Distribuciones estadísticas unidimensionales.

Variables estadísticas

<u>Carácter</u>

Se denomina carácter a cada uno de los atributos observables en los elementos que constituyen la población.

<u>Modalidades</u>

Las modalidades son las diferentes situaciones que puede presentar un carácter. Para que las técnicas estadísticas sean aplicables, se requiere que las modalidades sean exhaustivas y excluyentes, es decir que cada elemento de la población presente una y solamente una de las modalidades del carácter.

<u>Variable estadística</u>

Cualquiera que sea la escala de medida que se utilice para cuantificar un carácter, a cada modalidad le corresponderá un número, ya que dos elementos que presenten la misma modalidad serán considerados equivalentes en el sistema relacional empírico, y por tanto sus medidas serán iguales.

Este número que es fijo para cada modalidad, pero que varía de una modalidad a otra, se llama variable estadística. Como se ve, las características de los valores que puede tomar una variable estadística, dependen del esquema de medición que se utilice, de esta forma tenemos los siguientes tipos de variables:

<u>Variables nominales</u>

Son aquellas en que el carácter ha sido medido en una escala nominal. Como se vio, la única relación que puede establecerse entre sus valores es la de igualdad, actuando los números como meros identificadores, por ello no se pierde ninguna propiedad si las distintas modalidades son descritas por epígrafes no numéricos, lo que es utilizado habitualmente por razones de claridad. Denominándose también como variables cualitativas o caracteres cualitativos.

<u>Variables ordinales</u>

Son aquellas en que el carácter ha sido medido en una escala ordinal.

<u>Variables cuantitativas</u>

Son aquellas en que el carácter se ha medido, al menos, en una escala de intervalos. También se denominan caracteres cuantitativos, porque son los caracteres con cuyas medidas pueden realizarse las operaciones numéricas habituales. No obstante, desde el punto de vista de los métodos estadísticos aplicables a estas variables, conviene distinguir según que el número de los valores posibles sea numerable o no. Así distinguiremos entre variables discretas y continuas.

Variables estadísticas discretas

Son aquellas en que sus posibles valores constituyen un conjunto numerable. Es decir, existe un primer valor, un segundo valor, etc., esto implica que dados dos valores consecutivos no puede existir un valor intermedio.Por consiguiente, sus posibles valores son números aislados, aunque no tienen porqué ser números enteros. Por ejemplo, el número de aciertos, o errores, en una tarea de discriminación.

Variables estadísticas continuas

Son aquellas en que sus posibles valores son cualesquiera de los números reales comprendidos en un intervalo. Esto implica que el conjunto de sus posibles valores no es numerable y que dados dos valores cualesquiera de la variable, existe siempre uno intermedio. Ejemplo de este tipo de variables es el tiempo de reacción frente a un estímulo luminoso.

Aunque la distinción teórica entre estos dos tipos de variables es clara, en la práctica la clasificación de una variable en un tipo u otro puede, a veces, ser arbitraria. En realidad ninguna medida real es continua, pues la precisión limitada de los instrumentos de medida hace que el conjunto de valores posibles sea finito, aunque estemos midiendo un carácter esencialmente continuo.

Por ejemplo, si estamos midiendo el tiempo de reacción con un cronómetro digital que aprecia centésimas de segundo, podremos obtener para un sujeto 56 o 57 centésimas, pero nunca valores intermedios y el conjunto de nuestros posibles valores será discreto.

Por otra parte, aunque un carácter sea discreto, como en el caso del número de aciertos en un test, si el número de posibles valores de esa variable es muy alto, lo que ocurriría si el mencionado test constara, por ejemplo, de 500 ítems, entonces aunque la variable no deja de ser discreta, sería más útil tratarla con los métodos aplicables a las variables continuas, como veremos posteriormente.

Distribuciones estadísticas

Una vez que hemos medido un carácter en una determinada muestra o población, tendremos un conjunto de valores, como quiera que la estadística se aplica en general a conjuntos numerosos, el número de datos será lo suficientemente elevado como para impedir que la simple inspección, nos permita abarcar toda la información que contienen. El primer paso será ordenar y estructurar estos datos de una forma que mejore su comprensión y facilite su elaboración posterior. Esta ordenación se denomina distribución de frecuencias y su presentación se realiza mediante las tablas estadísticas.

Distribuciones estadísticas de un carácter

Las distribuciones estadísticas más sencillas son las que se limitan a reflejar los valores de un solo carácter.

Consideremos una población, o una muestra, que consta de n individuos. Sea A un atributo, observable en los individuos, que presenta k modalidades que designamos:

$$A_1, A_2, \ \dots \ A_i, \ \dots \ A_k$$

Una vez observado el carácter A en la totalidad de los individuos, tendremos n_1 individuos que presentan la modalidad A_1 , n_2 que presentan la modalidad A_2 , y en general n_i que presentan la modalidad A_i .

A n_i se le llama frecuencia absoluta de la modalidad A_i .

