TEMAS DE MATEMÁTICAS
(Oposiciones de Secundaria)

TEMA 60

PARÁMETROS ESTADÍSTICOS. CÁLCULO, SIGNIFICADO Y PROPIEDADES.

TEMA 60

PARÁMETROS ESTADÍSTICOS. CÁLCULO, SIGNIFICADO Y PROPIEDADES.

1. INTRODUCCIÓN.

Aunque la observación visual de cualquier representación gráfica de una misma distribución de frecuencias proporcionan una primera aproximación al análisis de los datos, este tipo de observación no nos permite comparar, con rigor, dos distribuciones del mismo carácter. Por tanto, se hace necesario estudiar procedimientos numéricos que obtengan, a partir de todos los datos de la distribución, unos valores que permitan deducir una información cuantitativa.

Por otra parte, si no se dispone de ninguna representación gráfica, es necesario, ante la imposibilidad del humano de retener gran cantidad de datos, el intentar resumir toda la información.

La idea de resumir la información del comportamiento global del fenómeno estudiado en unos pocos datos se realiza calculando una serie de parámetros. Este tipo de medidas descriptivas utilizadas son, principalmente, medidas de centralización o posición, medidas de dispersión, medidas de deformación o simetría y el apuntamiento.

2. MEDIDAS DE POSICIÓN.

La tabla estadística nos ofrece toda la información disponible, pero el investigador se encuentra incapaz, en numerosos casos, de interpretar toda esa extensa información, por lo que intenta resumirla en una serie de expresiones. Hacia la síntesis de esa información van dirigidas todas estas expresiones o medidas.

Toda síntesis de una distribución se considerará como operativa si:

1) Intervienen en su determinación todos y cada uno de los valores de la distribución.
2) Es siempre calculable.
3) Es única para cada distribución de frecuencias.

En este proceso de síntesis buscamos unos valores que nos fijen el comportamiento global del fenómeno estudiado a partir de los datos individuales recogidos en la información disponible. Estos valores sintéticos globales son las llamadas Medidas de Posición.

2.1. La Media Aritmética.

DEF Definimos la Media Aritmética como la suma de todos los valores de la distribución dividida por el número total de datos. Si el valor x_i de la variable X_i se repite n_i veces, hay que considerar estas repeticiones en la suma. Si representamos la media aritmética por $\bar{x}$, tendremos:

$$\bar{x} = \frac{x_1 n_1 + \ldots + x_n n_n}{N} = \sum_{i=1}^{n} \frac{x_i n_i}{N}$$

Pero esto sólo es válido en el supuesto más sencillo en que los datos de la variable estén sin agrupar. En el caso de que tuviéramos una distribución con datos agrupados, los valores individuales de la variable serían desconocidos y, por tanto, no podríamos hacer uso de la fórmula anterior. En este supuesto se postula la hipótesis de que el punto medio del intervalo de clase (marca de clase) representa adecuadamente el valor medio de dicha clase; y aplicaríamos la fórmula original de la media simple para dichos valores.

Otro tema al que tenemos que hacer referencia es el de la llamada Media Aritmética Ponderada, que se produce cuando se otorga a cada valor de la variable una ponderación o peso, distinto de la frecuencia o repetición. En este caso, en el cálculo de la media aritmética tendríamos en cuenta dichas ponderaciones.

En este caso, si ω_i son las ponderaciones, definimos la media como:

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

2.1.1. Propiedades.

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

Dem.

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

TEOREMA. Teorema de König.

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 es igual a la media aritmética.

Dem.

Consideremos la expresión:

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

que toma diferentes valores para una misma distribución de frecuencias, según los distintos valores de k.

Si sumamos y restamos $\bar{x}$ dentro del paréntesis, tenemos que:

$$D(k)=\sum_{i=1}^{n}(x_i-k)^2\,\frac{n_i}{N}=\sum_{i=1}^{n}(x_i-k+\bar{x}-\bar{x})^2\,\frac{n_i}{N}=\sum_{i=1}^{n}\left((x_i-\bar{x})-(k-\bar{x})\right)^2\,\frac{n_i}{N}=$$

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

y teniendo en cuenta la propiedad anterior, se transforma en

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

en donde el valor de k que hace que D(k) sea mínimo es $\bar{x}$.

PROP Si a todos los valores de una variable les sumamos una constante k, la media aritmética queda aumentada en esa constante. O lo que es lo mismo, la media aritmética se ve afectada por los cambios de origen.

Dem.

Inmediata.

PROP Si todos los valores de una variable los multiplicamos por una constante k, su media aritmética queda multiplicada por la misma constante. Es decir, la media aritmética se ve afectada por los cambios de escala.

Dem.

Inmediata.

2.1.2. Cálculo Abreviado.

Con objeto de facilitar el cálculo práctico de $\bar{x}$ cuando la distribución presenta numerosos valores o éstos están compuestos por bastantes dígitos, es aconsejable realizar el siguiente cambio de variable

$$x_i'=\frac{x_i-O}{c}$$

donde:

x_i son los valores de la distribución.

O un origen de trabajo arbitrario, que se procura sea un valor central de la distribución.

c una constante que es igual al máximo común divisor de las diferencias que existen entre cada dos valores consecutivos de la variable.

