Estadística Descriptiva unidimensional

Generalmente se pueden distinguir dos fases en la realización de cualquier experimento. Una primera, que consiste en la observación y análisis de los hechos que acontecen, y otra segunda, de interpretación y obtención de soluciones.
Cuando la experiencia se realiza en un contexto de incertidumbre, los resultados dependen del azar y nos encontramos ante lo que se denomina fenómeno aleatorio.

1.1. Estadística Descriptiva. Conceptos Generales.

Cuando se analiza un fenómeno aleatorio, ya sea de ingeniería o de otra rama del conocimiento humano, aparecen resultados del mismo que han de ser tratados de forma conveniente para conocer mejor los propios resultados obtenidos y el fenómeno en cuestión. Para ello será necesario emplear los resultados de la Estadística Descriptiva, cuya definición se presenta a continuación.

DEFINICION

*Se entiende por **Estadística Descriptiva** el conjunto de conceptos y técnicas que proporcionan una descripción numérica, ordenada y simplificada, a veces con la ayuda de representaciones gráficas, de la información obtenida en la observación y recogida de datos de un fenómeno aleatorio. Además, proporciona la base necesaria para construir los modelos matemáticos teóricos de los fenómenos aleatorios.*

En la observación y recogida de datos de un fenómeno aleatorio subyacen los siguientes conceptos.

Población Estadística es cualquier conjunto de personas, animales, objetos o acontecimientos sometido a estudio estadístico y que debe estar perfectamente definido, tanto en el tiempo como en el espacio.

Individuo o Unidad Estadística es cada uno de los elementos que componen una población estadística. Puede ser descrito por uno o varios caracteres, dependiendo de cuál sea el objeto del estudio estadístico.

Cualquier carácter o característica que se vaya a analizar en las unidades estadísticas presentará dos o más niveles exhaustivos e incompatibles, de forma que cada unidad estadística de la población presente uno y solamente uno de los niveles del carácter. Además, los caracteres se pueden clasificar en:

- **Cualitativos**: también denominados atributos. Los niveles de un atributo se llaman modalidades.
- **Cuantitativos**: también denominados variables estadísticas. Los niveles de una variable estadística se llaman valores observados.

Dentro de las variables estadísticas los caracteres pueden ser de dos tipos:

- **Discretos**: si hay una cantidad finita o infinita numerable de valores observados de la variable.
- **Continuos**: si la variable puede tomar cualquier valor de un intervalo.

Muestra es cualquier subconjunto observado de una población estadística, cuyo número de elementos se llama tamaño de la muestra.

Censo es la observación de todos los elementos de una población estadística.

EJEMPLOS

- En la población estadística de vehículos fabricados en una determinada factoría durante un día, se puede analizar el carácter cualitativo color, que presentará varias modalidades, así como los caracteres cuantitativos o variables: número de defectos en las pruebas finales a realizar en los vehículos antes de enviarlos al mercado (discreto) y consumo de combustible en una prueba simulada de circulación urbana (continuo).

- En la población estadística de estudiantes matriculados en un curso se puede analizar el carácter cualitativo sexo, que presenta dos modalidades; así como los caracteres cuantitativos edad (discreto) y estatura (continuo).

1.2. Organización de un conjunto de datos de una Variable Estadística en una Tabla de Frecuencias

Supongamos que se ha recogido un conjunto de datos sobre un carácter de las unidades de un colectivo. Por ejemplo, salarios de un grupo de trabajadores, número de horas que han estado 1000 vehículos aparcados en un aparcamiento, número de televisores vendidos cada día en una tienda a lo largo de un año, duración de una serie de bombillas, consumo de electricidad en una vivienda durante un trimestre, rendimiento de una máquina eléctrica, etc…Es necesario digerir y procesar tal información de manera que podamos extraer conclusiones sobre el comportamiento de dicho colectivo y, a veces, compararlo con el de otro colectivo con respecto a la misma variable.

