

Introducción a la Estadística

Tema 1: Variables estadísticas. Distribuciones de frecuencias

Tema 2: Distribuciones unidimensionales

Tema 3: Distribuciones bidimensionales

Tema 4: Regresión lineal

Tema 5: Números índice

LETANÍA

- Inteligencia, dame el nombre exacto de las cosas
- Astucia, dime qué errores no debo cometer
- Sabiduría, resista yo la dulce tentación de lo fácil
- Lucidez, asísteme en los momentos de pánico
- Estrategia, dime qué batallas no han de preocuparme
- Supervivencia, identifique yo al mortal enemigo
- Estupidez, no dé yo valor a lo que nada vale
- Fortaleza, dame sombra en el desierto
- Inmadurez, no te poses en mi hombro
- Desaliento, no serás mi confidente
- Miedo, sólo a ti temeré



netKEYNES.com

Introducción a la Estadística

Tema 1

Variables estadísticas Distribuciones de frecuencias

1.01	Estadística	2
1.02	Población. Muestra	2
1.03	Variable estadística. Atributo	3
1.04	Fases de un estudio estadístico	4
1.05	Tabulación de datos	5
1.06	Distribuciones de frecuencias	7
1.07	Gráfica de una tabla de datos	8
1.08	Gráfica de una tabla tipo II	8
1.09	Gráfica de una tabla tipo III	9
1.10	Gráficas para atributos	11

Nos vamos a meter en un jardín donde sólo hace falta saber sumar, restar, multiplicar y dividir



1.1 ESTADÍSTICA

Estadística descriptiva (LO QUE VAMOS A ESTUDIAR): conjunto de métodos necesarios para recoger, clasificar, representar y resumir **datos**.

Inferencia estadística (SE ESTUDIA EN CURSOS POSTERIORES): conjunto de métodos necesarios para hacer inferencias (sacar consecuencias) científicas a partir de los datos recogidos.

1.2 POBLACIÓN. MUESTRA

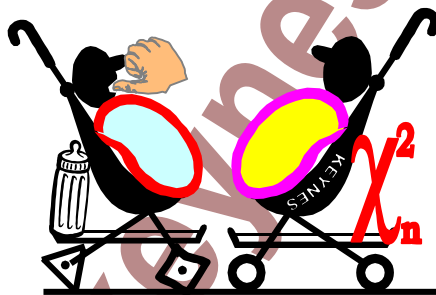
Llamamos **población** a todo conjunto de entes (personas, plantaciones de patatas, etc.) que poseen ciertas características comunes que se pretenden analizar.

Se llama **tamaño poblacional** al número de elementos que forman la población.

Llamamos **muestra** a todo subconjunto **finito** de elementos de la población. Se llama **tamaño muestral** al número de elementos que forman una muestra.

Cada característica "X" común a los elementos de la población se llama **carácter**, y permite **clasificar** los elementos de la población.

Por ejemplo, la población de los estudiantes de la Facultad, puede ser descrita mediante el carácter **edad**, mediante el carácter **sexo**, mediante el carácter **calificación en selectividad**, mediante el carácter **chuparse el dedo**, etc.



De cada forma en que puede presentarse un carácter, se dice que es una **modalidad** del carácter en cuestión.

Obvio: las modalidades de un mismo carácter han de ser incompatibles y exhaustivas: cada elemento de la población presenta una y sólo una de las diversas modalidades del carácter (mujer/hombre).

1.3 VARIABLE ESTADÍSTICA. ATRIBUTO

- Se dice que un carácter "X" es una **variable estadística** si es **cuantitativo**; o sea, **es medible**: a cada modalidad de "X" se le puede asociar un número, del que se dice que es el **valor de la variable estadística** (altura, peso, renta, calificación en selectividad).

Se dice que una variable estadística es **discreta** si su número de modalidades es finito (número de hijos en una familia) o infinito numerable (número de prospecciones que debe hacer Repsol hasta encontrar petróleo). Se dice que una variable estadística es **continua** si su número de modalidades es infinito no numerable (estatura, peso, renta).

