TEMA 2

2.1.- VARIABLE ESTADÍSTICA.

2.2.- DISTRIBUCIONES DE FRECUENCIAS.

Al observar los elementos en una determinada población o muestra se presentan distintos valores de la variable considerada, cada uno de los cuales puede repetirse una o más veces.

- Frecuencia absoluta o repetición : Es el número de veces que se repite cada valor o dato de la variable. Se representa por ni.
- Frecuencia relativa: es igual a la frecuencia absoluta dividida por el número total de datos. Se representa por fi. fi = ni / N siendo N el número total de datos.
- Frecuencia absoluta acumulada. Se denota por Ni y es la suma de las frecuencias absolutas de todos los valores menores o iguales que dicho valor.
- Frecuencia relativa acumulada. Fi = Ni/N, es resultado de dividir cada frecuencia absoluta acumulada entre el número total de datos, o la acumulación de fi.

Cuando sobre el conjunto de valores observados de una variable se realizan, las operaciones de ordenación y agrupación de los valores que se repiten (determinación de la frecuencia del valor) se obtiene una tabla estadística de distribución de frecuencias, o tabla de frecuencias. A dicho conjunto de operaciones se le denomina tabulación.

Unidimensional → significa que se está observando exclusivamente una sola variable de cada elemento de una población.

La suma de todas las frecuencias relativas es igual a la unidad.

La última frecuencia relativa acumulada es la unidad.

Distribución de frecuencias de una sola variable. Llamamos distribución de frecuencias al conjunto de valores que ha tomado una variable con sus frecuencias correspondientes (Xi,ni). Para que quede determinada una distribución de frecuencias debemos conocer las diferentes Xi y cualquiera de las columnas de frecuencias, ya que el paso de una a otra es inmediato. Como normalmente la primera columna que obtenemos es la de las ni, representaremos una distribución de frecuencias como los diferentes valores que en cada caso toma el par (Xi, ni).

Para que dos distribuciones de frecuencias sean iguales han de ser iguales los diferentes Xi y sus frecuencias relativas fi.

Recorrido, intervalos y marcas de clase.

Vamos a distinguir dos tipos fundamentales de distribuciones de frecuencias: las no agrupadas en intervalos y las agrupadas en intervalos.

- *Distribución no agrupada en intervalos*: una vez recogida y tabulada la información, ésta se dispone asociando a cada valor su frecuencia. Si las frecuencias son todas iguales a 1, la distribución se denomina de frecuencias unitarias.
- *Distribución agrupada en intervalos*. Pero si el número de valores distintos que ha tomado la variable es suficientemente grande parece aconsejable para mayor comodidad en el tratamiento de la información, agrupar estos valores en clases o intervalos, teniendo en cuenta que lo que ganamos en manejabilidad lo perdemos en información.

En la agrupación hay 3 aspectos que debemos tener en cuenta:

- Que el máximo de información lo tenemos al recogerla, disminuyendo al realizar la operación de agrupación por intervalos.
- Que las distribuciones agrupadas en intervalos no se presentan realmente así, sino que es el investigador el que las agrupa para manejar los datos más fácilmente.
- Que al agrupar hay que tener en cuenta las frecuencias.

En general, representaremos una distribución de frecuencias agrupada en intervalos por el par (L_{i-1}, L_i; ni) siendo L_{i-1}: extremo inferior del intervalo

L_i: extremo superior del intervalo.

Para agrupar los datos en intervalos o clases, debemos comenzar determinando el *recorrido* o rango de la variable, que se define como la diferencia entre el mayor y el menor valor de la variable.

$Re = \text{máx } X_i - \text{mín } X_i$

Amplitud del intervalo $C_i = L_i - L_{i-1}$ (diferencia entre los extremos superior e inferior del mismo). Los intervalos pueden ser de amplitud (o longitud) constante o variable. Para el mejor tratamiento de la información, es más cómodo que sean de amplitud constante. Si la amplitud es constante se verificará que : Re = Nº intervalos * C_i . Esta relación nos permite deducir el nº de intervalos si fijamos la amplitud, o ésta última si fijamos el nº de intervalos.

En la fijación del nº de intervalos no existen reglas fijas (suelen oscilar entre 5 y 15), hasta el punto de que a veces se hacen varios ensayos.

Un intervalo queda especificado por sus extremos, en general, para el intervalo i-ésimo se representará por L_{i-1} - L_i. Aparece un problema cuando un valor de la variable coincide exactamente con un extremo de intervalo, con lo que hay dudas sobre su inclusión o no en un determinado intervalo. Por esto se establece, como regla general, que los intervalos sean abiertos por la izquierda y cerrados por la derecha, es decir del tipo (a,b], es decir que el intervalo se compone de todos los puntos comprendidos entre a y b, incluido b y excluido a.

Como representativo de cada intervalo o clase, elegimos su punto medio llamado "*marca de clase*" (X_i). Así, en el intervalo i-ésimo la marca de clase será el valor central del intervalo:

$X_i = L_{i-1} + L_i/2$

2.3.- REPRESENTACIONES GRÁFICAS.

