Tema 59

Técnicas de obtención y representación de datos. Tablas y gráficas estadísticas. Tendenciosidad y errores más comunes

59.1 Introducción

Como ya comentamos en el tema 57, habitualmente el propósito de la Estadística Aplicada es el de sacar conclusiones de una población en estudio, examinando solamente una parte de ella denominada muestra.

Este proceso, denominado Inferencia Estadística, suele venir precedido de otro, denominado Estadística Descriptiva, en el que los datos son ordenados, resumidos y clasificados con objeto de tener una visión más precisa y conjunta de las observaciones, intentando descubrir de esta manera posibles relaciones entre los datos, viendo cuales toman valores parecidos, cuales difieren grandemente del resto, destacando hechos de posible interés, etc.

También están entre los objetivos de la Estadística Descriptiva el presentarlos de tal modo que permitan sugerir o aventurar cuestiones a analizar en mayor profundidad, así como estudiar si pueden mantenerse algunas suposiciones necesarias en determinadas inferencias como la de simetría, normalidad, homocedasticidad, etc.

59.2 Conceptos fundamentales

Comenzaremos definiendo algunos conceptos propios de la terminología de la Estadística Descriptiva.

Población e individuo

Los fenómenos aleatorios se presentan en un mundo real formado por individuos, en los que se observa el fenómeno aleatorio en estudio. El conjunto de todos los individuos recibe el nombre de población.

Así, el conjunto de pacientes de los que se extraen los que van a ser sometidos a tratamiento, constituyen la población, siendo cada uno de ellos un individuo. Las parcelas forman la población de otro ejemplo, etc.

Como se ve, los términos población e individuo, no deben ser entendidos necesariamente en un sentido de población humana y persona humana, sino, respectivamente, como colectivo del que queremos sacar conclusiones y como elemento o unidad que compone la población.

Una cuestión muy importante es la de determinar con precisión lo que constituye la población ya que de ella se elegirán unos cuantos individuos con objeto de obtener conclusiones acerca de toda la población.

Así, en el primer ejemplo, puede considerarse como población la de los enfermos españoles que padecen la enfermedad en estudio, o la de los enfermos en todo el mundo, o alguna de las dos anteriores pero con individuos que tienen una edad comprendida entre dos valores determinados. La definición de lo que constituye la población depende del experimentador y de la naturaleza del problema que se investiga. No obstante, una vez definida, de ella se tomarán las observaciones y se deberán sacar las conclusiones.

Al conjunto de individuos que elegimos de la población lo denominaremos muestra.

Insistimos en que es muy importante el fijar la población con toda precisión, ya que solamente la obtención de una muestra representativa de la población permitirá obtener conclusiones fiables sobre ella.

Habitualmente la muestra representativa se obtendrá por un procedimiento aleatorio (es decir, de azar), lo cual permitirá medir los posibles errores en términos de probabilidades, pero insistimos en que lo importante es obtener una muestra representativa de la población sea o no por un procedimiento aleatorio. La ventaja de utilizar un mecanismo de azar es que éste nos garantiza que la muestra será representativa, mientras que con otros procedimientos, en general, no tendremos tal garantía. No obstante, una población suficientemente homogénea puede soslayar este mecanismo; así todos tenemos experiencias de situaciones en las que al ir a comprar un producto y pedir información sobre él (es decir, que nos enseñen una muestra del producto que queremos comprar) el dependiente elige un individuo de forma subjetiva como representativo de la población en estudio y nosotros consideramos que éste es lo suficientemente representativo de la población como para poder decidir sobre su adquisición; aunque, dicho sea de paso, a veces insistamos en comprar justamente el que nos ha enseñado.