X_i' los nuevos valores de la variable.

Despejando x_i de la expresión anterior tenemos que $x_i = cx_i' + O$

Por tanto:

$$\bar{x} = \sum_{i=1}^{n} x_i \frac{n_i}{N} = \sum_{i=1}^{n} (cx_i' + O)\frac{n_i}{N} = \sum_{i=1}^{n} cx_i' \frac{n_i}{N} + O\sum_{i=1}^{n} \frac{n_i}{N} = c\bar{x}' + O$$

2.1.3. Ventajas e Inconvenientes.

Como ventajas podemos nombrar las tres que se le exigen a una medida de síntesis:

1) Consideración de todos los valores de la distribución.
2) Ser calculable.
3) Ser única.

También podemos considerar como ventaja la obtenida en la primera propiedad, que nos dice que la media aritmética es el centro de gravedad de la distribución, así como la obtenida de la segunda propiedad (Teorema de König).

Como inconvenientes podemos indicar que a veces da lugar a conclusiones no muy atinadas. Esto ocurre en el caso de que la variable presente valores anormalmente extremos que pueden distorsionar la media aritmética, haciéndola poco 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.

2.2. La Media Geométrica.

DEF Sea una distribución de frecuencias (x_i; n_i). Definimos la Media Geométrica, y la representaremos por G, como la raíz N-ésima del producto de los N valores de la distribución. Así:

$$G = \sqrt[N]{x_1^{n_1} \cdot \ldots \cdot x_n^{n_n}} = \sqrt[N]{\prod_{i=1}^{n} x_i^{n_i}}$$

2.2.1. Propiedades.

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

Dem.

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

2.2.2. Ventajas e Inconvenientes.

Como ventajas podemos señalar:

1) En su determinación intervienen todos los valores de la distribución.
2) Es menos sensible que la media aritmética a los valores extremos, por su carácter de producto.

Como inconvenientes tenemos:

1) Es de significado estadístico menos intuitivo que la media aritmética.
2) Su cómputo es más difícil.
3) En ocasiones no queda determinada. Esto ocurre cuando la variable toma en algún momento el valor 0. Si la variable toma valores negativos, se pueden presentar una amplia gama de casos particulares en los que tampoco queda determinada G. No es que G no exista, sino que no la podemos determinar.

El empleo más frecuente de la media geométrica es el de promediar porcentajes, tasas, números índices, etc. Es decir, en los casos en los que se supone que la variable presenta variaciones acumulativas.

2.3. La Media Armónica.

DEF Definimos la Media Armónica de una distribución de frecuencias (x_i; n_i), y se representa por H, como

$$H = \frac{N}{\frac{n_1}{x_1} + \ldots + \frac{n_n}{x_n}} = \frac{N}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$$

OBS La inversa de la media armónica es la media aritmética de los inversos de los valores de la variable.

2.3.1. Ventajas e Inconvenientes.

Como ventajas diremos que intervienen en su cálculo todos los valores de la distribución y que, en ciertos casos, es más representativa que la media aritmética. Por otra parte, siempre se puede pasar de una media armónica a una media aritmética transformando adecuadamente los datos.

Como inconvenientes hemos de citar la influencia de los valores pequeños, y su no determinación en las distribuciones con algunos valores iguales a cero. Por ello no es aconsejable su empleo en distribuciones en las que existan valores muy pequeños.

Se suele utilizar para promediar velocidades, tiempos, rendimientos, etc.

2.4. Relación entre los tres Promedios.

PROP Para una misma distribución de frecuencias (x_i; n_i), y siempre que existas, se verifica que:

$$H \le G \le \bar{x}$$

Dem.

Vamos a ver la demostración para el caso de una distribución con dos valores x_1 y x_2 con frecuencia unitaria. Para el caso de n valores, sólo hemos de aplicar inducción.

Las medias tienen como valores:

$$\bar{x} = \frac{x_1 + x_2}{2} \qquad G = \sqrt{x_1 x_2} \qquad H = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

Comencemos demostrando que $H \le G$

$$\frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} \le \sqrt{x_1 x_2} \quad \Rightarrow \quad \frac{2x_1 x_2}{x_1 + x_2} = \sqrt{x_1 x_2} \quad \Rightarrow$$

$$\Rightarrow \quad 2x_1 x_2 \le \sqrt{x_1 x_2}\,(x_1 + x_2) \quad \Rightarrow \quad 4x_1^2 x_2^2 \le x_1 x_2 (x_1 + x_2)^2$$

$$\Rightarrow \quad 4x_1 x_2 \le (x_1 + x_2)^2 \quad \Rightarrow \quad 4x_1 x_2 \le x_1^2 + 2x_1 x_2 + x_2^2$$

$$\Rightarrow \quad 0 \le x_1^2 - 2x_1 x_2 + x_2^2 \quad \Rightarrow \quad 0 \le (x_1 - x_2)^2$$

que es una desigualdad que claramente se verifica siempre, por tanto $H \le G$.

