista práctico satisfactoria al habernos simplificado, por otra parte, la dura tarea de trabajar con los datos iniciales. En esta situación, se supone que en cada intervalo de la distribución., las frecuencias se encontrarán repartidas de forma uniforme a lo largo del intervalo, lo que, como consecuencia, da lugar a que el valor medio de cada intervalo coincida exactamente con el punto medio del mismo, y que hemos llamado en el tema anterior como marca de clase o del intervalo correspondiente. En este caso, los valores de la variable, representados por Xi, corresponden a la marca de clase. Destaca también el hecho de que para poder disponer de la marca de clase de cada intervalo, se requiere que los intervalos estén perfectamente determinados por unos extremos concretamente definidos. Así, pues, no podríamos calcular la media de una distribución de datos que nos midiera el nº de habitantes de los municipios de una provincia si el grupo de municipios más poblados estuviese definido ambiguamente, diciendo sólo, por ejemplo que tiene más de 200.000 habitantes.

Propiedades de la media aritmética:

1.- La suma de las desviaciones de los valores de la variable respecto a su media es cero

$$\sum_{i=1}^{n}(x_i-\bar{x})\,n_i=\sum x_in_i-\bar{x}\sum_{i=1}^{n}n_i=N\frac{\sum_{i=1}^{n}x_in_i}{N}-N\bar{x}=N\bar{x}-N\bar{x}=0$$

2.- La media de las desviaciones al cuadrado de los valores de la variable respecto a una constante k cualquiera se hace mínima cuando esa constante k es igual a la media aritmética.

$$\sum_{i=1}^{n}(x_i-k)^2\frac{n_i}{N}$$

3.- Si a todos los valores de una variable le sumamos una constante k la media aritmética queda también aumentada en esa constante, es decir, que la media aritmética queda afectada por los cambios de origen.

$$Sea\ (xi,ni)\ \ y\ (xi',ni)\ donde\ xi'=xi+k\ entonces$$

$$\bar{x}'=\frac{\sum_{i=1}^{n}x_i'n_i}{N}=\sum_{i=1}^{n}\frac{(xi+k)n_i}{N}=\frac{\sum_{i=1}^{n}x_in_i}{N}+k\frac{\sum n_i}{N}=\bar{x}+k$$

4.- Si a todos los valores de una variable les multiplicamos por una constante k, su media aritmética también queda multiplicada por esa constante, es decir que la media aritmética también queda afectada por los cambios de escala.

$$Sea\ (xi,ni)\ \ y\ (xi',ni)\ donde\ xi'=xi*k\ entonces$$

$$\bar{x}'=\frac{\sum_{i=1}^{n}x_i'n_i}{N}=\sum_{i=1}^{n}\frac{x_ikn_i}{N}=k\frac{\sum_{i=1}^{n}x_in_i}{N}=k*\bar{x}$$

5.- Si de un conjunto de valores obtenemos 2 o más conjuntos disjuntos la media aritmética de todo el conjunto se relaciona con la media aritmética de los diferentes subconjuntos disjuntos. Esta propiedad está relacionada con una variante de la media aritmética conocida como media aritmética ponderada, de aplicación en aquellas circunstancias en las que se conoce que los valores de la variable no tienen todos la misma importancia para su tratamiento, sino que, por el contrario, existen observaciones que deben ser consideradas como más representativas que otras. Para su cálculo se le asocia a cada valor de la variable Xi un peso o ponderación (Wi), que nos medirá su grado de importancia o representatividad dentro de la distribución. Estos pesos Wi, serán valores positivos que representarán el nº de veces que sus correspondientes valores Xi son más representativos que un valor que tuviese peso asociado la unidad. Encontramos varios ejemplos de medias ponderadas en los números índices (que veremos en el tema 6). Hay que destacar que cuando se emplean ponderaciones proporcionales, la suma de las mismas es la unidad, y esto nos evita la división por su suma.

Subconjunto 1, i=1,2,....j con tamaño N_1

Subconjunto 2, i = j+1, ..., n con tamaño N_2 siendo $N_1 + N_2 = N$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{N_1 \frac{\sum_{i=1}^{j} x_i n_i}{N_1} + N_2 \frac{\sum_{i=j+1}^{n} x_i n_i}{N_2}}{N} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N}$$

$$\bar{x}_p = \frac{\sum_{i=1}^{n} x_i w_i}{\sum w_i}$$

<u>Ventajas e inconvenientes de la media aritmética</u>:

Ventajas:

- Ser calculable
- Es única y está perfectamente definida de forma objetiva.
- Para su cálculo se utilizan todos los valores de la distribución.
- Tiene un claro significado interpretativo.

Inconvenientes:

El que los valores extremos muy dispares influyen de forma notable en su valor, haciéndola menos representativa.

La media aritmética, como medida de posición, es la fórmula más adecuada para el resumen estadístico en caso de distribuciones en escala de intervalos o de proporción, con las cuales dicha medida alcanza su máximo sentido.