Caracteres

Cada uno de los individuos de la población en estudio posee uno o varios caracteres. Así por ejemplo, si la población en consideración es la de los estudiantes de una determinada universidad, éstos poseerán una serie de caracteres, o si se quiere características, que permiten describirlo. Los caracteres en este ejemplo pueden ser "facultad en la que está matriculado", "curso que sigue", "sexo" , "edad", etc. Precisamente la observación de uno o más de esos caracteres en los individuos de la muestra es lo que dará origen a los datos.

Los caracteres pueden ser de dos clases: cuantitativos, cuando son tales que su observación en un individuo determinado proporciona un valor numérico como medida asociada, como ocurre por ejemplo con los caracteres "edad" o "curso que sigue", y cualitativos, cuando su observación en los individuos no suministra un número, sino la pertenencia a una clase determinada, como por ejemplo el "sexo", o la "facultad en la que está matriculado".

Modalidades de los caracteres

Consideremos un carácter cualquiera como por ejemplo el "gusto". Este carácter, al ser observado en un individuo (una sustancia), puede presentar cuatro posibilidades, es decir, es posible percibir cuatro sensaciones diferentes: dulce, amargo, salado y ácido. Pues bien, a las posibilidades, tipos o clases que pueden presentar los caracteres las denominaremos modalidades.

Las modalidades de un carácter deben ser a la vez incompatibles y exhaustivas. Es decir, las diversas modalidades de un carácter deben cubrir todas las posibilidades que éste puede presentar y además deben ser disjuntas (un individuo no puede presentar a la vez más de una de ellas y además debe presentar alguna de ellas).

Así, al estudiar algún carácter, como por ejemplo la raza, el investigador deberá considerar todas las posibles modalidades del carácter (todas las posibles razas), con objeto de poder clasificar a todos los individuos que observe.

La matriz de datos

Habitualmente, la información primaria sobre los individuos, es decir, la forma más elemental en la que se expresan los datos es la de una matriz, en la que aparecen en la primera columna los individuos identificados de alguna manera y en las siguientes columnas las observaciones de los diferentes caracteres en estudio para cada uno de los individuos, tal y como aparece en la tabla 2.1.

Dicha matriz recibe el nombre de matriz de datos.

	carácter 1	carácter 2	...	carácter p
individuo 1	•	•	...	•
individuo 2	•	•	...	•
...	...	...	...	...
individuo n	•	•	...	•

Tabla 2.1

Así, los datos correspondientes a una investigación llevada a cabo para el estudio de una posible contaminación radioactiva en un determinado lugar produjeron como resultado la matriz de datos de la tabla 2.2, en donde se recogen las observaciones de los caracteres "edad", "sexo", "cáncer", "caída anormal del cabello" y "profesión" en los 100 individuos seleccionados en la muestra.

	edad	*sexo*	*cáncer*	*caída cabello*	*profesión*
individuo 1	32	masculino	no	no	agricultor
individuo 2	29	femenino	no	no	maestra
...	...	...	...	...	...
individuo 100	61	masculino	sí	sí	agricultor

Tabla 2.2

En algunas ocasiones se reserva el nombre de matriz de datos a la obtenida de la anterior eliminando la primera columna.

Clases de datos

Es habitual denominar a los caracteres variables estadísticas o simplemente variables, calificándolas de cualitativas o cuantitativas según sea el correspondiente carácter, y hablar de los valores de la variable al referirnos a sus modalidades, aunque de hecho solamente tendremos verdaderos valores numéricos cuando analicemos variables cuantitativas.

Una variable estadística es discreta si sus valores posibles son valores aislados, y diremos que una variable es continua si sus valores posibles están en número infinito y a priori son cualesquiera en un intervalo de valores.

En ocasiones, con objeto de facilitar la toma de los datos, el investigador los agrupa en intervalos. Así por ejemplo, resulta más sencillo averiguar cuantos individuos hay en una muestra con una estatura, por ejemplo, entre 1.70 y 1.80 que medirlos a todos, en especial si tenemos marcas en la pared cada 10 cm.