Por otro lado, veamos que $G \le \bar{x}$

$$\sqrt{x_1 x_2} \le \frac{x_1 + x_2}{2} \quad \Rightarrow \quad 4x_1 x_2 \le (x_1 + x_2)^2 \quad \Rightarrow \quad 0 \le (x_1 - x_2)^2$$

que es el mismo resultado que en el caso anterior.

Luego se verifica la desigualdad entre las tres medias.

2.5. La Mediana.

Dado que la definición de Mediana puede entrañar múltiples dificultades, vamos a dar una definición operativa diciendo:

DEF La Mediana es el valor de la distribución, supuesta ésta ordenada de menor a mayor, que deja a su izquierda y a su derecha el mismo número de frecuencias. Es decir, el valor que ocupa el lugar central, supuesto un número impar de datos. Si el número de

datos fuese par puede decirse que hay dos valores medianos, y se toma la media aritmética de ellos.

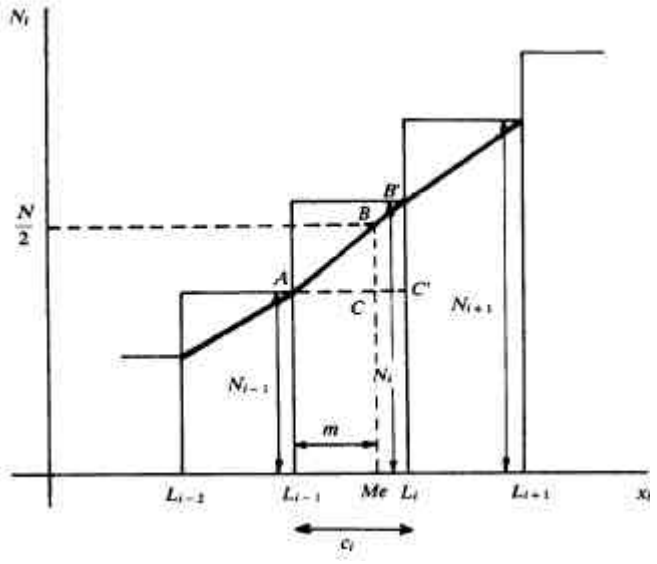
La Mediana también se puede definir como el valor de la distribución cuya frecuencia acumulada es N/2.

No obstante, hemos de tener en cuenta que para los diferentes casos particulares (número impar de datos, número par de datos, distribuciones de frecuencias unitarias) se pueden establecer diferentes criterios.

En el caso de distribuciones agrupadas en intervalos, no es necesario distinguir si los intervalos se han construido de la misma o distinta amplitud. Siguiendo el método general de búsqueda del valor que ocupa el lugar N/2, en este caso, nos encontramos con un intervalo mediano, en lugar de un valor mediano.

Con el objeto de fijar la mediana en un valor, seleccionaremos un representante del intervalo mediano al que llamaremos mediana. El criterio usualmente seguido es el siguiente.

Suponemos, en primer lugar, que todos los valores comprendidos dentro del intervalo mediano se encuentran distribuidos uniformemente a lo largo de él. A continuación, vamos a considerar la poligonal de frecuencias acumuladas correspondiente al intervalo mediano y a sus dos contiguos, y determinamos gráficamente la mediana.



Vemos que $M_e = L_{i-1} + m$. Determinaremos m en base a la hipótesis fijada que nos permite escribir

$$\frac{\overline{AC}}{\overline{AC'}} = \frac{\overline{BC}}{\overline{B'C'}}$$

ya que los triángulos ABC y AB'C' son semejantes.

Pero

$$\overline{AC} = m \qquad \overline{AC'} = c_i \qquad \overline{BC} = \frac{N}{2} - N_{i-1}$$

Por tanto

$$\overline{B'C'} = N_i - N_{i-1} = n_i \qquad \frac{m}{c_i} = \frac{\frac{N}{2} - N_{i-1}}{n_i}$$

Es decir

$$m = \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i$$

Con lo que tenemos

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

2.5.1. Propiedades de la Mediana.

PROP La Mediana hace mínima la suma de todas las desviaciones absolutas. Es decir, si representamos la mediana por Me, tenemos que

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

cuando la constante respecto a la cual se toman las desviaciones, k, es igual a la mediana.

Dem.

Transformamos la distribución en otra de frecuencias unitarias, tal que

$$x_1 \le x_2 \le ... \le x_{m-1} \le Me \le x_m \le ... \le x_{a-1} \le k \le x_a \le ... \le x_n$$

siendo k > Me.

Por definición de Mediana, tendremos igual número de valores iguales o inferiores que iguales o superiores, Supongamos que hay m-1 en cada lado. Tendremos que:

$$\sum_{i=1}^{n} |x_i - k| = \sum_{i=1}^{m-1} (k - x_i) + \sum_{i=m}^{a-1} (k - x_i) + \sum_{i=a}^{n} (x_i - k) \qquad (1)$$

$$\sum_{i=1}^{n} |x_i - Me| = \sum_{i=1}^{m-1} (Me - x_i) + \sum_{i=m}^{a-1} (x_i - Me) + \sum_{i=a}^{n} (x_i - Me) \qquad (2)$$

$$\sum_{i=1}^{n} |x_i - k| - \sum_{i=1}^{n} |x_i - Me| = \sum_{i=1}^{m-1} (k - Me) + \sum_{i=m}^{a-1} (k + Me - 2x_i) + \sum_{i=a}^{n} (Me - k) \qquad (3)$$