<u>2.- MEDIA GEOMÉTRICA (G)</u>

Sea una distribución de frecuencias (Xi,ni). La media geométrica (G), se define como la raíz de índice la frecuencia total (nº total de datos) N, y cuyo radicando es el producto de las potencias de cada variable, elevada a su respectiva frecuencia absoluta.

$$G = \sqrt[N]{{x_1}^{n_1} * {x_2}^{n_2} *** \ \ {x_n}^{n_n}} = \sqrt[N]{\prod_{i=1}^{n} {x_i}^{n_i}} = ({x_1}^{n_1} * {x_2}^{n_2} *** \ \ {x_n}^{n_n})^{\frac{1}{N}}$$

Propiedad:

El logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

$$\log G = log \sqrt[N]{\prod_{i=1}^{n} {x_i}^{n_i}} = \frac{1}{N} log \left(\prod_{i=1}^{n} {x_i}^{n_i} \right) = \frac{1}{N} \sum_{i=1}^{n} log\, {x_i}^{n_i} = \frac{1}{N} \sum_{i=1}^{n} (log\, xi) * n_i$$

Ventajas:

- En su determinación intervienen todos los valores de la distribución.
- Es menos sensible que la media aritmética a los valores extremos, por su carácter de producto.
- Está definida de forma objetiva y es única.

Inconvenientes:

- Es de significado estadístico menos intuitivo que la media aritmética.
- Su cálculo es más difícil.
- En ocasiones no queda determinada. Algún valor Xi = 0, hace que la media geométrica se anule.

Su empleo más frecuente es el de promedia porcentajes, tasas, números índices, etc, es decir en los casos en que se supone que la variable presenta variaciones acumulativas (para variables que evolucionan exponencialmente).

Aplicación de la media geométrica:

Suponemos un dinero colocado durante n años a unos tantos unitarios de interés anual i_1, i_2, …, i_n. Si deseamos saber el interés medio del periodo, no debemos calcular la media aritmética de los n tantos de interés anuales, pues nos daría un valor erróneo.

Sea Do el dinero inicial. Al final de cada año obtendré D_1, D_2, …, D_n.

$D_1 = D_o\ (1+i_1) = D_o + D_o i_1$

$D_2 = D_1\ (1+i_2) = D_o(1+i_1)(1+i_2)$

$D_n=D_{n-1}(1+i_n) = D_o(1+i_1)(1+i_2)\dots(1+i_n)$

El tanto medio i que buscamos tiene que ser el que cumpla la igualdad

$D_n = D_o(1+i)^n$, con lo que igualando obtenemos

$$(1+i)^n = (1+i_1)(1+i_2)\dots(1+i_n)$$

$$(1 + i) = \sqrt[n]{(1 + \mathrm{i1})(1 + \mathrm{i2}) \dots (1 + \mathrm{in})}$$

De donde el tanto i:

$i = \sqrt[n]{(1 + \mathrm{i1})(1 + \mathrm{i2}) \dots (1 + \mathrm{in})}$ - 1 = G-1 valor que se determina en función de la media geométrica de los rendimientos unitarios de cada año y que difiere sensiblemente del que se obtiene al calcular la media aritmética de los tantos unitarios de interés anual. Este i se conoce como tasa media anual acumulativa, rendimiento medio (media geométrica de los incrementos de cada periodo por unidad menos 1.

3.- LA MEDIANA (Me).

Definición: Aquel valor de la distribución, tal que una vez ordenados todos los valores observados de menor a mayor (en sentido creciente), ocupa el lugar central, es decir que deja a un lado y a otro de sí, el mismo nº de frecuencias, supuesto un número impar de datos. Si el número de datos fuese par puede decirse que hay 2 valores medianos, y se toma la media aritmética entre ellos. También se puede definir como aquel valor de la distribución cuya frecuencia acumulada es N/2. Para su cálculo en una distribución agrupada en intervalos no es necesario distinguir si los intervalos se han construido de la misma o diferente amplitud.

$$Me = L_{i-1} + c_i \frac{\left(\frac{N}{2} - N_{i-1}\right)}{n_i}$$

Propiedad: la mediana hace mínima la suma de todas las desviaciones absolutas.

$$\min \sum_{i=1}^{n} |x_i - k|\, n_i = \sum_{i=1}^{n} |x_i - Me|\, n_i$$

4.- LA MODA (Mo).

Es el valor de la variable que más veces se repite y en consecuencia, en una distribución de frecuencias, es el valor de la variable que viene afectada por la máxima frecuencia de la distribución.

Llamamos moda absoluta, a aquel valor de la variable cuya frecuencia absoluta no es superada por ningún otro valor de la variable.

Llamamos moda relativa, a aquel valor de la variable cuya frecuencia absoluta asociada no es superada por las de sus valores contiguos.

De estas definiciones se deduce que tanto la moda absoluta como la relativa no tienen porque ser necesariamente únicas. Podría ocurrir que hubiese varios valores que se repitiesen un nº de veces no superado por ningún otro valor, y en ese caso todos esos valores serían modas absolutas. Según una distribución tenga una, dos o más modas absolutas, recibe los nombres de distribución unimodal, bimodal o multimodal, respectivamente.

Cálculo de la moda:

A) Distribuciones no agrupadas en intervalos: la determinación de la media es inmediata, se observa la columna de las frecuencias absolutas y el valor de la distribución al que corresponde la mayor frecuencia será la moda.
B) Distribuciones agrupadas en intervalos. Aquí si es necesario distinguir si los intervalos son de la misma o diferente amplitud.