Observemos, no obstante, que siempre se producirá una pérdida de información al agrupar los datos en intervalos y dado que hoy en día la utilización

del ordenador suele ser de uso corriente, un agrupamiento en intervalos es en general desaconsejable.

No obstante, por razones docentes admitiremos esta posibilidad, ya que precisamente el agrupamiento en intervalos traerá complicaciones adicionales en el cálculo de algunas medidas representativas de los datos.

Consideraremos, por tanto, tres tipos posibles de datos:

I. Datos correspondientes a un carácter cualitativo.

II. Datos sin agrupar correspondientes a un carácter cuantitativo.

III. Datos agrupados en intervalos correspondientes a un carácter cuantitativo.

59.3 Obención de datos

Realizar una encuesta significa proceder a la realización de ciertas averiguaciones o pesquisas, o, en términos más estadísticos a la obtención de cierta información.

La obtención de la información necesaria sólo es posible si se tiene fijado el objetivo u objetivos que fundamentarán la encuesta. Aunque los objetivos puedan fijarse inicialmente de forma difusa, es conveniente hacer un esfuerzo para expresarlos de forma cómoda. En este sentido, es conveniente disponer de una documentación inicial que permita establecer, aunque sea de "forma aproximada", al menos los siguientes elementos básicos en la realización de cualquier encuesta:

- Población a estudiar
- Objetivos específicos del trabajo: aspectos de la población que se desean observar y analizar, y expresarlos mediante variables
- Identificación de las cuestiones a cubrir con el estudio.

En todo trabajo estadístico es necesario conocer las limitaciones que se tendrán al realizarlo. Para la realización de una encuesta se considerarán los siguientes aspectos:

- Medios económicos: es necesario conocer el presupuesto disponible
- Medios materiales y humanos
- Temporalidad: tiempo del que se dispone.

59.3.1 Métodos de recolección

Para recolectar información en una encuesta existen métodos muy diversos, los cuales difieren en su aplicación, en la precisión de la información que obtienen o en su metodología. Además, el método de recolección incide en el diseño del cuestionario e incluso en su contenido.

Los métodos pueden agruparse en dos grandes bloques:

a) Auto-administrados

Entre otros puden citarse los siguientes: encuestas por correo, por fax, por disquete, transmisión de información (telemática, telefónica,...), contadores de audiencia, lápices ópticos, correo electrónico, red internet, etc.

b) Administrados

Observación directa, entrevista domiciliaria, telefónica o de intercepción,...

59.3.2 El cuestionario

El cuestionario es el medio que fija la comunicación entre el entrevistador y el entrevistado, así como un documento de trabajo para los codificadores, depuradores y personal que introduce los datos para su posterior tratamiento.

Debe satisfacer las exigencias del que realiza el estudio, del entrevistador (el que recoge la información) y del entrevistado (el que da la información).

59.4 Tablas estadísticas

La tabulación tiene como objetivo presentar de forma ordenada y clara la información referente a uno o más caractéres observados en una población.

Las tablas más simples son las que constan de una primera columna, donde se reflejan las distintas modalidades que presenta el carácter en estudio. A dicha columna se añaden una o más columnas a su derecha en las que indicamos sus respectivas frecuencias.

Consideremos una población estadística de n individuos descrita según el carácter C, cuyas k modalidades son $C_1, ..., C_k$. Designemos por n_i el número de individuos que presentan la modalidad C_i.

La frecuencia absoluta de la modalidad C_i es n_i, que representa el número de individuos que presentan dicha modalidad.