- Se dice que un carácter "Y" es un **atributo** si es **cualitativo**; o sea, **no es medible** (sexo, tipo de Bachillerato estudiado). De cada una de las formas en que se puede presentar "Y", se dice que es una **modalidad**. A veces, para facilitar la gestión de la información relacionada con un carácter cualitativo, sus modalidades se **codifican**, sin que eso suponga cuantificación ni ordenación alguna.

ATRIBUTO: TIPO DE BACHILLERATO

Modalidades	Código
Ciencias Sociales	1
Ciencias de la Salud	2
:	:

La **Estadística Unidimensional** estudia la población atendiendo a un único carácter (estatura), y la **Estadística Multidimensional** la estudia atendiendo a más de un carácter (estatura y peso).



1.4 FASES DE UN ESTUDIO ESTADÍSTICO

Todo estudio estadístico consta de las siguientes fases:

- 1) Definición de los objetivos del estudio.
- 2) Elaboración de datos:
 - * Definición de la población a estudiar y del carácter a observar.
 - * Recogida de datos, eligiendo uno de los posibles métodos de observación de la población.
 - * Clasificación de los datos observados, atendiendo a los objetivos fijados.
 - * Presentación los datos observados mediante tablas y gráficos.
- 3) Utilización de los datos:
 - * Análisis de los datos observados, obteniendo nuevos datos (media, varianza, etc.) a partir de ellos.
 - * Interpretación del resultado del análisis.
 - * Predicción.

©netkeynes.com

1.5 TABULACIÓN DE LOS DATOS

A la hora de estudiar una variable estadística "X" (cuantitativa), tras observar la muestra seleccionada, cabe distinguir tres tipos de tablas.

- **Tipo I: Pocos datos y pocos valores de la variable.** Carecen de interés.
- **Tipo II: Muchos datos y pocos valores de la variable.**

Contamos el número de veces que se ha observado cada uno de los valores que puede tomar "X".

Por ejemplo, siendo "X" el número de aseos en las viviendas de un tipo de familias, en una muestra de tamaño 48 se observa:

1	2	2	1	1	1	3	1	2	2	1	3
4	1	2	3	1	3	2	2	2	1	1	1
3	3	3	4	1	1	1	3	2	2	3	2
1	4	3	3	2	1	2	1	2	1	2	1

⇒

N° de aseos	N° de viviendas
1	19
2	15
3	11
4	3

⇒

- **Tipo III: Muchos datos y muchos valores de la variable.**

Agrupamos los valores observados de "X" en **intervalos de clase** (no necesariamente todos con la misma amplitud) de la forma (a;b], considerando iguales los valores que pertenezcan a un mismo intervalo de clase; para ello, lo primero es determinar el **recorrido ó rango** $R_c = \text{Máx} \{x_i\} - \text{Mín.} \{x_i\}$ de la variable "X", que es la diferencia entre el mayor y el menor valor observado; así, si todos los intervalos de clase son de igual amplitud, ha de ser:

$$R_c = (\text{N° de intervalos}) \cdot (\text{Amplitud})$$

Por tanto:

$$\text{N° de intervalos} = \frac{R_c}{\text{Amplitud}}$$

$$\text{Amplitud} = \frac{R_c}{\text{N° de intervalos}}$$

A veces, para facilitar la gestión de los datos, el mayor y el menor valor observados de "X" se sustituyen por otros próximos que sean más cómodos; en tal caso, el recorrido o rango es la diferencia entre esos nuevos valores.

Del **punto medio de cada intervalo de clase**, que es la semisuma de los extremos del intervalo, diremos que es la **marca de clase** del intervalo en cuestión: será el número que **representará** a su correspondiente intervalo de clase la hora de hacer ciertas "cosas" de las que hablaremos más adelante.

Por ejemplo, siendo "X" la longitud (en milímetros) de un tipo de espárragos, en una muestra de tamaño 40 se observa:

160	169	173	167	166	172	169	166
170	173	187	181	179	180	177	168
163	169	162	173	161	170	164	167
172	167	164	190	167	173	163	179
170	168	180	179	166	167	168	174

Así las cosas, es claro que organizar una tabla del tipo II sería un petardo, pues cada valor observado se repite pocas veces.