   Intervalos de la misma amplitud. Una vez determinada la mayor frecuencia observamos que a esta frecuencia no le corresponde un valor sino un intervalo, así no tendremos un valor modal sino un intervalo modal. De entre todos los valores comprendidos en el intervalo modal seleccionamos uno que desempeñe el papel de valor modal. Para esto, podemos utilizar distintos criterios, entre los cuales citamos 4:

   a) Tomar como valor modal el extremo inferior del intervalo
   b) Considerar como moda el extremo superior del intervalo.
   c) Hacer la moda igual a la marca de clase
   d) Suponer que todos los valores del intervalo están distribuidos uniformemente dentro de él, y la moda estará más cerca de aquel intervalo contiguo cuya frecuencia sea mayor. Este es el criterio utilizado y al que corresponde la fórmula

$$Mo = L_{i-1} + c_i \frac{n_{i+1}}{n_{i-1} + n_{i+1}}$$

   Intervalos de distinta amplitud. Aquí para determinar el intervalo modal no es significativa la mayor frecuencia (ni) puesto que al ser diferentes las amplitudes de cada intervalo las ni no nos dirá nada sobre la abundancia de valores en cada intervalo, ya que podría ocurrir que el intervalo con mayor frecuencia fuese muy amplio, utilizamos entonces las densidades de frecuencias (di), que nos dan el nº de valores que hay en cada unidad de intervalo, para cada intervalo. La mayor di nos determina el intervalo modal. Una vez determinado el intervalo modal, y con el objetivo de operar con valores y no con intervalos, podemos aplicar cualquiera de los 4 criterios mencionados anteriormente, y aplicamos también el 4º.

$$Mo = L_{i-1} + c_i \frac{d_{i+1}}{d_{i-1} + d_{i+1}}$$

   En las distribuciones de frecuencias normales (distribución estadística por excelencia que tiene forma de campana y es simétrica con respecto de su media), la media aritmética, mediana y moda coinciden. Sin embargo, si la distribución es acampanada, pero no es simétrica, entonces la mediana suele estar situada entre la moda y la media aritmética.

MEDIDAS DE POSICIÓN NO CENTRAL

Hemos visto que las medidas de posición central trataban de mostrar en algún sentido el valor más representativo, común o central de la distribución. Las medidas que vamos a ver ahora se llaman medidas de posición no central, porque, aunque tratan de posicionar sobre la escala de valores posibles de la variable algún punto característico de la distribución, ese punto no tiene porque ser, en ningún sentido central, sino un punto cualquiera.

La idea es análoga a la definición de mediana: si el valor mediano caracterizaba un valor central dado que dividía en 2 mitades la distribución, la idea es considerar puntos que nos dejen a su derecha una determinada posición de la distribución (no necesariamente la mitad, y, por tanto, a su izquierda el resto. A estas medidas en general se llaman cuantiles y se definen como aquellos valores que dividen la distribución en intervalos, de forma que cada uno de ellos comprendan el mismo número de valores y lo denominaremos cuantil r-ésimo de orden k y se representa por Qr/k. Los cuantiles más utilizados son:

- Cuartiles. Son 3 valores de la distribución que la dividen en 4 partes iguales, es decir, de igual frecuencia N/4. (C1 = Q1/4, C2 = Q2/4, C3= Q3/4). C1 deja a su izquierda el 25 % de la distribución y el 75 % a su derecha, C2 deja a su izquierda el 50% de la distribución y el 50% a su derecha, C3 deja a su izquierda el 75% de la distribución y el 25% a su derecha.
- Deciles. Son 9 valores de la distribución que la dividen en 10 partes iguales, es decir de igual frecuencia N/10. (D1=Q1/10, D2=Q2/10……, D9=Q9/10). D1 deja a su izquierda el 10% de la distribución y el 90% a su derecha, D2 deja a su izquierda el 20% y a su derecha el 80% y D9 deja el 90% a su izquierda y el 10% a su derecha.
- Centiles o percentiles. Son 99 valores de la distribución que la dividen en 100 partes iguales, es decir de igual frecuencia N/100. (P1 = Q1/100, P2=Q2/100, …., P99=Q99/100). P1 deja a su izquierda el 1% de la distribución y el 99% a su derecha, P2 deja a su izquierda el 2% y el 98% a su derecha,…., P99 deja el 99% a su izquierda y el 1% a su derecha.

Observar que los cuantiles medios C2, D5 y P50 coinciden con la mediana de la distribución.

Cálculo de los cuantiles:

- Para distribuciones no agrupadas en intervalos

  Cuartiles: C1, valor de la distribución que ocupa el lugar N/4, C2, valor de la distribución que ocupa el lugar 2N/4, C3, valor de la distribución que ocupa el lugar 3N/4.

  Deciles; D1, valor de la distribución que ocupa el lugar N/10, D2, valor de la distribución que ocupa el lugar 2N/10, D9, valor de la distribución que ocupa el lugar 9N/10.

  Percentiles: P1, valor de la distribución que ocupa el lugar N/100, P99, valor de la distribución que ocupa el lugar 99N/100.

- Distribuciones agrupadas en intervalos, fórmula igual que la mediana.

$$Q_{r/k} = L_{i-1} + c_i \frac{\left(\frac{rN}{k} - N_{i-1}\right)}{n_i}$$

MOMENTOS DE UNA DISTRIBUCIÓN DE FRECUENC IAS

Los momentos son medidas que caracterizan a una distribución de frecuencias, son unos valores de tal modo que dos distribuciones son iguales si tienen todos sus momentos iguales.

Momentos respecto al origen. Se define el momento de orden r de la variable X con respecto al origen

$a^r = \frac{\sum_{i=1}^{n} x_i^r n_i}{N}$, siendo $N = \sum_{i=1}^{n} n_i$

r=1 $a^1 = \frac{\sum_{i=1}^{n} x_i^1 n_i}{N} = \bar{x} = media\ aritmética$

r=2 $a^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N}$

Momentos respecto a la media. Momento de orden r de la variable X con respecto a la media aritmética (media).