La frecuencia relativa, f_i, de la modalidad C_i es la proporción de individuos de la población que presentan dicha modalidad. Se obtiene, por tanto, dividiendo la frecuencia absoluta entre el total de elementos de la población:

$$f_i = \frac{n_i}{n}$$

Como las modalidades son incompatibles y exhaustivas, se tiene:

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} f_i = 1$$

Se llama frecuencia absoluta acumulada, N_i, de la modalidad C_i, a la suma de las frecuencias absolutas hasta la de la i-ésima modalidad:

$$N_i = n_1 + ... + n_i = \sum_{j=1}^{i} n_j$$

Se llama frecuencia relativa acumulada, F_i, de la modalidad C_i a la suma de las frecuancias relativas hasta la de la i-ésima modalidad:

$$F_i = f_1 + ... + f_i = \sum_{j=1}^{i} f_j$$

Esquemáticamente obtenemos una tabla estadística:

Modalidades	Frecuencias absolutas		Frecuencias relativas	
C		acumuladas		acumuladas
C_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
C_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\vdots$				
C_i	n_i	$N_i = n_1 + ... + n_i$	f_i	$F_i = f_1 + ... + f_i$
$\vdots$				
C_k	n_k	$N_k = n_1 + ... + n_k = n$	f_k	$F_k = f_1 + ... + f_k = 1$
Suma	n		1	

Caracteres cualitativos

Cuando el carácter C es cualitativo, la tabla estadística es de la forma general anterior.

Example 1 *Casos del lanzamiento de una moneda 25 veces:*

Modalidad	n_i	N_i	f_i	F_i
Cara	*13*	*13*	$\frac{13}{25}$	$\frac{13}{25}$
Cruz	*12*	*25*	$\frac{12}{25}$	$\frac{13}{25} + \frac{12}{25} = 1$
	25		1	

Caracteres cuantitativos

a) <u>Variable estadística discreta</u>

Este caso es igual que para caracteres cualitativos.

Example 2 *Resultados de lanzar un dado 50 veces:*

	n_i	N_i	f_i	F_i
1	*10*	*10*	$\frac{10}{50}$	$\frac{10}{50}$
2	*9*	*19*	$\frac{9}{50}$	$\frac{19}{50}$
3	*8*	*27*	$\frac{12}{50}$	$\frac{27}{50}$
4	*12*	*39*	$\frac{7}{50}$	$\frac{39}{50}$
5	*7*	*46*	$\frac{7}{50}$	$\frac{46}{50}$
6	*4*	*50*	$\frac{4}{50}$	1
	50		1	

b) Variable estadística continua

Cuando la variable estadística es continua, las modalidades del carácter son las clases de valores posibles definidas por los extremos del intervalo.

En general, si se designan por $e_0, ..., e_i, ..., e_k$ los extremos de clase, la clase i-ésima estará definida por el intervalo $[e_{i-1}, e_i[$.

La tabla estadística es, abreviadamente, del tipo:

Clase i-ésima	$[e_{i-1}, e_i[$	n_i	N_i	f_i	F_i

Si se quiere precisar en la tabla estadística,se pueden incluir las marcas de clase, las amplitudes de clase, las distancias entre marcas de clase, etc., y éstas entrarán a formar parte de la tabla. Definimos

- la $i-$ ésima marca clase como el punto medio de la $i-$ ésima clase.

$$c_i = \frac{e_i + e_{i-1}}{2}$$

- la distancia entre las marcas de clase i e $i+1$ es:

$$d_i = c_{i+1} - c_i = \frac{e_{i+1} - e_i}{2} - \frac{e_i - e_{i-1}}{2} = \frac{e_{i+1} - e_{i-1}}{2}$$

- la amplitud de la clase i por:

$$a_i = e_i - e_{i-1}$$

59.5 Representaciones gráficas

Aunque una tabla estadística encierra toda la información disponible, es necesario traducirla mediante gráficos para realizar síntesis visuales. Según la naturaleza del carácter estudiado se utilizan varios tipos de representaciones.