Consideremos $x_1, \ldots, x_n$ los datos recogidos de una variable estadística, y $x_1, \ldots, x_k$ los diferentes valores observados de la variable.
Designamos por n_i al número de veces que aparece el dato x_i, lo que se denomina ***frecuencia absoluta*** del valor x_i, siendo $i=1, \ldots, k$.
Designamos por f_i a la proporción de veces que aparece el valor x_i, entre los n datos, lo que se llama ***frecuencia relativa*** del valor x_i. Es decir, el cociente entre la frecuencia absoluta y el número total de observaciones realizadas

$$f_i = \frac{n_i}{n}\ , i:1,2,.....k$$

Llamamos ***frecuencia absoluta acumulada*** en el valor x_i a la suma de las frecuencias absolutas de los valores inferiores o iguales a él y es denotada por N_i.

$$N_i = \sum_{j=1}^{j=i} n_j = N_{i-1} + n_i\ , i:1,2,.....k$$

Es claro que $N_1=n_1$ y que $N_k=N$.

Llamamos ***frecuencia relativa acumulada*** en el valor x_i al cociente entre la frecuencia absoluta acumulada y el número de observaciones
realizadas, es denotada por F_i

$$F_i = \frac{N_i}{n} = \sum_{j \le i} f_i\ , i:1,2,.....k$$

Se llama ***tabla de frecuencias o distribución de frecuencias*** a una tabla con cinco filas o columnas en la que la primera fila o columna contiene los valores diferentes en observación, ordenados de menor a mayor, la segunda contiene las frecuencias absolutas de dichos valores, la tercera las frecuencias relativas, en la cuarta las frecuencias absolutas acumuladas y en la última las frecuencias relativas acumuladas. Su estructura será, por tanto, la siguiente: Si se utilizan filas:

x	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
……	……	…..	…..	……
x_k	n_k	f_k	N_k	F_k

En las observaciones realizadas puede ocurrir que la variable estadística tome pocos valores diferentes, ya sea grande o pequeño el tamaño muestral, o bien puede pasar que la variable tome muchos valores diferentes, lo cual suele ocurrir para tamaños muestrales grandes. En el primer caso confeccionaremos la tabla de frecuencias como hemos visto anteriormente. En la segunda situación trataremos de agrupar los valores de la variable estadística en intervalos.

Agrupación de un conjunto de datos en intervalos

Los intervalos cuando agrupamos datos se denominan ***intervalos de clase***, los extremos de los intervalos se llaman ***extremos de clase*** y los puntos medios de los intervalos reciben el nombre de ***marcas de clase***.
A la hora de construir una tabla de frecuencias con datos agrupados se debe intentar seguir los siguientes pasos:

- La tabla de frecuencias no puede presentar intervalos de clase con frecuencia nula, ya que supondría una ruptura artificial en la representación de la frecuencia, fruto de una inadecuada agrupación por intervalos, que no se correspondería, con la manera en que se presenta la frecuencia de aparición de la variable estudiada. Es decir, no puede haber intervalos vacíos.

- Siempre que sea posible las clases deberán tener la misma longitud, con el fin de no enmascarar la realidad del fenómeno. No obstante, cuando se manejan datos de una magnitud económica este criterio no suele ser posible cumplirlo y hay que plantear intervalos de longitud diferente.

- El extremo superior del último intervalo debe ser mayor que el mayor valor observado.

- Cuando el dato más pequeño (resp. grande) se encuentra muy alejado del resto, se dirá que se trata de una observación anómala o extrema, (outlier).

- Se recomienda elegir los intervalos de clase de forma que sus marcas de clase coincidan con datos observados.

- Los extremos de clase de los intervalos se deben definir con precisión, de forma que los intervalos sean contiguos, pero no solapados. Así, una observación queda perfectamente encasillada en sólo un intervalo. Generalmente los intervalos se cierran por los extremos inferiores.

- La agrupación de los datos en intervalos de clase supone una pérdida de información, al no tratar de forma directa las observaciones, pero esta pérdida de información es compensada por la comodidad y facilidad de interpretación de la tabla de frecuencias.