Aunque el par de valores (Xi; ni) encierra toda la información disponible, parece útil traducirla en gráficos, de modo que su inspección visual sirva de punto de partida para el análisis estadístico.

A) ***Variables cuantitativas***.

A.1) **Distribuciones no agrupadas en intervalos**. En el caso de datos sin agrupar la representación gráfica se realiza fácilmente mediante un sistema de ejes de coordenadas cartesianas, representando en el eje de abscisas los valores de la variable y en el de ordenadas las frecuencias.

A.1.1.- *Diagrama de barras*. Consiste en levantar para cada valor de la variable Xi una barra cuya altura sea su frecuencia absoluta ni. De esta forma se obtiene una visión en conjunto de la distribución de frecuencias absolutas. Se puede hacer la representación utilizando frecuencias absolutas (ni) o relativas (fi) indistintamente, en el caso de utilizar frecuencias relativas la suma de las longitudes de las barras es igual a 1, lo que permite comparar gráficamente dos distribuciones de efectivos desiguales.

A.1.2.- *Diagramas acumulativos*. Estos diagramas permiten la representación gráfica de las distribuciones de frecuencias acumuladas (tanto absolutas N_i como relativas F_i). Se construye levantando sobre cada valor de la variable, (con altura igual a su frecuencia acumulada), y uniendo mediante trazos horizontales el extremo de cada ordenada con el siguiente. La última ordenada será de magnitud N (nº total de observaciones), y la ordenada correspondiente a un valor X_i de la abscisa indicará el nº de observaciones para los cuales la variable ha tomado valores menores o iguales a la abscisa. A partir de la última ordenada se suele prolongar con un trazo horizontal hacia la derecha, mientras que a partir del menor de la variable se prolonga sobre el eje X, hacia la izquierda. La interpretación del diagrama acumulativo es fácil. Por ejemplo, para X = 5,5 la ordenada vale 51, ya que existen 51 datos correspondientes a valores de la variable X que son menores a 5,5, concretamente a los valores 0,1,2,3,4 y 5 que son los que toma la variable. Este diagrama nos muestra conjuntamente el crecimiento, en términos de frecuencia, que se ha dado en cada valor de la variable con respecto al anterior, en forma de saltos.

A.2) **Distribuciones agrupadas en intervalos**. Las gráficas utilizadas para este tipo de distribuciones de frecuencias son similares a las anteriores, en cuanto a que persiguen la misma idea, pero incorporando las modificaciones pertinentes.

Así, por ejemplo, el diagrama de barras experimenta la siguiente transformación: puesto que los valores de la variable están ahora agrupados en intervalos, la base deja de ser un punto o un único valor de la variable para ser ahora un intervalo, lo que implica que en lugar de levantar una barra ahora se levantará un rectángulo, de modo que si el diagrama de barras visualiza las frecuencias como longitudes, ahora el nuevo tipo de diagrama, que se denomina "histograma", lo hará mediante áreas de rectángulos.

Por otra parte, los diagramas acumulativos que ofrecían una forma de escalera, pasarán ahora a tener forma poligonal, y se suelen denominar "polígonos acumulativos".

A.2.1.- *Histograma o histogramas de frecuencias*. Se construye levantando sobre cada intervalo (L_{i-1}, L_i) un rectángulo cuya área sea su frecuencia absoluta n_i. Para ello, puesto que conocemos la base de cada rectángulo, habremos de determinar cómo se obtiene su altura d_i. Observemos que la base de ese rectángulo mide L_{i-1}- L_i = C_i, esto es, la amplitud del intervalo. El área del rectángulo levantado sobre el intervalo i-ésimo (L_{i-1}, L_i) vale:

A_i = (base) * (altura) = C_i * d_i; pero como el área debe ser la frecuencia absoluta del intervalo, se tiene:

$A_i = C_i * d_i$

$A_i = n_i$ $\rightarrow n_i = C_i * d_i \rightarrow d_i = n_i/c_i$

Así, pues, la altura de cada intervalo será su frecuencia absoluta dividida por su amplitud, cantidades que se conocen a partir de la tabla. Al resultado de esta operación, esto es a d_i, se le denomina "densidad de frecuencia del intervalo" o altura del intervalo. Por tanto, para construir un histograma basta con levantar sobre cada intervalo en el eje de abscisas un rectángulo cuya altura sea d_i. Se consigue así una gráfica que ofrece el modo de distribuirse la masa de datos entre los intervalos de valores de la variable, comparando áreas de rectángulos.

Existe un caso en el que la construcción del histograma puede simplificarse, en el sentido de que hay que realizar menos cálculos, es el caso en el que los intervalos son de amplitud constante, es decir $C_1 = C_2 = \ldots = C_k = C$, donde C es un número constante. En este caso, las densidades de frecuencia son las frecuencias absolutas divididas todas por el mismo número C, con lo que la representación gráfica de los rectángulos con altura n_i, resulta la misma que si la altura de los rectángulos fuese d_i, salvo que la unidad de medida en el eje de ordenadas es diferente, pero la impresión visual provocada es la misma. Por tanto, en el caso en que los intervalos sean de amplitud constante, no es necesario calcular las densidades de frecuencia d_i, sino que basta tomar como altura de los rectángulos la propia frecuencia absoluta n_i, bien entendido que ahora el área de los rectángulos no es n_i, sino que es proporcional, pero la forma del gráfico, que es lo que en último extremo interesa, es la misma.