El recorrido de la variable "X" es de 30 milímetros:

$$R_c = \text{Máx} \{x_i\} - \text{Mín.} \{x_i\} = 190 - 160 = 30$$

Por tanto, si elegimos los intervalos de clase con amplitud de 5 milímetros, resultan $30/5 = 6$ intervalos de clase.

Intervalos de clase	Marca de clase	Nº de observaciones
(160;165]	162'5	6
(165;170]	167'5	18
(170;175]	172'5	6
(175;180]	177'5	6
(180;185]	182'5	2
(185;190]	187'5	2

A veces, el primero y el último de los intervalos de clase se sustituyen respectivamente por otros de la forma $\leq c$ y $> d$ y se hace esto cuando el número de observaciones que pertenecen a esos intervalos es tan pequeño que el desglose en intervalos sería poco significativo. Cuando se hace tal cosa, la correspondiente tabla **sólo se emplea para presentar los datos observados**, pues al no estar determinada la amplitud de los intervalos primero y último, se reduce mucho la posibilidad de tratarlos numéricamente.

1.6 DISTRIBUCIONES DE FRECUENCIAS

La **frecuencia absoluta de un dato** es el **número de veces** que ese dato aparece en la muestra. Si los valores de la variable estadística se agrupan en intervalos de clase, la **frecuencia absoluta de un intervalo de clase** es el **número de datos** de la muestra que pertenecen a ese intervalo. Escribiremos n_i para denotar la frecuencia absoluta del dato x_i o del intervalo de clase $(L_{i-1}; L_i]$.

La **frecuencia relativa** de x_i ó $(L_{i-1}; L_i]$ es la **proporción de veces** que x_i ó $(L_{i-1}; L_i]$ se han observado en la muestra, y se denota f_i . Así, siendo la muestra de tamaño "N", es $f_i = n_i/N$. Naturalmente, la suma de todas las frecuencias relativas es la unidad. El número $p_i = 100 \cdot f_i$ es el porcentaje de veces que x_i ó $(L_{i-1}; L_i]$ se ha observado en la muestra.

La **frecuencia absoluta acumulada** e x_i ó $(L_{i-1}; L_i]$ se denota N_i , y es el **número de veces** que se han observado valores no superiores x_i ($\leq x_i$) o no superiores a L_i ($\leq L_i$); o sea: $N_i = n_1 + n_2 + \dots + n_i = N_{i-1} + n_i$

La **frecuencia relativa acumulada** de x_i ó $(L_{i-1}; L_i]$ se denota F_i , y es la **proporción de veces** que se han observado valores no superiores x_i ($\leq x_i$) o no superiores a L_i ($\leq L_i$); o sea: $F_i = f_1 + f_2 + \dots + f_i = F_{i-1} + f_i$

Llamamos **distribución de frecuencias** al conjunto de valores (o intervalos de clase) observados en la muestra, con sus correspondientes frecuencias; quedará determinada por dos columnas $(x_i; n_i)$ ó $((L_{i-1}; L_i]; n_i)$

Ejemplos

TABLA TIPO II

x_i	n_i	$f_i = \frac{n_i}{40}$	$N_i = n_1 + \dots + n_i$	$F_i = f_1 + \dots + f_i$
19	20	0'500	20	0'500
23	10	0'250	30	0'750
41	5	0'125	35	0'875
47	5	0'125	40	1
	40			

TABLA TIPO III

Intervalos de clase	n_i	$f_i = \frac{n_i}{50}$	$N_i = n_1 + \dots + n_i$	$F_i = f_1 + \dots + f_i$
(150;160]	10	0'2	10	0'2
(160;200]	20	0'4	30	0'6
(200;300]	15	0'3	45	0'9
(300;500]	5	0'1	50	1
	50			

1.7 GRAFICA DE UNA TABLA DE DATOS

Una vez tabulados los datos observados en la muestra, se emplean diversas representaciones gráficas de la tabla resultante (ya sea del tipo II o del tipo III), para así dar una visión de conjunto.