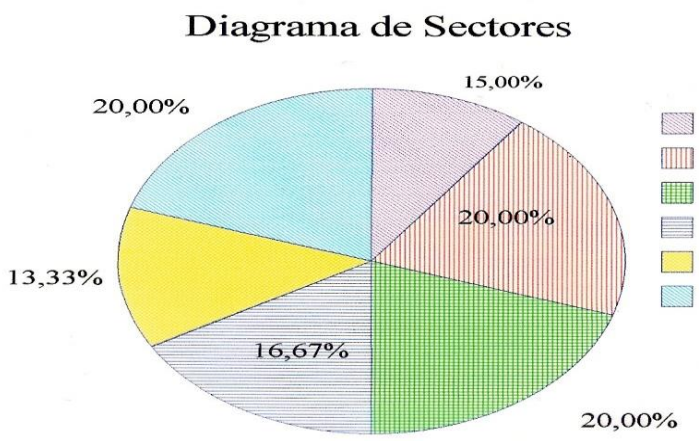
Intervalos de clase	Marcas de clase	n_i	f_i	N_i	F_i
	x				
$[a_1, a_2)$	x_1	n_1	f_1	N_1	F_1
$[a_2, a_3)$	x_2	n_2	f_2	N_2	F_2
……	……	……	…..	…..	……
$[a_k, a_{k+1})$	x_k	n_k	f_k	N_k	F_k

1.4 Representaciones gráficas de una Tabla de frecuencias.

1.4.1. Caso de un atributo

- Diagrama de sectores

  Consideramos un círculo cuya longitud de arco o ángulo equiparamos al número de observaciones. Así asignamos a cada modalidad un sector de longitud de arco, ángulo o área proporcional a su frecuencia.



1.4.2. Caso de una Variable cualitativa

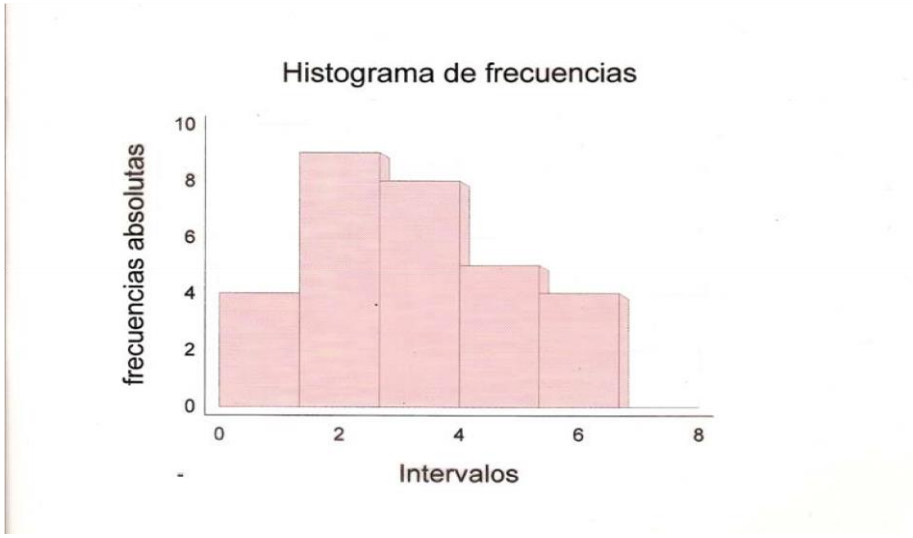
- **Diagrama de barras**

  Se utiliza para representar frecuencias absolutas o relativas cuando el conjunto de datos no está agrupado, las alturas de las barras deben ser proporcionales a las frecuencias por lo que la suma de estas alturas será n ó 1 dependiendo de que se tomen frecuencias absolutas o relativas. También se puede construir de carácter acumulado.