Sumando y restando $\sum_{i=m}^{a-1}(k-Me)$ en (3):

$$\sum_{i=1}^{n}|x_i-k|-\sum_{i=1}^{n}|x_i-Me|=\sum_{i=1}^{m-1}(k-Me)+\sum_{i=m}^{a-1}(k-Me)+\sum_{i=m}^{a-1}(k+Me-2x_i)-\sum_{i=m}^{a-1}(k-Me)+$$
$$+\sum_{i=a}^{n}(Me-k)$$

$$\sum_{i=1}^{n}|x_i-k|-\sum_{i=1}^{n}|x_i-Me|=\sum_{i=1}^{m-1}(k-Me)-\left[\sum_{i=m}^{a-1}(k-Me)+\sum_{i=a}^{n}(k-Me)\right]+$$
$$+\sum_{i=m}^{a-1}(k+Me-2x_i)+\sum_{i=m}^{a-1}(k-Me)=$$

$$=(m-1)(k-Me)-(m-1)(k-Me)+\sum_{i=m}^{a-1}(2k-2x_i)=2\sum_{i=m}^{a-1}(k-x_i)$$

Es decir

$$\sum_{i=1}^{n}|x_i-k|-\sum_{i=1}^{n}|x_i-Me|=2\sum_{i=m}^{a-1}(k-x_i)>0$$

Luego

$$\sum_{i=1}^{n}|x_i-k|>\sum_{i=1}^{n}|x_i-Me|$$

por tanto $\sum_{i=1}^{n}|x_i-Me|$ es mínimo para cualquier k>Me.

La demostración para cualquier k<Me es análoga y la omitimos.

2.5.2. Ventajas e Inconvenientes.

Las ventajas y los inconvenientes son los mismos que posteriormente veremos en el caso de la Moda.

A pesar de la fórmula que hemos estado viendo para el caso de distribuciones en escala por intervalos, la mediana tiene un mayor sentido en casos de distribuciones en escala ordinal (datos susceptibles de ser ordenados), de la cual es la medida más representativa por describir la tendencia central de la misma.

2.6. La Moda.

DEF Llamaremos Moda al valor de la variable que más se repite. Se representa por Mo.

Por tanto, en una distribución de frecuencias, es el valor de la variable que viene afectada por la máxima frecuencia de distribución.

Para calcular la Moda, distinguiremos entre distribuciones no agrupadas en intervalos y distribuciones agrupadas en intervalos.

2.6.1. Distribuciones no agrupadas en Intervalos.

En este caso, la determinación de la Moda Mo 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.

A veces aparecen distribuciones con más de una moda (bimodales, trimodales, etc.) e incluso una distribución de frecuencias que presente una moda absoluta y una relativa.

2.6.2. Distribuciones agrupadas en Intervalos.

a) Intervalos de la misma amplitud.

En este caso, una vez determinada la mayor frecuencia, observamos que a ésta no le corresponde un valor sino un intervalo, luego realmente no tendremos un valor modal sino un intervalo modal.

De entre todos los valores comprendidos en el intervalo modal vamos a seleccionar uno que desempeñe el papel de valor modal. Para esto, podemos utilizar diferentes criterios, entre los cuales citamos los cuatro siguientes:

1) Tomar como valor modal el extremo inferior del intervalo. $Mo = L_{i-1}$.
2) Considerar como moda el extremo superior. $Mo = L_i$.
3) Hacer la moda igual a la marca de clase. $Mo = x_i$.
4) O bien, suponiendo que:
   - Todos los valores del intervalo están distribuidos uniformemente dentro de él.
   - La moda estará más cerca de aquel intervalo contiguo cuya frecuencia sea mayor.

   Lo anterior se puede resumir diciendo que las distancias de la moda Mo a los intervalos contiguos son inversamente proporcionales a las frecuencias de dichos intervalos.

   La Moda será $Mo = L_{i-1} + m$

   Pero $$\frac{m}{c_i - m} = \frac{n_{i+1}}{n_{i-1}}$$

   Que teniendo en cuenta las propiedades de las proporciones queda:

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

de donde

$$m = \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i$$

Por tanto

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

b) Intervalos de Distinta Amplitud.

Si recurrimos a la definición que hemos dado como moda (valor que más se repite), al ser ahora los intervalos diferentes la frecuencia absoluta no nos dirá nada sobre la abundancia de valores en cada intervalo, ya que podría suceder que el intervalo al que correspondiese la mayor frecuencia fuera muy amplio y entonces, fuera más denso otro intervalo con menor frecuencia pero menor amplitud. Por tanto, ahora, las frecuencias no son significativas para resolver el problema.

Recordemos que las densidades de frecuencia se obtenían dividiendo las frecuencias absolutas por los recorridos o amplitudes de sus correspondientes intervalos. Las densidades de frecuencias nos dan el número de valores que hay en cada unidad de intervalo, para cada intervalo. La mayor densidad de frecuencia nos determinará el intervalo modal.