La representación gráfica de la **distribución de frecuencias absolutas o relativas** da lugar a los llamados **diagramas diferenciales**.

La representación de la distribución de **frecuencias absolutas o relativas acumuladas** da lugar a los llamados **diagramas integrales o acumulativos**.

1.8 GRAFICA DE UNA TABLA TIPO II

1) Diagrama diferencial: Diagrama de barras

En el eje de abscisas posicionamos los diversos valores x_i que puede tomar la variable estadística, y sobre cada uno levantamos un segmento o barra de altura igual a la frecuencia absoluta (relativa) de x_i en la muestra observada.

2) Diagrama integral: Diagrama acumulativo de frecuencias

Según la frecuencia acumulada (absoluta o relativa) representada, ubicamos en el plano los puntos $(x_i; N_i)$, $(x_i; F_i)$ y con ellos formamos una **escalera**.

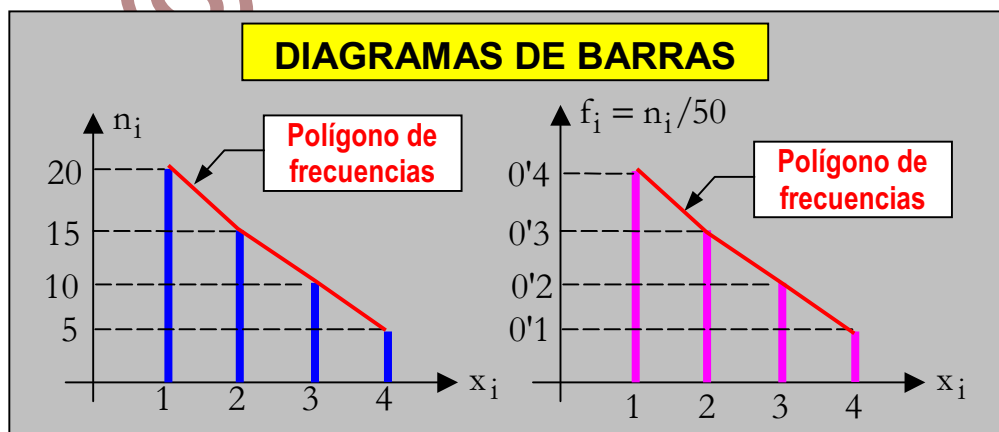
Ejemplo

Al observar el número de hijos en 50 familias se registran los siguientes resultados:

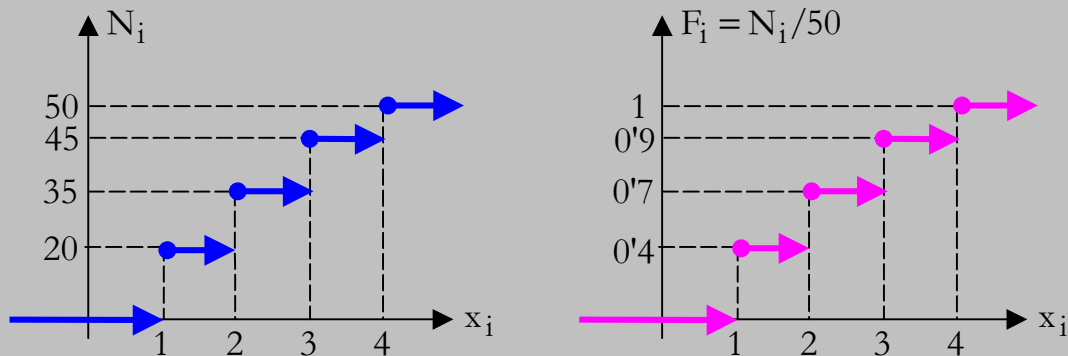
x_i	1	2	3	4
n_i	20	15	10	5

La siguiente tabla recoge la información necesaria para representar sus correspondientes diagramas.