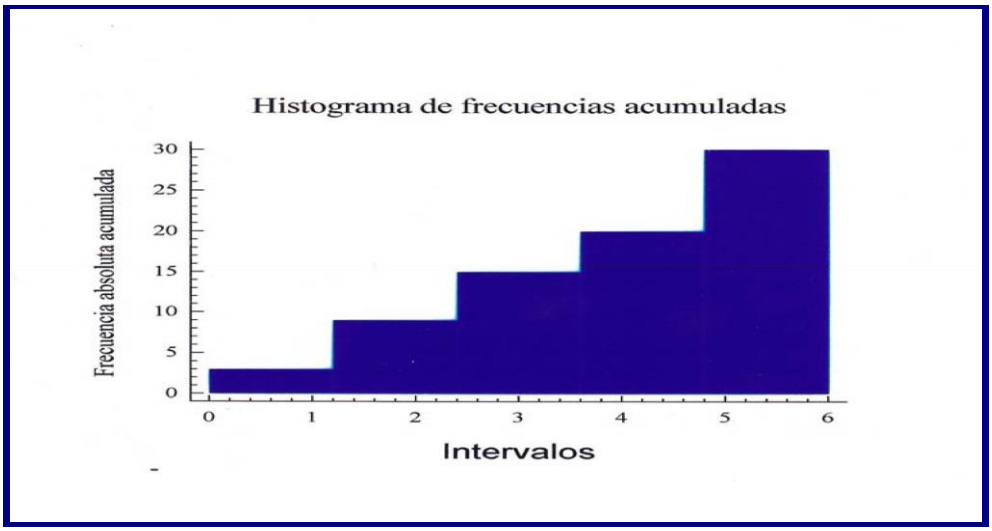
- **Histograma**

  Se utiliza para representar frecuencias absolutas o relativas cuando los datos están agrupados, sobre cada intervalo se levanta un rectángulo de área proporcional a la frecuencia de dicho intervalo. La suma de todas las áreas será n ó 1 según se representen frecuencias absolutas o

frecuencias relativas. Para las alturas de los rectángulos se tiene que $n_i=(a_{i+1}-a_i)h_i$ o bien $f_i=(a_{i+1}-a_i)h_i$



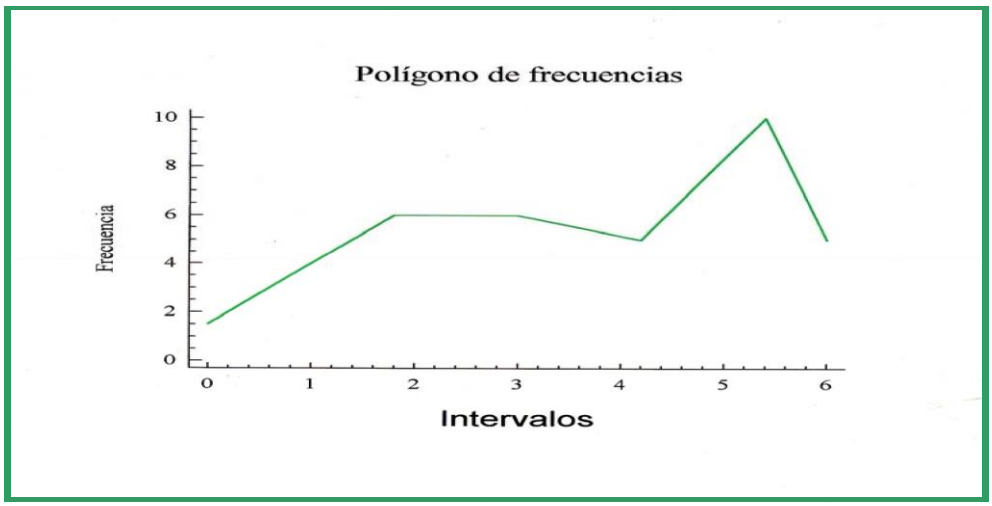
El histograma también puede representar frecuencias acumuladas, en este caso se construye levantando sobre cada intervalo un rectángulo de altura proporcional a la frecuencia que se acumula en dicho intervalo, tanto si los intervalos tienen la misma amplitud como si es distinta.



- **Polígono de frecuencias**

Los polígonos de frecuencias se construyen uniendo los extremos de las barras de un diagrama de barras (frecuencias relativas o absolutas) si los datos están sin agrupar, o bien, si los datos están agrupados y todos los intervalos de clase tienen la misma amplitud, uniendo los puntos medios de las bases superiores de los rectángulos de un histograma sin acumular incluyendo un intervalo anterior al primero y un intervalo posterior al último, de modo que la figure quede cerrada.