59.5.1 Caracteres cualitativos

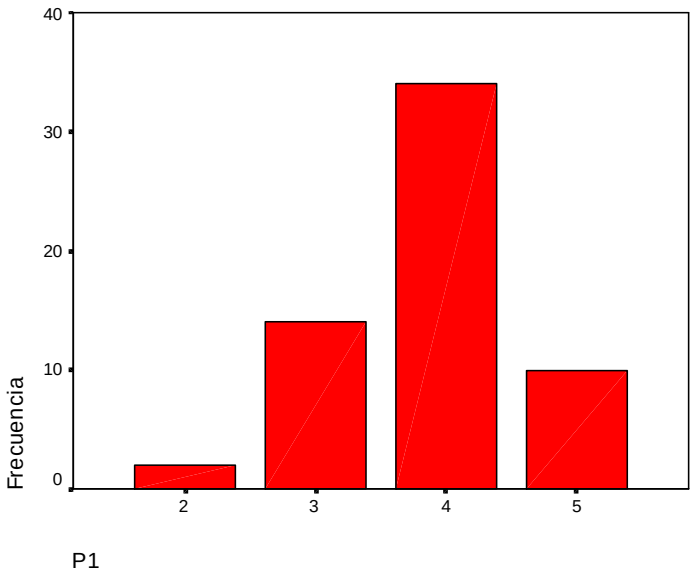
Por lo general, se utilizan figuras geométricas más o menos complicadas donde la idea importante a utilizar es la de proporcionalidad de las áreas a las frecuencias absolutas.

a) Diagrama de rectángulos

Tienen una base constante y una altura proporcional a la frecuencia absoluta correspondiente.

El lugar de colocación sobre el eje OX, la longitud de la base y la separación

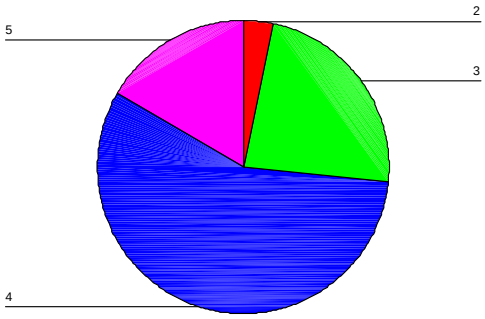
de una modalidad a otra, carece de todo significado que no sea estético.



b) <u>Diagrama de sectores</u>

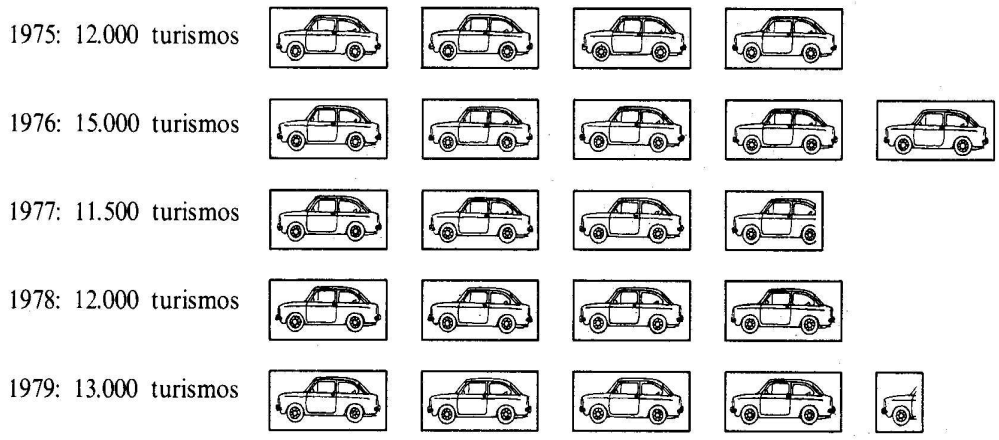
En esta sistema de representación, los sectores circulares tienen un ángulo central proporcional a la frecuencia absoluta correspondiente, y por consiguiente un área proporcional a la frecuencia absoluta.

$$\frac{2\pi}{n} = \frac{\alpha_i}{n_i} \text{ (donde } \alpha_i = \text{ángulo)} \Rightarrow \alpha_i = 2\pi \frac{n_i}{N} = 2\pi f_i$$



c) <u>Pictogramas</u>

Consiste en la representación de figuras alegóricas al carácter que se está estudiando tal que el área resultante sea proporcional a la frecuencia absoluta de cada modalidad.