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r n_i}{N}$$

r=0

$$m_0 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^0 n_i}{N} = \frac{\sum n_i}{N} = 1$$

r=1

$$m_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^1 n_i}{N} = 0\ por\ la\ propiedad\ 1)\ de\ la\ \overline{x}$$

r=2 $m_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 n_i}{N}$ = VARIANZA

Los momentos respecto a la media se pueden expresar en función de los momentos respecto al origen de igual e inferior orden.

$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}\ x_i) n_i}{N} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2\, n_i}{N} + \bar{x}^2 \frac{\sum n_i}{N} - 2\bar{x}\frac{\sum_{i=1}^{n} x_i\ \ n_i}{N} = a_2 + a_1^2 - 2a_1 a_1 =$$

$$= a_2 - a_1^2$$

En general, un momento de orden r respecto a la media se puede expresar como función de los primeros r momentos respecto al origen, por el binomio de Newton.

3.2.- MEDIDAS DE DISPERSIÓN

Con una medida de posición central tratamos de sintetizar mediante un solo número toda la información suministrada por una distribución de frecuencias. Ahora bien, no sabemos hasta que punto esa síntesis representa bien al conjunto de información.

Medir la representatividad del promedio equivale a cuantificar la dispersión de los valores respecto a dicho promedio.

"cuantos más cercanos estén los valores al promedio más representativo será". Esta variabilidad, heterogeneidad, esparcimiento o dispersión de los valores es lo que tratamos de medir con las MEDIDAS DE DISPERSIÓN. Resulta pues necesario acompañar de la información que nos den los promedios de alguna medida de dispersión de la variable y así decimos que cuando mayor dispersión exista respecto a la media, habrá una menor representatividad.

Las medidas de dispersión miden el grado medio de alejamiento o separación de las observaciones, con respecto de las medidas de posición central (media aritmética). Según este grado de alejamiento o separación se mida en unidades análogas a las de las observaciones, o por el contrario, no tenga unidades, las medidas de dispersión se llamarán ABSOLUTAS o RELATIVAS. La propiedad fundamental de las medidas de dispersión relativas estriba en que, al no venir expresada en ningún tipo de unidades, permite comparar distintas distribuciones (en cuanto a su dispersión), aunque viniesen medidas en distintos tipos de unidades.

MEDIDAS DE DISPERSIÓN ABSOLUTAS:

1.- Recorrido o rango. Diferencia entre el mayor valor y el menor valor de la variable. Esta medida tiene el inconveniente de que viene determinada por sólo 2 valores de la variable, siendo, por tanto, muy sensible a la fluctuación de estos valores extremos. Es aplicable sólo a algunos casos, como variables distribuidas uniformemente.

$$Re = x_n - x_1 = \max(x_i) - min(x_i)$$

2.- Recorrido intercuartílico. Diferencia entre el tercer cuartil y el primero. Esta medida nos da realmente el campo de variación del 50 % central de las observaciones de la distribución.

$$Re_I = C_3 - C_1$$

Sin embargo, estas 2 medidas de dispersión adolecen de un gran defecto: el no considerar la totalidad de los valores observados, con lo cual es fácil que distribuciones sustancialmente diferentes puedan dar mismas medidas de dispersión (estas 2 medidas no hacen referencia a ningún promedio). Para evitar este problema se recurre a la idea intuitiva de medir separaciones medias de los valores de la variable a las distintas medidas de posición central de la distribución, y así surgen las siguientes medidas de dispersión absolutas

3.- Desviación media respecto a la media aritmética. Se define como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la media aritmética. Si el resultado obtenido es pequeño se interpreta en el sentido de que los valores de la variable se distribuyen en valores próximos a la media. Si el resultado obtenido fuese un valor grande significaría que los valores de la variable, se distribuirán en valores alejados de la media.

$$D_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{n} |x_i - \overline{x}| n_i$$

4.- Desviación media respecto a la mediana. Se define como la media aritmética de las desviaciones en valor absoluto de los valores de la variable a la mediana. Para un valor grande de D_{Me}, la mediana no será representativa. Para cualquier distribución, siempre se verifica que $D_{Me} \leq D_{\bar{x}}$. Pero la desviación media respecto a la mediana tiene una desventaja a tener en cuenta respecto a la desviación media respecto a la media aritmética, y es precisamente la mayor complejidad de cálculo de la Mediana con respecto del cálculo de la media aritmética, por lo que suele usarse más la desviación media respecto a la media aritmética.

$$D_{Me} = \frac{1}{N}\sum_{i=1}^{n} |x_i - Me| n_i$$

A pesar de todo, estas medidas (3 y 4) tienen un inconveniente común, sobre todo a la hora de operar analíticamente con sus expresiones. El inconveniente es el uso de valores absolutos, lo cual las hace difícilmente manejables para cálculos algebraicos.

Por ello se intenta definir otras medidas de dispersión que, teniendo buenas propiedades algebraicas, sigan manteniendo la idea de medir distancias medias o similares (es decir desviaciones respecto a la media, con su signo correspondiente), pero por la primera propiedad de la media, el numerador sería nulo por tanto no podríamos usar esta medida como medida de dispersión.