x_i	n_i	$f_i = n_i/50$	$N_i = n_1 + \dots + n_i$	$F_i = f_1 + \dots + f_i$
1	20	0'4	20	0'4
2	15	0'3	35	0'7
3	10	0'2	45	0'9
4	5	0'1	50	1
	50	1		



DIAGRAMAS ACUMULATIVOS DE FRECUENCIAS



El que $N_i(2'6) = 35$ indica que, entre las 50 familias, hay 35 cuyo número de hijos no es superior a 2'6

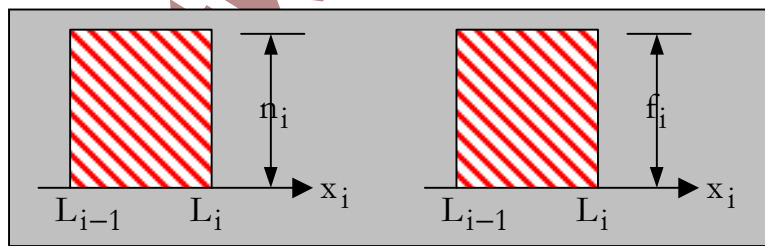
El que $F_i(2'6) = 0'7$ indica que 0'7 es la proporción de familias de la muestra cuyo número de hijos no es superior a 2'6

1.9 GRAFICA DE UNA TABLA TIPO III

1) Diagrama diferencial: Histograma

La frecuencia absoluta (relativa) de cada intervalo de clase se expresa mediante el área de un rectángulo cuya base es dicho intervalo de clase, y eso de modo que el área sea proporcional a dicha frecuencia absoluta (relativa).

- Si **los intervalos de clase tienen igual amplitud**, la altura del rectángulo correspondiente a cada intervalo es la frecuencia absoluta n_i (relativa $f_i = n_i/N$) de éste.



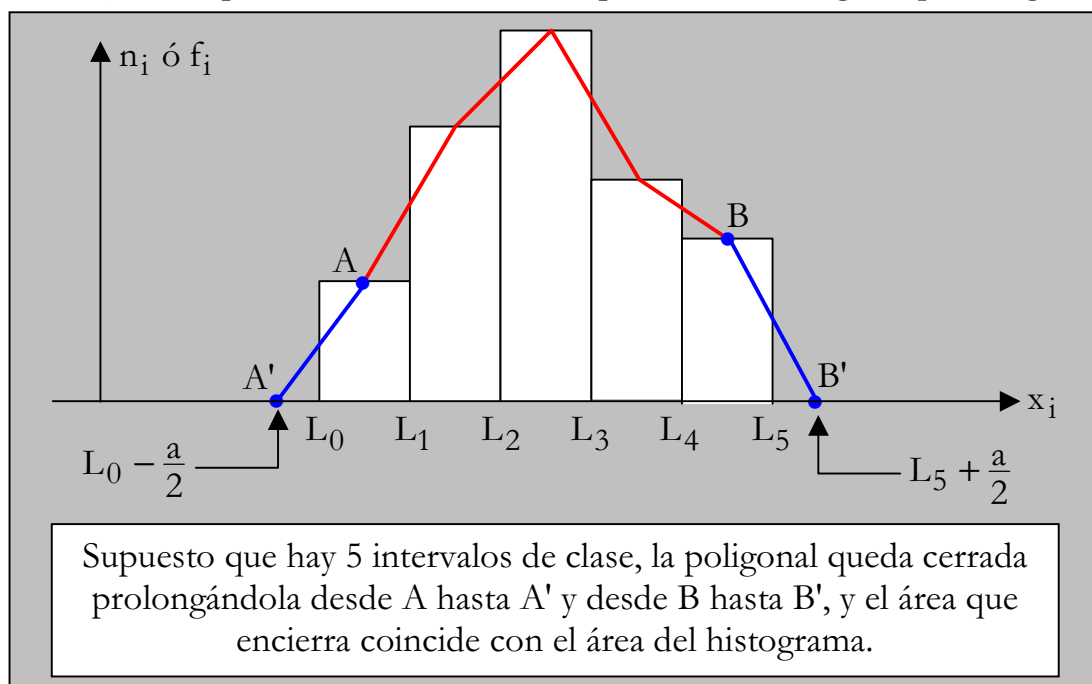
Siendo "N" el tamaño de la muestra, si la distribución está formada por "k" intervalos de clase de amplitud "a", el área del histograma de frecuencias absolutas es "a.N", y el área del histograma de frecuencias relativas es "a":