Una vez determinado el intervalo modal, y siempre en la línea de operar con valores y no con intervalos, podemos aplicar cualquiera de los cuatro criterios expuestos en el caso anterior. Si seleccionamos, por parecer el más razonable, el cuarto, tendremos que:

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

La deducción de esta fórmula es análoga al caso anterior.

Por último, diremos que la moda es la medida más representativa en caso de distribuciones en escala nominal. Esto es debido a que las distribuciones de este tipo presentan los datos no susceptibles de ordenación, de tal forma que para estas distribuciones no es posible realizar operaciones elementales con sus observaciones.

2.7. Medidas de Posición no Centrales.

Podemos nombrar otros valores notables pero que no van a reflejar ninguna tendencia central: los Cuantiles. Son valores de la distribución que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores.

Entre los Cuantiles podemos citar, por ser de uso más frecuente, los Cuartiles, los Deciles y los Percentiles.

DEF Llamaremos Cuartiles a los tres valores de la distribución que la dividen en cuatro partes iguales. Es decir, en cuatro intervalos dentro de cada cual están incluidos el 25% de los valores de la distribución.

DEF Llamaremos Deciles a los nueve valores de la distribución que la dividen en diez partes iguales. Cda parte contendrá el 10% de la distribución.

DEF Llamaremos Percentiles a los noventa y nueve valores que dividen a la distribución en cien partes iguales.

Para calcular los valores anteriores hemos de distinguir entre:

a) Distribuciones no agrupadas en intervalos.

Cuartiles: C_i es el valor que ocupa el lugar $\frac{iN}{4}$ con i:1,2,3.

Deciles: D_i es el valor que ocupa el lugar $\frac{iN}{10}$ con i: 1,...,9

Percentiles: P_i es el valor que ocupa el lugar $\frac{iN}{100}$ con i:1, , 99

Para determinarlos, se calculan previamente las frecuencias acumuladas, y se busca el valor que ocupe el lugar $\frac{iN}{k}$ de la distribución.

b) Distribuciones agrupadas en intervalos.

El problema que se presenta es el mismo que el que teníamos al calcular la mediana. Para elegir el representante para un determinado cuantil seguiremos el criterio:

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i} \cdot c_i$$

donde

1) Para k=4 y r=1,2,3 obtenemos los cuartiles.
2) Para k=10 y r=1,2,...9 obtenemos los deciles.
3) Para k=100 y r=1,2,...,99 obtenemos los percentiles.

La fórmula anterior se obtiene de forma análoga a la desarrollada para la mediana.

3. MOMENTOS POTENCIALES.

Al considerar las diferentes características de una distribución haremos referencia a unos valores específicos, deducidos de todos los valores de la distribución y que, como

se verá, serán la base de alguna de las características de cada distribución de frecuencia. Estos valores específicos reciben el nombre de Momentos.

Los Momentos de una distribución son unos valores que la caracterizan, de tal modo que dos distribuciones son iguales si tienen todos sus momentos iguales, y son tanto más parecidas cuanto mayor sea el número de momentos iguales que tengan.

Conviene advertir que existen dos tipos de momentos: los potenciales y los factoriales. Sólo vamos a tratar los momentos potenciales y a partir de ahora los designaremos simplemente por momentos.

El momento de orden r respecto a un origen arbitrario O se define como

$$M_r = \sum_{i=1}^{n} (x_i - O)^r \cdot \frac{n_i}{N}$$

Pero dentro de los momentos potenciales podemos distinguir, a su vez, dos tipos: los momentos respecto al origen y los momentos respecto a la media aritmética.

3.1. Momentos respecto al Origen.

Los momentos respecto al origen se representan por a_r y se obtienen haciendo O=0. Por tanto:

$$a_r = \sum_{i=1}^{n} x_i^r \cdot \frac{n_i}{N}$$

Los primeros momentos serán

$a_0=1$ $a_1=\bar{x}$ etc.

3.2. Momentos respecto a la Media Aritmética.

Son también llamados momentos centrales. Se representan por m_r y se obtienen al hacer $O=\bar{x}$. Por tanto:

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

siendo $\bar{x}$ la media aritmética de la distribución, que coincide con el momento de primer orden respecto al origen.

$m_0=1$ $m_1=0$ $m_2=s^2$ etc.

Observemos que los términos del sumatorio son de la forma $(x_i - \bar{x})^r \cdot n_i$. Si llamamos

$$u_i = x_i - \bar{x}$$

el momento central de orden r de la distribución será:

$$m_r = \sum_{i=1}^{n} u_i^r \cdot \frac{n_i}{N}$$

que es por definición el momento de orden r respecto al origen para la distribución (u_i; n_i). Por tanto, conceptualmente no existe diferencia entre los momentos respecto al origen y respecto a la media. La única diferencia existente entre ambos comsiste en que mientras en los momentos respecto al origen se toma como origen de medidas el cero de la escala correspondiente a la característica en estudio, en los momentos centrales se hace una traslación del origen de medidas, para situarlo precisamente en la media aritmética.

PROP Todos los momentos respecto a la media se pueden representar en función de los momentos respecto al origen.

Dem.