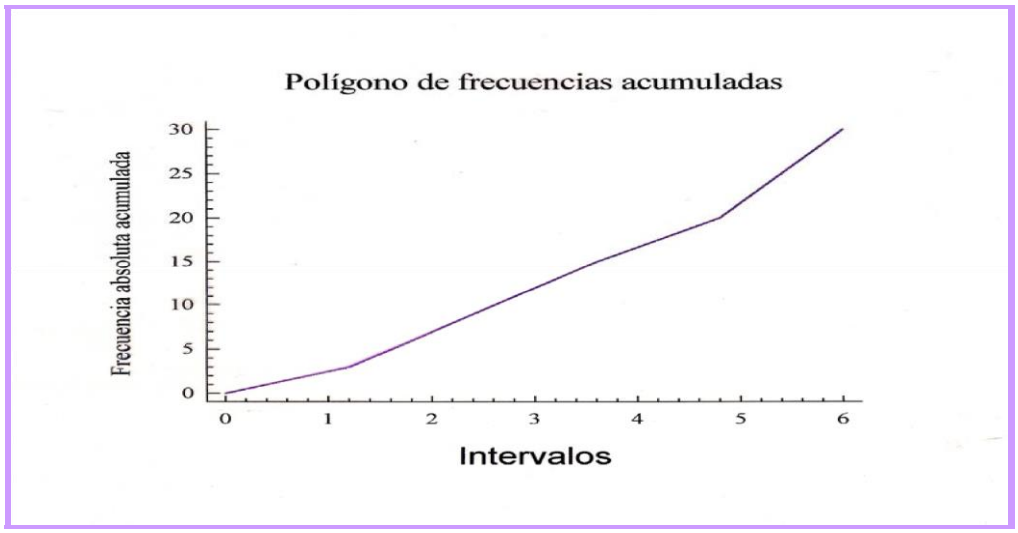
Si los datos están sin agrupar, la suma de las alturas será n ó 1, según sean las frecuencias absolutas o relativas. Si los datos están agrupados y todos los intervalos tienen la misma amplitud, entonces el área encerrada por el polígono será n ó 1, dependiendo de que la frecuencia representada sea absoluta o relativa.



Los polígonos también pueden ser acumulados, en este caso se obtienen uniendo los vértices superiores derechos de los rectángulos de un histograma acumulado, incluyendo un intervalo anterior al primero.

Al hacer esta representación por medio de rectas suponemos que en cada intervalo las observaciones están uniformemente distribuidas.

Con esta representación podemos ver el porcentaje de observaciones menores a cada valor, cosa que en el histograma no podíamos.



1.5. Medidas asociadas a una distribución de frecuencias

Una vez que se han recogido las observaciones, se han organizado en una tabla de frecuencias y representado gráficamente; procede resumir esa información a través de medidas que nos permiten conocer diferentes aspectos de la distribución de frecuencias como puede ser su comportamiento central o posición, su variabilidad o su forma. Estas medidas se llaman **estadísticos,** reservando el nombre de **parámetros** para sus correspondientes en la población.

Medidas de posición

Este tipo de medidas tienen por objeto el dar valores alrededor de los cuales se encuentran las observaciones muestrales. Entre las medidas más importantes se encuentran los diferentes tipos de medias, la mediana, la moda, los cuarteles y los percentiles.

Media aritmética: es el promedio de las observaciones muestrales.

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \sum_{i=1}^{k} x_i \cdot f_i$$

donde $x_1, x_2, \ldots, x_k$ son los diferentes valores observados, $n_1, n_2, \ldots, n_k$ las frecuencias absolutas de los mismos y $f_1, f_2, \ldots, f_k$ las frecuencias relativas.

Mediana: es el valor de la variable estadística (observado o no) que deja igual número de observaciones por debajo y por encima de ella, ordenadas las observaciones en forma creciente. Si el número de observaciones es impar se trata del dato central, si el número de observaciones es par se toma el promedio de los centrales.
Por tanto,

- Datos no agrupados:

$$Me = \begin{cases} x_{\left(\left[\frac{n}{2}\right]+1\right)} & si \;\; n \text{ es impar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & si \text{ n es par} \end{cases}$$

- Datos agrupados:

$$Me = a_i + \left(a_{i+1} - a_i\right) \cdot \frac{\frac{n}{2} - N_{i-1}}{n_i}$$

Centil de orden K: es el valor no necesariamente observado de la variable estadística que deja las k/100 partes de las observaciones por debajo o igual que él, con k=1, 2, …, 99, una vez ordenadas de forma creciente las observaciones.