59.5.2 Caracteres cuantitativos

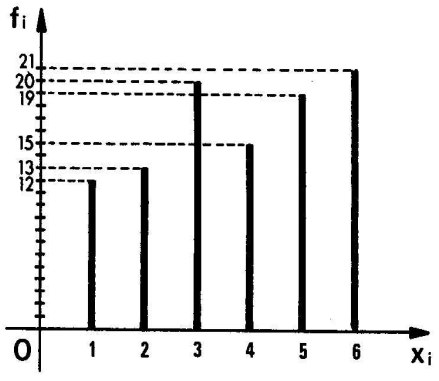
Si el carácter en estudio es cuantitativo, se utilizan dos clases de representaciones:

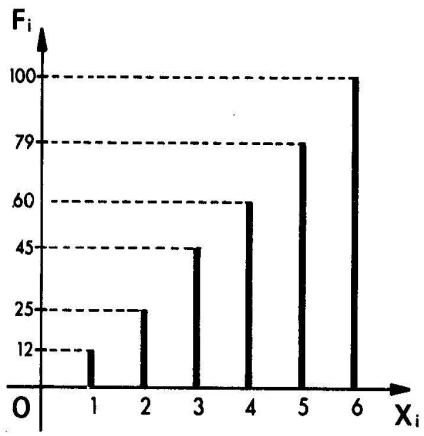
- Diagrama diferencial: diagrama de barras, histograma
- Diagrama integral: curva acumulativa o de distribución

Variables estadísticas discretas

a) Diagrama de barras

El diagrama de barras representa en función de los valores de las modalidades, las frecuencias relativas correspondientes.





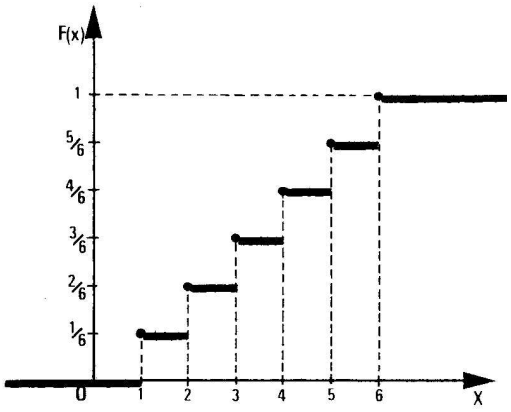
Como $\sum f_i = 1$, la suma de las longitudes de las barras es 1, lo que permite comparar gráficamente dos distribuciones de efectivos desiguales.

b) Curva acumulativa o de distribución (Diagrama de barras acumulativo)

Sea $F(x)$ la proporción de individuos de la población cuyo carácter es inferior a x. Esta función, llamada función acumulativa o de distribución, está definida para todo $x \in \mathbb{R}$ y es constante en cada intervalo entre dos valores posibles consecutivos. Así:

$$F(x) = \sum_{j=1}^{i} f_j \qquad \forall x_i < x \leq x_{i+1}$$

La curva de distribución o curva de frecuencias acumuladas es la curva representativa de $F(x)$ en función de x. Es una curva en escalera cuyos saltos corresponden a los valores posibles de x_i y son iguales a las frecuencias f_i. En la tabla, los valores de $F(x)$ corresponden a los de la columna de frecuencias relativas acumuladas.