El binomio de Newton nos dice que

$$(a-b)^r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} a^{r-k} b^k$$

y aplicándolo a la expresión de los momentos centrales

$$m_r = \sum_{i=1}^{n} (x_i - \bar{x})^r \frac{n_i}{N} = \sum_{i=1}^{n} \sum_{k=0}^{r} (-1)^k \binom{r}{k} x_i^{r-k} \bar{x}^k \frac{n_i}{N} =$$

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

$$= \sum_{k=0}^{r} (-1)^k \binom{r}{k} \bar{x}^k a_{r-k} = \sum_{k=0}^{r} (-1)^k \binom{r}{k} a^k a_{r-k}$$

Como casos particulares podemos expresar

$$m_2 = a_2 - a^2$$

$$m_3 = a_3 - 3a_2 \cdot a + 2a^3$$

$$m_4 = a_4 - 4a_3 \cdot a + 6a_2 a^2 - 3a^4$$

4. MEDIDAS DE DISPERSIÓN.

En los dos apartados anteriores definíamos una serie de medidas de tendencia central cuyo objetivo era sintetizar toda la información de que se disponía. En este apartado veremos hasta que punto, para una determinada distribución de frecuencias, estas medidas de tendencia central son representativas como síntesis de toda la información.

Medir la representatividad de estas medidas equivale a cuantificar la separación de los valores de la distribución respecto de dicha media. A la mayor o menor separación de los valores entre sí se le llama Dispersión o Variabilidad.

Llamaremos Medidas de Dispersión a los coeficientes que nos miden el grado de dispersión de la distribución de la variable.

Para una mejor clasificación, vamos a distinguir entre medidas de dispersión absolutas y relativas.

4.1. Absolutas.

4.1.1. Recorrido.

Una primera aproximación para medir la dispersión en una distribución es calcular su recorrido.

DEF Llamaremos Recorrido a la diferencia entre el mayor valor y el menor valor de una distribución.

$$R = x_n - x_1$$

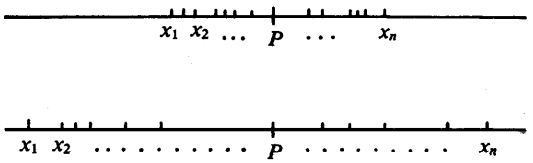
DEF Llamaremos Recorrido Intercuartílico a la diferencia existente entre el tercer cuartil y el primero.

$$R_I = C_3 - C_1$$

El Recorrido Intercuartílico nos indica que en el intervalo de longitud R_I están comprendidos el 50% central de los valores. Si R_I es pequeño, siempre en términos relativos de acuerdo con las unidades en que venga dada la distribución, podemos intuir una pequeña dispersión.

4.1.2. Desviaciones.

Supongamos que tenemos un promedio P del que vamos a estudiar su representatividad. Consideremos que tenemos dos distribuciones que originan este mismo promedio P (supongámoslas de frecuencias unitarias por sencillez) y que son tales como las que se representan en el siguiente gráfico:



Si queremos saber cual de los dos promedios es más representativo, a simple vista parece que el primero, porque el error que se comete utilizando P (en lugar de los valores de la distribución) es menor en la primera que en la segunda. Luego, cuanto más agrupados estén los valores alrededor del promedio, más útil será.

Para poder medir esto en una distribución genérica tenemos que considerar las desviaciones de cada valor con respecto al promedio, pero para evitar errores, se tomarán en valor absoluto o elevadas al cuadrado.

4.1.2.1.Desviación media respecto a la Media Aritmética.

Tomamos ahora como promedio genérico P la media aritmética $\bar{x}$ y tomaremos las desviaciones en valor absoluto. Así pues, tendremos

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

que es la desviación media respecto de la media aritmética. Un valor grande nos dirá la existencia de una gran dispersión en la distribución.

La desviación media respecto la media aritmética se puede definir como la media aritmética de los valores absolutos de las diferencias entre los valores de la variable y la media aritmética.

4.1.2.2.Desviación Media respecto a la Mediana.

Si el promedio cuya eficacia queremos medir es ahora la mediana tendremos:

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

que es la desviación media respecto a la Mediana. Para un valor grande, la mediana no será representativa. En la mediana demostramos que:

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

era mínima para k=Me, luego se verifica que $D_{Me} < D_{\bar{x}}$. Cuando la distribución está agrupada en intervalos, para calcular Me seguíamos el criterio:

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

mientras que para $\bar{x}$ utilizábamos las marcas de clase. En esta doble operación utilizamos unas hipótesis de trabajo incompatibles. Para la Me la hipótesis era que los valores dentro del intervalo estaban distribuidos uniformemente, mientras que para $\bar{x}$, al utilizar las marcas de clase, se hace implícitamente la hipótesis de que todos los valores de cada intervalo son iguales a x_i. Debemos, en este caso, optar por una de las dos hipótesis para ambos cálculos.

Las desviaciones medias tienen un significado preciso como promedio de las desviaciones, aunque tienen el inconveniente de no ser adecuadas al cálculo algebraico.

4.1.2.3.La Varianza. Propiedades.

De todas las medidas de dispersión absolutas respecto a la media aritmética, la varianza y su raíz cuadrada, la desviación típica, son las más importantes.