Por otra parte, puesto que el cuadrado de una distancia es independiente del signo con que ésta se tome, parece lógico considerar, en lugar de distancias, cuadrados de distancias, y así surge una nueva medida de dispersión:

5.- Varianza. Se define como la media aritmética de los cuadrados de las diferencias (distancias) de los valores de la variable con respecto a su media aritmética y coincide con el momento de segundo orden respecto a la media aritmética. La varianza nos medirá la mayor o menor dispersión de los valores respecto a la media aritmética, si la dispersión es muy grande la media no será representativa. La varianza valdrá cero si todos los valores de la variable observados son exactamente iguales a la media aritmética, y será más grande cuanto más alejados estén los valores de ella.

$$S^2 = m_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i}{N}$$

<u>Propiedades de la varianza</u>:

1.- La varianza nunca puede ser negativa, ya que al ser una suma de cuadrados, a lo sumo, tomará el valor cero, y ello ocurrirá precisamente cuando todos los valores de la variable sean iguales, y entonces lo serán a la media aritmética.

$$S^2 \geq 0$$

2.- La varianza es igual al momento de 2º orden respecto al origen menos el de primer orden elevado al cuadrado. $S^2 = m_2 = a_2 - a_1^2$

3.- La varianza es la medida cuadrática de dispersión óptima.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i}{N} < \frac{\sum_{i=1}^{n}(x_i - k)^2 n_i}{N} \quad \forall k \neq \overline{x}$$

4.- La varianza está acotada inferior y superiormente en cada distribución de frecuencias.

5.- Si en la distribución de frecuencias sumamos a todos los valores de la variable una constante, la varianza no varía, es decir, un cambio de origen en la variable no afecta a la varianza.

$$Sea\ (xi, ni)\ \ y\ (xi', ni)\ donde\ xi' = xi + k\ y\ \ \bar{x}' = \bar{x} + k\ entonces$$

$$S'^2 = \frac{\sum_{i=1}^{n} \left(x'_i - \overline{x'}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{n} (x_i + k - \overline{x} - k)^2 n_i}{N} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i}{N} = S^2$$

6.- Al multiplicar los valores de una distribución de frecuencias por una constante k, la varianza queda multiplicada por la constante al cuadrado, es decir, la varianza cambia al realizar un cambio de escala.

$$Sea\ (xi, ni)\ \ y\ (xi', ni)\ donde\ xi' = xi * k\ y\ \ \bar{x}' = \bar{x} * k\ entonces$$

$$S'^2 = \frac{\sum_{i=1}^{n} \left(x'_i - \overline{x'}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{n} (x_i * k - \overline{x} * k)^2 n_i}{N} = k^2 \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i}{N} = k^2 S^2$$

Todas estas propiedades son ventajas de la varianza. Además, unido a su facilidad de cálculo esta medida de dispersión utiliza para su obtención todos los valores observados. Tiene sin embargo, un pequeño defecto de carácter práctico, y es que sus unidades de medida no coinciden con las de la variable en estudio, sino que son exactamente sus cuadrados. Esto dificulta su interpretación, y se hace necesario resolver este problema mediante la definición de la medida llamada desviación típica, que se define, como la raíz cuadrada positiva de la varianza.

$$S = +\sqrt{S^2}$$

6.- Desviación típica o desviación estándar (S). Esta medida ya sí viene expresada en las mismas unidades que la variable en estudio. Análogamente a como ocurre con la varianza, cuando la desviación típica sea próxima a cero, indicará una pequeña dispersión de las observaciones con respecto de la media aritmética, y valdrá exactamente cero si todas las observaciones presentan el mismo valor.

Propiedades de la desviación típica:

1.- Siempre será positiva, o igual a cero. $S \geq 0$

2.- Es una medida de dispersión óptima.

3.- es la raíz cuadrada del momento de 2 orden respecto al origen menos el de 1 orden elevado al cuadrado. $S = \sqrt{a_2 - a_1^2}$

4.- está acotada superior e inferiormente en cada distribución de frecuencias.

5. No le afectan los cambios de origen.

6.- Le afectan los cambios de escala, pero solo por la constante.

$$S' = \sqrt{S'^2} = \sqrt{k^2 S^2 = kS}$$

Finalmente, con respecto de las medidas de dispersión absolutas se demuestra que existe la siguiente relación:

$$D_{Me} \leq D_{\bar{x}} \leq s$$

Para distribuciones simétricas o moderadamente simétricas se cumple aproximadamente que:

- Entre $\bar{x} - S \; y \; \bar{x} + S \; quedan \; aproximadamente \; el \; 68 \; \% \; de \; las \; observaciones$
- Entre $\bar{x} - 2S \; y \; \bar{x} + 2S \; quedan \; aproximadamente \; el \; 95 \; \% \; de \; las \; observaciones$
- Entre $\bar{x} - 3S \; y \; \bar{x} + 3S \; quedan \; aproximadamente \; el \; 68 \; \% \; de \; las \; observaciones$

Tipificación de una variable.

Tipificar es expresar la distancia entre un valor individual de la variable y la media en términos de desviación típica. Por tanto permite hacer comparaciones, de posición relativa entre valores de distribuciones heterogéneas y decimos que una variable estadística se denomina tipificada, estandarizada o reducida si su media es cero y su varianza la unidad.

Sea una variable X con $\bar{x} \; y \; Sx^2$ y decimos que la variable Z

$$Z = \frac{X - \bar{X}}{S_x} \; es \; su \; tipificada \; si \; su \; \bar{z} = 0 \; y \; S_z^2 = 1$$

Todas las anteriores medidas de dispersión venían afectadas de unidades de medida y, como hemos visto para el caso de la varianza y desviación típica, un cambio de escala afectaba decisivamente al valor de dichas medidas de dispersión. Unas medidas con tales características mal pueden servir para comparar distintas distribuciones cuyas medidas sean sustancialmente diferentes, lo que nos lleva a definir

MEDIDAS DE DISPERSIÓN RELATIVAS

Son medidas adimensionales, que no se ven afectadas por las unidades de medida, y por lo tanto aparecen en forma de cociente.