Esta simplificación no puede llevarse a cabo cuando los intervalos son de amplitud variable, ya que la gráfica quedaría totalmente distorsionada. No olvidemos que, en este caso, las densidades de frecuencias se obtienen dividiendo las frecuencias absolutas entre las amplitudes, que ahora serían diferentes para cada intervalo.

A.2.2.- *Polígonos acumulativos*. La forma de construir los polígonos acumulativos es la misma que para los diagramas acumulativos pero con una diferencia proveniente del hecho de que los valores de la variable ahora están agrupados en intervalos. La diferencia radica en la forma en que se unen estos puntos para formar la gráfica. Puesto que los valores de la variable están agrupados en intervalos, no hay manera de saber la frecuencia concreta de cada valor individual de la variable en cada intervalo, y entonces se recurre a suponer que la frecuencia se distribuye de un modo uniforme a lo largo del intervalo en cuestión; en otras palabras, que todos los posibles valores de la variable contenidos en el intervalo poseen la misma frecuencia, y por tanto, la frecuencia acumulada variará de un modo lineal entre los extremos del intervalo. Así, pues, todos los puntos Li Ni se unirán entre sí mediante tramos rectos, dando como resultado una línea poligonal, de la que toma el nombre este tipo de representación gráfica. Aquí, no desempeña ningún papel la diferencia entre intervalos de amplitud variable o constante, en la construcción de los polígonos acumulativos.

B) **Atributos**

Los atributos se pueden representar utilizando el diagrama de barras, no obstante, puesto que los valores que toman los atributos no pueden ordenarse, es decir, no puede decirse si uno precede o no a otro de un modo único, y , por ello, no tiene sentido el obtener distribuciones de frecuencias acumuladas, no se usan los diagramas acumulativos.

Para los atributos se suelen usar también representaciones de tipo espacial, que consisten en repartir la superficie de alguna figura geométrica plana proporcionalmente a las frecuencias absolutas de cada uno de los valores que toma. Entre éstos, uno de los más conocidos es el diagrama de sectores. Estos diagramas consisten en repartir la superficie total de un círculo en tantos sectores circulares como modalidades tome el atributo, correspondiendo a cada valor un número de grados directamente proporcional a su frecuencia absoluta. En general se construye como sigue. Sea X un atributo que puede tomar las modalidades a_1,a_2, …,a_k de acuerdo con la siguiente tabla:

a_i	n_i
a_1	n_1
a_2	n_2
.	.
a_k	n_k

Repartiendo proporcionalmente los 360 º a las frecuencias absolutas, le corresponde al valor a_i

Número de grados $a_i = n_i/N*360$ º i = 1,2,…,k

Los diagramas de rectángulos tienen una base constante y una altura proporcional a la frecuencia absoluta correspondiente.

El pictograma. Se emplean dibujos alegóricos a la variable que se estudia. Suelen ser pocos precisos. Por ej maletas de diferente tamaño para representar el número de viajeros.

Algunas normas finales sobre interpretación de gráficas:

Las representaciones gráficas deben considerarse un instrumento auxiliar importante, pero no definitivo, para interpretar los datos y, por tanto extraer conclusiones. Un aspecto de gran relevancia a tener en cuenta en las representaciones gráficas es la subjetividad de quien lo elabora y la subjetividad de quien lo interpreta. Ambos pueden influir, de forma que las conclusiones extraídas por dos observadores distintos podrían conducir a conclusiones diferentes sobre el mismo hecho investigado. Por eso es fundamental estudiar previamente algunas características básicas en la elaboración del gráfico, por ejemplo, la escala de medida elegida para los ejes, el redondeo de datos, la amplitud de los intervalos si la distribución es agrupada, los aspectos relevantes que se desean destacar y otras cuestiones semejantes.

Básicamente 2 normas que deben tenerse siempre presentes al trabajar con gráficos:

1. Se debe tener siempre presente la escala de medida utilizada en el eje de ordenadas, ya que una misma distribución de frecuencias puede presentar aspectos muy diferentes en gráficos con diferente escala de medida.

2. Si las distribuciones de frecuencias corresponden a datos agrupados en intervalos, se debe tener en cuenta, que un cambio en la elección de los intervalos en que se agrupan los valores de la variable produce una distribución distinta y, por tanto, representaciones gráficas diferentes.

Las consideraciones anteriores no deben, en ningún caso, cuestionar la utilidad y fiabilidad de la mayoría de las representaciones gráficas. Simplemente, intentan introducir la dosis de profesionalidad y prudencia que cualquier elaborador o usuario de gráficos debe tener en cuenta.