DEF Llamamos Varianza a la medida de dispersión que surge como media aritmética de los cuadrados de las desviaciones de los valores de la variable a la media aritmética. Es decir, el momento de segundo orden respecto a la media aritmética. Se denota por S^2 y es:

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

Evidentemente, S^2 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.

En el caso extremo de que todos los valores de la variable fuesen iguales, la media coincidiría con el valor común de las mismas y las desviaciones serían todas nulas, dando $S^2=0$. En general, cuanto más dispersas sean las observaciones, mayores serán las desviaciones respecto de la media, y mayor el valor numérico de la varianza.

A continuación enunciamos unas propiedades que verifica la varianza. Las demostraciones son inmediatas y no las damos para no agrandar en exceso el tema.

PROP La varianza no es negativa.

PROP La varianza es la medida cuadrática de dispersión óptima.

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

PROP La varianza es igual al momento de segundo orden respecto al origen menos el de primer orden elevado al cuadrado.

$$S^2 = m_2 = a_2 - a^2$$

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

PROP Si en la distribución de frecuencias sumamos a todos los valores de la variable una constante, la varianza no varía.

PROP Al multiplicar los valores de una distribución de frecuencias por una constante k, la varianza queda multiplicada por el cuadrado de la constante.

También podemos utilizar como medida de dispersión respecto a la media el coeficiente

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

denominado Cuasivarianza.

4.1.2.4.Desviación Típica o Standard. Propiedades.

Así como las desviaciones medias vienen expresadas en las mismas unidades de medida que la distribución, la varianza no, lo cual dificulta su interpretación. Es por ello que aparece la desviación típica.

DEF Llamamos Desviación Típica a la raíz cuadrada, con signo positivo, de la varianza. Se representa por S, y es

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

Al ser la raíz cuadrada de la varianza, vendrá expresada en las mismas unidades de medida que la distribución, lo cual la hace más apta como medida de dispersión.

Sus propiedades las podemos deducir fácilmente de las propiedades de la varianza.

PROP La desviación típica no es negativa.

PROP La desviación típica es una medida de dispersión óptima.

PROP La desviación típica verifica $S = \sqrt{a_2 - a^2}$

PROP La desviación típica esta acotada superior e inferiormente.

PROP A la desviación típica no le afectan los cambios de origen.

PROP A la desviación típica le afectan los cambios de escala, siendo S'=k·S

DEF Diremos que una variable estadística X está Tipificada, Estandarizada o Reducida si su media es cero y su varianza es 1.

Dada una variable X, su tipificada Z, se define como $Z = \frac{X - \bar{x}}{S_X}$

4.2. Relativas.

Supongamos que tenemos dos distribuciones de frecuencias cuyos promedios son P_1 y P_2 y queremos saber cuál de las dos es más representativa. Esta comparación no la podemos efectuar por sus respectivas medidas de dispersión , ya que las distribuciones, en general, no vendrán dadas en las mismas unidades de medida. Tampoco se podrá efectuar en el caso de que las unidades de medida sean las mismas, si los promedios son numéricamente diferentes.

Por tanto, resulta necesario construir medidas adimensionales. Estas medidas de dispersión, llamadas relativas, siempre vendrán dadas en forma de cociente.

Podemos destacar:

DEF Llamamos coeficiente de Apertura a la relación por cociente entre el mayor y menor valor de una distribución.

$$A = \frac{x_n}{x_1}$$

Este coeficiente es el más fácil de calcular, pero presenta varios inconvenientes. Entre ellos podemos nombrar:

1) Mide la dispersión de la distribución sin hacer referencia a ningún promedio, por lo que no resuelve el problema de la comparación entre éstos.
2) Sólo tiene en cuenta dos valores de la distribución (los dos extremos), lo que nos dará una gran dispersión en el caso de que estén muy separados.

DEF Llamamos Recorrido Relativo al cociente entre el recorrido y la media aritmética.

$$R_r = \frac{R_e}{\bar{x}}$$

Nos indica el número de veces que el recorrido contiene a la media aritmética.

DEF Llamamos Recorrido Semi-intercuartílico al cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil.

$$R_S = \frac{C_3 - C_1}{C_3 + C_1}$$

4.2.1. Coeficiente de Variación de Pearson.

Para poder comparar las medias aritméticas de dos distribuciones que vengan dadas en unidades diferentes tenemos el coeficiente de variación de Pearson.

DEF Definimos el coeficiente de variación de Pearson como la relación por cociente entre la desviación típica y la media aritmética.

$$V = \frac{S}{\bar{x}}$$

En primer lugar, tenemos que dicha medida es adimensional. En segundo lugar, V representa el número de veces que S contiene a $\bar{x}$. Cuanto mayor sea V, más veces contendrá S a $\bar{x}$, luego relativamente a mayor V menor representatividad de $\bar{x}$.

Este coeficiente se suele expresar en tanto por ciento, siendo

$$V = \frac{S}{\bar{x}} \cdot 100$$

Como tanto en S como en $\bar{x}$ han intervenido todos los valores de la distribución, V presenta la garantía de que utiliza toda la información.

La cota inferior de V es cero, al ser éste el menor valor que puede tomar S, y es el valor de V que indica la máxima representatividad de $\bar{x}$.