1.- Coeficiente de apertura. Se define como el cociente entre los 2 valores extremos de una distribución. Nos indica el nº de veces que el mayor valor contiene al menor valor.

$$C_{Ap} = \frac{X_n}{X_1}$$

Es de fácil cálculo, pero tiene algunos inconvenientes:

- No tiene en cuenta más que los valores extremos y no el resto de la distribución ni a sus medidas de posición central.

- No puede aplicarse si $X_1 = 0$.

- Queda afectado por cambios de origen.

Se aplica frecuentemente al estudiar la dispersión de salarios de una empresa.

2.- Recorrido relativo. Se define como el cociente entre el recorrido de la variable y su media aritmética, y nos indica el nº de veces que el recorrido contiene a la media aritmética. También queda afectado por los cambios de origen y no puede utilizarse si la media aritmética es cero. El problema que presentan el coeficiente de apertura y el recorrido relativo es que pueden ser muy influenciables negativamente por observaciones anómalas extremas.

$$R_r = \frac{Re}{\bar{X}} = \frac{\max(X_i) - min(X_i)}{\bar{X}}$$

3.- Recorrido semi-intercuartílico. Se define como el cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil. Esta medida no está afectada por cambios de origen ni escala y se debería de aplicar cuando se utilizara la mediana como promedio, ya que $\frac{1}{2}(C_3 - C_1)$ es el punto central de la zona de la distribución, que contiene el 50 % central de las observaciones, el recorrido semi-intercuartílico puede considerarse como la mitad del nº de veces que el recorrido intercuartílico contiene a su punto central.

$$Re_{SI} = \frac{C_3 - C_1}{C_3 + C_1}$$

Las 3 medidas anteriores tienen el mismo inconveniente de no utilizar todos los valores de la distribución.

4.- Coeficiente de variación de Pearson (CV). Es el más empleado de los índices de dispersión relativa y se define como el cociente entre la desviación típica y la media aritmética, representando, por tanto, el nº de veces que la desviación típica la contiene. Es adimensional, y tiene en cuenta todos los valores observados de la distribución, por lo que se puede considerar el índice de dispersión relativa más completo.

$$Cv = \frac{S_x}{\bar{X}}$$

A mayor coeficiente de variación, menor representatividad de la media, cuánto más cercano a cero sea CV, evidentemente más cercano a cero debe ser S, indicándonos que la dispersión es más pequeña (máxima representatividad de la media). No debe aplicarse este coeficiente cuando la media aritmética sea cero →CV = ∞, falseando la verdadera dispersión de la distribución y no aportando ninguna información válida para su conocimiento. También se puede expresar en porcentaje. El coeficiente de variación variará ante cambios de origen.

La gran utilidad de estas medidas estriba en que nos permitirán comparar distribuciones que están medidas en distintas escalas o unidades, e incluso, aquellas que, aun estando medidas en las mimas, sus valores son sustancialmente diferentes, en cuyo caso las medidas de dispersión absolutas quedas influenciadas por ello y son poco adecuadas a tal fin.

3.3.- MEDIDAS DE FORMA.

Es necesario definir una serie de medidas que permitan cuantificar en lo posible la forma de la distribución de frecuencias. Esta cuantificación se realiza en 2 sentidos principales:

- La asimetría o asimetría de la distribución de frecuencias
- Curtosis o grado de apuntamiento de una distribución.

Medidas de asimetría

Se dirigen a elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presente la distribución, sin necesidad de llevar a cabo su representación gráfica, es decir con los coeficientes de asimetría tratamos de medir si las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central (en general, la media aritmética) y el grado de esta asimetría. Este tipo de medidas de forma deben ser adimensionales.

Decimos que una distribución es simétrica si existe el mismo número de valores a la derecha que a la izquierda de la media aritmética, y por tanto el mismo número de desviaciones con signo positivo que con signo negativo.

No podemos tomar potencias pares porque perderíamos los signos, que nos interesa conservar, así tomaremos una potencia impar de dichas desviaciones, la más simple el cubo, formando m_3, medida expresada en las mismas unidades que la variable elevadas al cubo, por lo que varía ante un cambio de escala. Si $m_3=0$ la distribución es simétrica, si $m_3>0$ la distribución es asimétrica positiva y si $m_3<0$ la distribución es asimétrica negativa. Para conseguir un indicador adimensional, dividimos la expresión anterior por una cantidad que venga en sus mismas unidades de medida (cubo de la desviación típica), obteniéndose el Coeficiente de Asimetría de Fisher , como la desviación típica siempre es positiva se seguirán la misma regla de signos que con m_3.

$$As_F = \frac{m_3}{S^3}$$

Otras medidas de asimetría:

. Pearson propuso para distribuciones campaniformes, unimodales y moderamente asimétricas un coeficiente de Asimetría, teniendo en cuenta que en una distribución campaniforme simétrica, la media aritmética, moda y mediana coinciden y por lo tanto $A_{Sp} = 0$.

$$As_P = \frac{\overline{x} - Mo}{S}$$

Si la distribución es asimétrica positiva o a derechas, la media aritmética se desplaza a la derecha de la moda y por lo tanto $A_{Sp}>0$.

Si la distribución es asimétrica negativa o a izquierdas la media aritmética se sitúa por debajo de la Moda y por lo tanto $A_{Sp}<0$.