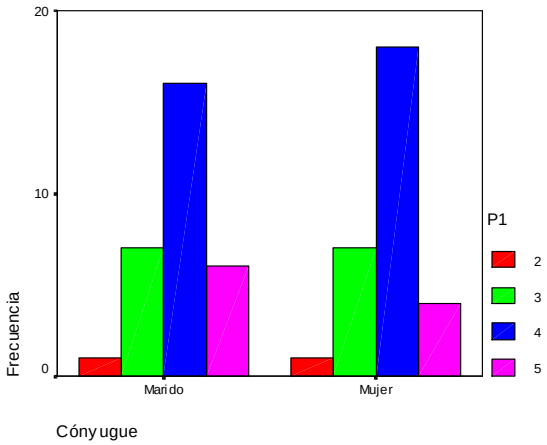


Variables estadísticas continuas

a) Histograma

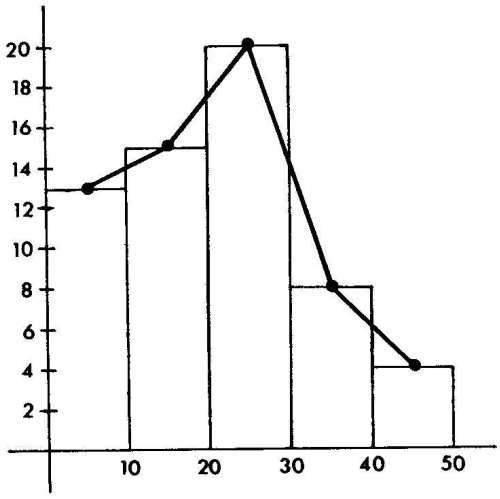
El histograma está formado por rectángulos yuxtapuestos o no, cuyas bases son las diferentes clases y cuyas alturas son las frecuencias medias $\frac{f_i}{a_i}, \frac{n_i}{a_i}$ (donde

$a_i = e_i - e_{i-1}$) por unidad de amplitud.



b) Polígono de frecuencias

Es la línea poligonal que resulta de unir los puntos correspondientes a las marcas de clase de los intervalos.



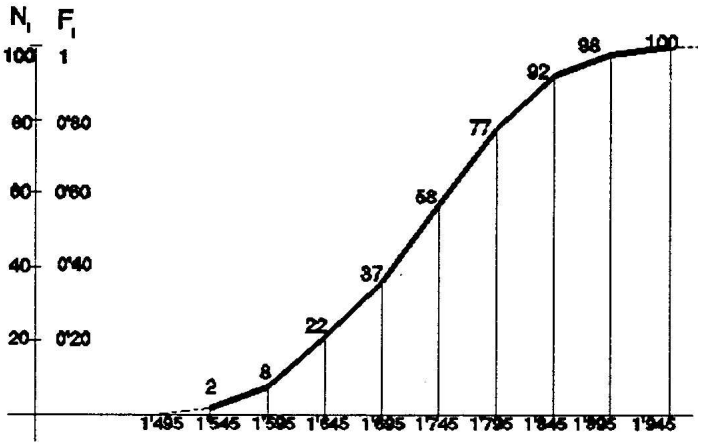
c) Curva acumulativa o de distribución (Histograma de frec. acumuladas)

Análoga al caso de variables discretas. La función acumulativa o de distribución $F(x)$ es la proporción de individuos de la población cuyo carácter es inferior a x. Esta función se conoce únicamente para los valores de x que son

extremos de clase: $x = e_0, e_1, ..., e_k$

$$F(e_i) = \sum_{j=1}^{i} f_j$$

La función de distribución es monótona no decreciente con $F(-\infty) = 0$ y $F(+\infty) = = 1$.



59.6 TENDENCIOSIDAD Y ERRORES MÁS COMUNES

En un estudio estadístico es importante que el informe que se haga del mismo sea suficientemente claro y exprese con precisión y exactitud los resultados obtenidos, de forma que los mismos no puedan ser malinterpretados.