- Si los datos no están agrupados:

$$C_k = \begin{cases} x_{\left(\left[\frac{kn}{100}\right]+1\right)} & si \;\; kn \notin \mathbb{Z} \\ \dfrac{x_{\left(\frac{kn}{100}\right)} + x_{\left(\frac{kn}{100}+1\right)}}{2} & si \; kn \in \mathbb{Z} \end{cases}$$

- Si los datos están agrupados:

$$C_k = a_i + \left(a_{i+1} - a_i\right) \cdot \frac{\frac{kn}{100} - N_{i-1}}{n_i}$$

Cuartiles: es el valor no necesariamente observado de la variable que deja las k/4 partes de las observaciones por debajo o igual que él, con k=1, 2, 3, una vez que las observaciones están ordenadas de forma creciente. Se tiene que $Q_1=C_{25}$, $Q_2=C_{50}=Me$ y $Q_3=C_{75}$.

Deciles: es el valor no necesariamente observado de la variable que deja las k/10 partes de las observaciones por debajo o igual que él, con k=1, 2, …, 9, una vez que las observaciones están ordenadas de forma creciente. Se tiene que $D_1=C_{10}$, …, $D_9=C_{90}$.

Moda: es el valor de la variable que tiene máxima frecuencia. No tiene por qué ser única; así, si hay dos modas, la distribución se llama bimodal, si tres, trimodal, etc.
En una distribución de frecuencia agrupada por intervalos, denominamos intervalo modal al intervalo de mayor frecuencia, como una primera aproximación se puede tomar la marca de clase como la moda.
De una forma más exacta se suele tomar la moda de la forma siguiente:

$$Mo = a_i + \left(a_{i+1} - a_i\right) \cdot \frac{d_1}{d_1 + d_2}$$

Observación: existen las siguientes relaciones entre la media, mediana y moda:

- Distribución simétrica unimodal: $\bar{x} = Me = Mo$
- Distribución asimétrica unimodal: $\bar{x} \neq Me \neq Mo$
- Empíricamente se ha comprobado que en distribuciones asimétricas $\left|\bar{x} - Mo\right| \cong \left|\bar{x} - Me\right|$.
  Por lo tanto, conociendo dos de ellas podemos conocer aproximadamente la tercera.

Medidas de dispersión

Las medidas de dispersión tienen como objeto el cuantificar si los datos están próximos o separados entre sí o respecto algún punto.

Momento de orden r respecto a la media o momento central: es el promedio de potencias de las diferencias entre las observaciones y la media.

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

Observación. Se tiene que $m_0=1$

Varianza: es el momento central de orden 2.

$$s^2 = m_2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^2$$

Desviación típica: es la raíz cuadrada positiva de la varianza.

$$s = \sqrt{s^2}$$

PROPIEDAD 1

La varianza se puede escribir como

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot {x_i}^2 - \bar{x}^2 = a_2 - a_1$$

Es decir, la varianza es la diferencia entre la media de los cuadrados y el cuadrado de la media.

Coeficiente de variación de Pearson: es una medida de dispersión relativa, no conlleva unidades. Su expresión es la siguiente:

$$CV = \frac{s}{|\bar{x}|} \; si \; \bar{x} \neq 0$$

Sirve para medir la variabilidad relativa una vez eliminado el efecto de la unidad utilizada.

Recorrido: es la diferencia entre el máximo valor observado y el mínimo.

$$R = x_n - x_1$$

Recorrido intercuartílico: es la diferencia entre el tercero cuartil y el primero. Nos da idea de la longitud de un intervalo central de valores que contiene el 50% de las observaciones.

$$RI = Q_3 - Q_1$$

Recorrido interdecílico: es la diferencia entre el noveno decil y el primero. Nos da idea de la longitud de un intervalo central de valores que contiene el 80% de las observaciones.

$$RID = D_9 - D_1$$