. Coeficiente de Asimetría de Bowley, con la misma regla de signos que los casos anteriores.

$$As_B = \frac{C_1 + C_3 - 2Me}{C_3 - C_1}$$

Medidas de apuntamiento o curtosis

Las medidas de curtosis se aplican a distribuciones campaniformes, es decir unimodales simétricas o con ligera asimetría.

Las medidas de curtosis tratan de estudiar la distribución de frecuencias en la zona central de la distribución, por lo que la mayor o menor concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada. Por esta razón a las medidas de curtosis se les llama también de *apuntamiento o concentración central* .

Para estudiar la curtosis de una distribución es necesario definir previamente una distribución tipo que se toma como modelo de referencia. Esta distribución es la llamada distribución Normal, cuya representación gráfica es la campana de Gauss. Se presenta en numerosos casos, e implica que la mayoría de los valores de la variable están cerca de la media, y aquellos que se encuentran muy distanciados de ellos, a ambos lados, son poco numerosos. Este comportamiento es lógico y normal en numerosos fenómenos estudiados (peso, estatura, etc.) de ahí su nombre.

Así, tomando esta distribución como referencia diremos que una distribución puede ser más apuntada que la normal (LEPTOCÚRTICA) o menos apuntada (PLATICÚRTICA). A la distribución normal, se la llama MESOCÚRTICA.

En la distribución normal se verifica que

$$m_4 = 3S^4$$

Siendo m_4 el momento de orden 4 respecto a la media y S la desviación típica.

Si consideramos el cociente

$$\frac{m_4}{S^4} = 3$$

su valor será siempre igual a 3 cuando se trate de una distribución normal su valor será cero para la distribución normal, por ello como coeficiente de apuntamiento o curtosis se utiliza la siguiente expresión:

$$C_c = \frac{m_4}{S^4} - 3$$

Siendo una distribución:

- Mesocúrtica (normal) si $C_c = 0$
- Leptocúrtica si $C_c > 0$
- Platicúrtica si $C_c < 0$

3.4.- MEDIDAS DE CONCENTRACIÓN

Tratan de poner de relieve el mayor o menor grado de igualdad en el reparto del total de los valores de la variable. Son indicadores del grado de equidistribución de la variable y tienen especial aplicación a variables socioeconómicas (distribución de la renta, salarios, distribución de la propiedad agraria, distribución de la producción entre las empresas pertenecientes a un sector industrial). Como vimos la dispersión hace referencia a la variabilidad de los datos, a las diferencias que entre ellos existen, y por tanto a la representatividad de los promedios.

Llamamos concentración a la mayor o menor equidad en el reparto de la variable considerada.

En el cálculo de la media aritmética, el numerador es la suma de los valores que toma la variable (multiplicada o no, según sea el tipo de distribución, por las frecuencias correspondientes). En muchos casos dicho numerador no tiene un sentido estadístico claro, por ejemplo, en una distribución de estaturas o pesos sería la suma de las estaturas o pesos. Pero en otros casos, en particular cuando se trata de variables de carácter socio-económico, si lo tiene. Por ejemplo, en una distribución de salarios el numerador de la media aritmética representaría la masa total de los salarios.

Las medidas de concentración tienen por finalidad, precisamente, medir la uniformidad del reparto de dicha masa total: si todos los trabajadores percibiesen el mismo salario, la uniformidad de dicho reparto sería absoluta, y diremos que la concentración sería mínima, por el contrario, en un caso hipotético, si la masa total de los salarios fuese percibida por un solo trabajador, no existiría uniformidad en el reparto y en este caso diremos que la concentración es máxima.

Lógicamente, cuando se tiende hacia una uniformidad absoluta, es decir, la concentración es mínima, la media aritmética es perfectamente representativa de la distribución de frecuencias, contrariamente a lo que sucede cuando la falta de uniformidad es total, es decir la concentración es máxima.

Es decir, si suponemos que la distribución es de salarios y que tenemos n trabajadores cuyos salarios son

$X_1 \leq X_2 \leq X_3 \leq .... \leq X_n$ ordenados nos interesa estudiar hasta que punt la suma total de salarios $\sum X_i$ está equitativamente repartida.

No cabe duda que las infinitas posiciones que se nos pueden presentar estarán entre las 2 siguientes situaciones extremas:

1. Concentración máxima: cuando de los n trabajadores, solo 1 percibe el total de la masa salarial y los demás nada

   $X_1=X_2=X_3=....=X_{n-1}=0$ para $X_n \# 0$

2. Concentración mínima o equidistribución: todos los trabajadores perciben la misma cantidad

   $X_1=X_2=X_3=....=X_n=0$

Existen diferentes estudios teóricos sobre el tema de concentración pero nos centramos en el Índice de concentración de Gini y la curva de Lorenz.

a) INDICE DE GINI

Conviene disponer de un indicador que nos permita valorar numéricamente la concentración y al mismo tiempo, facilite la comparación entre 2 distribuciones. A esta finalidad responde el Índice de Gini o índice de concentración. Sea una distribución de frecuencias (Xi, ni) de la que vamos a formar las siguientes columnas:

1.- los productos Xini que nos indicará la masa salarial total percibida por los ni trabajadores de salario individual Xi.

2.- las frecuencias absolutas acumuladas Ni.

3.- los totales acumulados Ui calculados de la siguiente forma:

$U_1 = X_1 n_1$

$U_2 = X_1 n_1 + X_2 n_2$

$U_n = X_1 n_1 + X_2 n_2 + ... + X_n n_n = \sum X_i n_i$

Ui será el salario total percibido por los Ni primeros trabajadores.