Las causas que provocan la no utilidad de un informe estadístico pueden ser muy diversas. Esencialmente se pueden decribir tres tipos de errores:

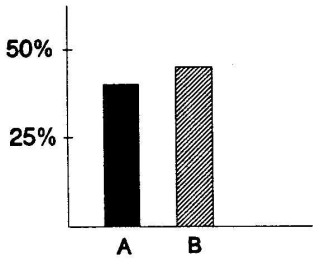
- Errores en la toma de datos
- Errores en la expresión de los datos
- Tendenciosidad de la expresión de los datos

Los errores en la toma de datos suelen deberse a que la muestra elegida no es representativa de la población, a que la técnica de obtención de los datos de un elemento de la muestra no sea adecuada, a no haber depurado adecuadamente los resultados, o en el caso de datos agrupados, que no se hayan elegido correctamente los intervalos de agrupamiento.

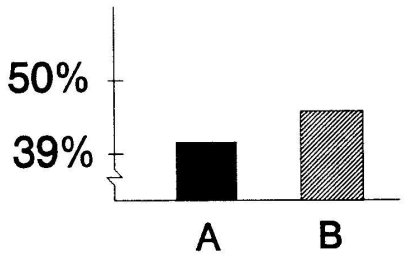
Dentro de este último caso el error más habitual es que los intervalos de agrupación no constituyan una partición del recorrido de la variable a estudiar. En algunos casos hay un dato que podría pertenecer a dos intervalos o un dato que no pertenezca a ninguno.

Cuando el informe se hace de forma gráfica es posible que la presentación del mismo no sea legible. En una representación gráfica el informe debe ser totalmente explicativo.

El principal problema que se puede presentar en una estadística es la tendenciosidad, intencionada o no. Consiste en que la información, aún siendo verdadera, se presenta a modo que puede inducir a error.



Gráfica correcta

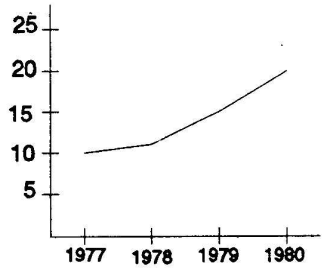


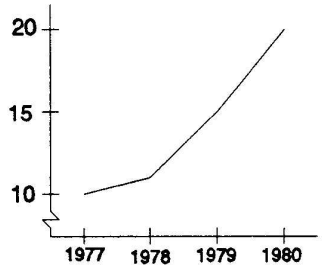
Gráfica tendenciosa

En la gráficas anteriores se representa un mismo estudio, pero con objeto de dar una mayor precisión a la figura se ha recortado la parte inferior de la gráfica de la izquierda para poder ampliar la figura. Si el que se presenta es el de la gráfica de la derecha, se obtiene un resultado tendencioso, ya que puede dar la impresión que la diferencia de ambos datos es mayor que la que realmente se da, por ser la barra B mucho mayor, en proporción, que la barra A.

Otra manifestación de la tendenciosidad es debida al hecho de que en muchos casos no existe una escala predeterminada a la hora de representar las variables que se estudian. Como consecuencia de ello las dos figuras de abajo pueden

representar el mismo estudio.





En la segunda gráfica da una sensación de crecimiento de la variable mucho mayor que en la primera.

En un pictograma gráfico el área de la figura que representa a cada variable ha de ser proporcional al valor de la variable. Un error que se suele dar es que lo que se hace proporcional es el diámetro de la figura, con lo que las desproporciones reales de los valores se elevan al cuadrado.

Esto puede conllevar una mala interpretación de los resultados, aunque se escriba el dato, por lo que la gráfica puede considerarse tendenciosa.

El último ejemplo que comentaremos sobre tendenciosidad es aquél en que, después de haber tomado los datos, éstos se agrupan de manera (normalmente intencionada) que un grupo grande de elementos se incluya cerca del extremo de uno de los intervalos, de modo que, al tomar estos valores como iguales a la marca de clase, el resultado se ve ligeramente alterado.