En caso de que la media aritmética sea nula, el valor de V no es significativo, ya que su resultado numérico nos puede hacer tomar conclusiones estadísticamente equivocadas.

4.2.2. Índice de Dispersión respecto a la Mediana.

Para comparar medianas podemos definir un coeficiente similar a V.

DEF Definimos el índice de dispersión respecto a la mediana como:

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

5. MEDIDAS DE FORMA. ASIMETRÍA Y CURTOSIS.

En los apartados anteriores hemos realizado el análisis estadístico sintetizando la información mediante medidas de posición y visto la dispersión en la distribución. Pero analizar los datos no consiste sólo en hallar una media y una varianza. En este apartado vamos a ver una tipología de distribuciones según la forma de su representación gráfica. El motivo es que, aunque resumamos la distribución mediante medias, no debemos

proceder a una interpretación que implique un comportamiento de todos los elementos uniformemente constante e igual a la media.

Las medidas de la forma de la distribución se pueden clasificar en dos grandes grupos: medidas de asimetría y medidas de curtosis.

5.1. Asimetría.

Las medidas de asimetría se dirigen a elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta la distribución, sin necesidad de llevar a cabo su representación gráfica.

Si representamos gráficamente la distribución y trazamos una vertical que pase por la media aritmética, diremos que ésta es simétrica si deja a ambos lados el mismo número de valores. Será asimétrica en caso contrario.

Vamos a buscar una medida que nos diga si la distribución es simétrica o no sin necesidad de representarla. Tomaremos la expresión

$$m_3 = \sum_{i=1}^{n} (x_i - \bar{x})^3 \frac{n_i}{N}$$

así, si:

$m_3=0$ la distribución es simétrica.
$m_3>0$ la distribución es asimétrica positiva.
$m_3<0$ la distribución es asimétrica negativa

Si la distribución es asimétrica a derechas o positiva, sería lógico pensar que la suma de las desviaciones positivas será mayor que la suma de las desviaciones negativas. En caso de que la distribución sea asimétrica a la izquierda o negativa, lo anterior se repetirá, pero a la inversa.

Esta medida está expresada en las mismas unidades que las de la variable, pero elevadas al cubo, por lo que no es invariante ante un cambio de escala. Para poder obtener un indicador adimensional, debemos dividir la expresión anterior por una cantidad que venga en sus mismas unidades. Tomaremos como dicha cantidad el cubo de la desviación típica, obteniéndose así

DEF Llamaremos Coeficiente de Asimetría de Ficher a $g_1 = \frac{m_3}{S^3}$

Como S no es negativa, el signo de g_1 coincide con el de m_3 y entonces:

$g_1=0$ la distribución es simétrica.
$g_1>0$ la distribución es asimétrica positiva.
$g_1<0$ la distribución es asimétrica negativa

Otras medidas de asimetría que siguen el mismo criterio de signos que las dos anteriores son las siguientes:

DEF Llamamos Coeficiente de Asimetría de Pearson a $A_p = \dfrac{\bar{x} - Mo}{S}$

DEF Llamamos Coeficiente de Asimetría de Bowley a $A_B = \dfrac{C_3 + C_1 - 2Me}{C_3 - C_1}$

DEF Llamamos Coeficiente Absoluto de Asimetría a

$$A = \frac{(C_3 - C_2) - (C_2 - C_1)}{S} = \frac{C_3 + C_1 - 2C_2}{S} = \frac{C_3 + C_2 - 2Me}{S}$$

5.2. 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. 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 vamos a tomar como modelo de referencia.

Esta distribución es la llamada distribución Normal, que corresponde a fenómenos muy corrientes en la naturaleza, y cuya representación gráfica es una campana de Gauss, dada por la fórmula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\alpha)^2}{\sigma^2}}$$

donde α y σ son la media y desviación típica, respectivamente.

Se trata de ver la deformación existente entre una distribución, en sentido vertical, y la Normal.

Diremos que una distribución puede ser más apuntada que la normal si es más alta y recibe el nombre de Leptocúrtica. En caso contrario recibe el nombre de Platicúrtica. La propia distribución normal recibe el nombre de Mesocúrtica.

La idea del apuntamiento de una distribución surgió de la comparación de frecuencias de los valores centrales de la distribución considerada con la frecuencia de dichos valores en una distribución normal con media y varianza iguales a las de la distribución que se compara.

DEF Llamaremos Coeficiente de Apuntamiento o Curtosis a

$$g_2 = \frac{m_4}{S^4} - 3.$$

Si $g_2 = 0$ Mesocúrtica.
$g_2 > 0$ Leptocúrtica
$g_2 > 0$ Platicúrtica

BIBLIOGRAFÍA RECOMENDADA.

Introducción a la Teoría de la Estadística. Aut.: Mood/Graybill. Ed. Aguilar.

Introducción a la Probabilidad y la Medida. Aut. Procopio Zoroa. Ed. PPU

Algoritmo. Matemáticas II. Cou. Aut.: Vizmanos y Anzola. Edit. SM.

Estadística para Ingenieros. Aut.: Ramón Ardanuy y Quintín Martín. Edit.: Hespérides