4.- Columna de frecuencias acumuladas relativas, que expresamos en tano por ciento, y que llamamos p_i=Ni/N * 100

5.- Calculado el salario total para todos los trabajadores, que será Un, expresamos cada Ui en tantos por ciento de Un y a este porcentaje le llamamos qi = Ui/Un * 100.

Relacionando las dos últimas columnas, obtenemos la información que nos indica el "reparto" de los salarios, poniéndonos de relieve la concentración de los

En el ejemplo, ordenados los trabajadores de menor a mayor salario resulta que el 30 % de los trabajadores se reparte el 26 % de total de salarios de la empresa, el 60 % de los trabajadores recibe el 54 % del total de los salarios, el 80% de los trabajadores el 75 % del total de los salarios, etc.

Si los salarios estuvieran equidistribuidos el 30 % de los trabajadores recibiría el 30 % de los salarios, el 60 % de los trabajadores recibiría el 60 % del total de los salarios., etc.

$$I_G = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

Observar que en la fórmula del Índice de Gini, tanto el sumatorio del numerador como el del denominador se extiende desde i=1 hasta n-1, ya que para i=n→pi-qi=100-100 = 0.

Si la concentración es mínima, es decir, si los salarios están repartidos por igual pi=qi→I_G=0.

En el caso de concentración máxima, solamente el último trabajador percibe el salario $q_1=q_2=....=q_{n-1}=0 \rightarrow I_G=1$.

$$I_G = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{\sum_{i=1}^{n-1}(p_i - 0)}{\sum_{i=1}^{n-1} p_i} = 1$$

El campo de variación del I_G oscila de 0 a 1 y el Índice de Gini responderá a una distribución tanto más justa de salarios o rentas, cuanto más próximo esté de cero.

b) CURVA DE LORENZ

La distribución de salarios vista en el ejemplo se puede representar gráficamente obteniéndose la llamada "curva de concentración o curva de Lorenz".

La curva de Lorenz se obtiene representando los pares de valores pi y qi y uniendo mediante líneas rectas cada par consecutivo y, además (p_1,q_1) con el origen. Para pi utilizaremos el eje de abscisas y para qi el eje de ordenadas. Por ser idénticos tanto la escala como el campo de variación en cada uno de los ejes, la curva de Lorenz encaja perfectamente en un cuadrado, por lo que dibujamos un cuadrado cuyos lados están divididos en una escala de 1 a 100. Se representa también la diagonal que arranca del origen, esta diagonal se toma como elemento de referencia de la curva de Lorenz. La curva que se obtiene necesariamente habrá de situarse por debajo de esa diagonal, ya que sería imposible que el pi % de los primeros trabajadores superara ese mismo porcentaje en cuanto a la masa salarial acumulada qi, con lo que no puede haber ningún punto por encima de esa diagonal. También por la ordenación de salarios de menor a mayor, la curva tendrá que ser siempre creciente, ya que se consideran porcentajes obtenidos de totales acumulados.

Empezará en el origen (0,0) ya que el 0 % de los trabajadores dispondrá del 0 % de los salarios y terminará en B (100 %, 100 %) por un razonamiento análogo.

La curva que nos indicará la concentración mínima coincidirá con la diagonal, ya que en ella pi=qi→ la concentración mínima es equivalente a equidistribución y en este caso todos los trabajadores percibirían la misma cantidad, es decir, si todos los empleados percibiesen el mismo salario, a un porcentaje dado de trabajadores le correspondería idéntico porcentaje de la masa salarial, con lo que todos los puntos de la curva de Lorenz se encontrarían en la diagonal, este caso corresponde con un valor cero del índice de Gini

El caso más desfavorable sería aquel en que el 99 % de los trabajadores percibiese el 0 % de la masa salarial y el 1 % restante percibiese el 100 % de la masa salarial, entonces la curva de Lorenz estaría formada por los lados inferior y derecho del cuadrado, correspondiendo a un caso de concentración máxima, y a un valor uno del Índice de Gini. Por tanto, la curva de Lorenz deberá estar situada en el triángulo inferior del cuadrado que hemos formado y se deduce que cuanto más próxima esté la curva a la diagonal mejor será la distribución de la variable económica considerada, en nuestro ejemplo una distribución de salarios.

Es evidente que entre los 2 casos extremos, existen casos intermedios en los que la curva de Lorenz es más curvada cuanto más desigual sea la distribución, y más fuerte la concentración.

Finalmente una consideración importante: dos distribuciones pueden tener el mismo grado de concentración global (Índice de Gini igual), pero la representación de las curvas sea distinta (es decir, la estructura del reparto no sea la misma).

¿Cómo varía el Índice de Gini a cambios de origen y de escala?

EJEMPLO CLASE I. GINI

SALARIOS			Nº DE EMPLEADOS						
L_{i-1}	L_i	xi	ni	xi*ni	Ui	Ni	Pi=(Ni/N)*100	Qi=(Ui/Un)*100	pi-qi
						0	0	0	
140	150	145	15	2175	2175	15	30,00	26,36	3,64
150	160	155	15	2325	4500	30	60,00	54,55	5,45
160	180	170	10	1700	6200	40	80,00	75,15	4,85
180	200	190	5	950	7150	45	90,00	86,67	3,33
200	220	210	3	630	7780	48	96,00	94,30	1,70
220	250	235	2	470	8250	50	100,00	100,00	0,00
	Sumas		50	8250			356,00		18,97

I.GINI = **0,053**