$$\sum_{i=1}^k a \cdot n_i = a \cdot \sum_{i=1}^k n_i = a \cdot N \quad ; \quad \sum_{i=1}^k a \cdot f_i = a \cdot \sum_{i=1}^k f_i = a$$

$$\sum_{i=1}^k n_i = N$$

$$\sum_{i=1}^k f_i = 1$$

El **polígono de frecuencias** se obtiene uniendo el punto medio de la base superior de cada rectángulo (corresponde a la marca de clase del su intervalo asociado) con el punto medio de la base superior del rectángulo que le sigue.



- Si **los intervalos de clase no tienen igual amplitud** y a_i es la amplitud del intervalo $(L_{i-1}; L_i]$, el área de su rectángulo asociado coincide con la frecuencia $\left\{ \begin{array}{l} \text{absoluta } n_i \\ \text{relativa } f_i \end{array} \right\}$ de $(L_{i-1}; L_i]$ si la altura del rectángulo es $\left\{ \begin{array}{l} n_i/a_i \\ f_i/a_i \end{array} \right\}$.

De $\left\{ \begin{array}{l} n_i/a_i \\ f_i/a_i \end{array} \right\}$ se dice que es la frecuencia $\left\{ \begin{array}{l} \text{absoluta} \\ \text{relativa} \end{array} \right\}$ **rectificada** de $(L_{i-1}; L_i]$; o que es la **densidad de frecuencia** $\left\{ \begin{array}{l} \text{absoluta} \\ \text{relativa} \end{array} \right\}$ de dicho intervalo.

Siendo "N" el tamaño de la muestra, si la distribución está formada por "k" intervalos de clase, el área del histograma de frecuencias absolutas es "N", y el área del histograma de frecuencias relativas es "1":

$$\sum_{i=1}^k a_i \cdot \frac{n_i}{a_i} = \sum_{i=1}^k n_i = N \quad ; \quad \sum_{i=1}^k a_i \cdot \frac{f_i}{a_i} = \sum_{i=1}^k f_i = 1$$

El **polígono de frecuencias** se obtiene uniendo el punto medio de la base superior de cada rectángulo (corresponde a la marca de clase del su intervalo asociado) con el punto medio de la base superior del rectángulo que le sigue.

2) Diagrama integral: Polígono acumulativo de frecuencias

En el extremo superior de cada intervalo levantamos una ordenada igual a la frecuencia absoluta (relativa) acumulada correspondiente, uniendo después dichas ordenadas. La primera ordenada se une al extremo inferior del primer intervalo, prolongando la poligonal desde ese punto hacia la izquierda sobre el eje de abscisas. Desde la ordenada del extremo superior del último intervalo, que será "N" ó "1" según que la frecuencia acumulada sea absoluta o relativa, trazamos paralela al eje de abscisas.

1.10 GRAFICAS PARA ATRIBUTOS

1) Diagrama de sectores

A cada modalidad del atributo se le asocia un sector circular de área proporcional a la frecuencia de la modalidad. Así, si n_i es la frecuencia absoluta de la i -ésima modalidad, el ángulo correspondiente es $\alpha_i = 360 \cdot n_i / N$.

2) Diagrama de rectángulos

A cada modalidad del atributo se le asocia un rectángulo, de modo que la base es igual para todos y la altura es proporcional a la frecuencia absoluta (relativa) correspondiente.

3) Pictograma

Son gráficos figurativos donde los fenómenos estudiados se representan por los objetos relacionados con el carácter estudiado.

©netkeynes.com