Es evidente que por exigirle a las modalidades que sean exhaustivas y excluyentes, se verificará que:

$$n_1 + n_2 + \ldots + n_i + \ldots + n_k \quad = \quad n$$

o en notación más compacta:

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

Si ahora consideramos del total de individuos de la población que proporción de ellos pertenecen a la modalidad A_i , tendremos:

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

llamaremos a f_i frecuencia relativa de la modalidad A_i .

Es evidente que este cociente de la frecuencia absoluta dividida por el total, que define la frecuencia relativa, tiene que ser menor o igual que la unidad, ya que la frecuencia con que se presenta una modalidad, será menor, o en todo caso, igual que el número total de individuos de la población. Por consiguiente:

$$0 \quad \leq \quad f_i \quad \leq \quad 1$$

Es sencillo comprobar que las frecuencias relativas gozan de la siguiente propiedad:

$$f_1 + f_2 + \ldots + f_i + \ldots + f_k \quad = \quad \sum_{i=1}^{k} f_i \quad = \quad 1$$

En efecto

$$\sum_{i=1}^{k} f_i \quad = \quad \sum_{i=1}^{k} \frac{n_i}{n} \quad = \quad \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_i}{n} + \ldots + \frac{n_k}{n} \quad =$$

$$= \frac{n_1 + n_2 + \ldots + n_i + \ldots + n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i = \frac{n}{n} = 1$$

Tablas estadísticas

La presentación ordenada de los datos se realiza en una tabla, denominada tabla estadística. La tabla estadística que describe la población, según el carácter A , será de forma general:

Identificadores de las modalidades	Frecuencias absolutas de las modalidades	Frecuencias relativas de las modalidades
A_1	n_1	f_1
...	...	...
A_i	n_i	f_i
...	...	...
A_k	n_k	f_k
Totales	n	1

La forma general de la tabla presentará pequeñas modificaciones y añadidos en función del tipo de variable estadística que se represente en ella. A continuación pasamos revista a estas peculiaridades.

<u>Variables nominales</u>

Como ya indicamos, en estas variables los distintos valores numéricos son meras etiquetas, por lo que generalmente se emplean como identificadores de las modalidades epígrafes no numéricos que proporcionan mayor información sobre las características de los elementos incluidos en esa modalidad.

<u>Variables ordinales</u>

Cuando el carácter ha sido medido en una escala ordinal, en general, se coloca en la columna de identificadores de las modalidades, los k posibles valores de la variable, correspondientes a las k modalidades A_1, A_2, ... A_k del carácter A. No obstante, a veces pueden utilizarse etiquetas no numéricas que indiquen la relación de orden, como bajo, medio, alto, etc.

<u>Variables cuantitativas discretas</u>

Cuando la variable es discreta, en la columna modalidades se colocan los k valores correspondientes a cada una de las modalidades.

<u>Variables cuantitativas continuas</u>

En el caso de variables continuas, los valores posibles se agrupan en intervalos o clases. Un intervalo queda definido cuando se determinan sus extremos; inferior e_{i-1} y superior e_i . Habitualmente, los intervalos se consideran cerrados a la izquierda y abiertos a la derecha, es decir que el extremo inferior está incluido en el intervalo, pero el extremo superior no.

La diferencia entre el extremo superior e inferior de cada intervalo, se le llama amplitud del intervalo:

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

Al punto medio del intervalo se le llama marca de clase y viene determinado por la semisuma de los extremos de esa clase

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

En el caso de una variable cuantitativa continua, la tabla al incorporar esta información complementaria presentará la forma siguiente:

Intervalos	Amplitudes	Marcas de clase	Frec. absolutas	Frec. relativas
e_0 - e_1	a_1	c_1	n_1	f_1
e_1 - e_2	a_2	c_2	n_2	f_2
...	...	...	...	...
e_{i-1} - e_i	a_i	c_i	n_i	f_i
...	...	...	...	...
e_{k-1} - e_k	a_k	c_k	n_k	f_k
Totales			n	1

Frecuencias acumuladas

Además de las frecuencias absolutas y relativas, en las tablas de variables ordinales o cuantitativas se suele añadir una, o varias, columnas más donde se reflejan las frecuencias acumuladas. Las frecuencias acumuladas pueden referirse a frecuencias absolutas o relativas y ser ascendentes o descendentes, reflejando por consiguiente el número o la proporción de individuos que presentan valores inferiores o iguales a un valor de la variable, en el caso de las ascendentes, y valores superiores en el caso de las descendentes. Si, por ejemplo, añadimos una columna de frecuencias absolutas acumuladas ascendentemente, tendríamos:

Valores de la variable	Frecuencias absolutas	Frec. absolutas acumuladas
x_1	n_1	$N_1 = n_1$
...	...	...
x_i	n_i	$N_i = n_1 + n_2 + \cdots + n_i$
...	...	...
x_k	n_k	$N_k = n_1 + n_2 + \cdots + n_k$

Registro discreto de variables continuas

Cuando la variable considerada es continua ocurre, como veíamos en el ejemplo de la medida de los tiempos de reacción con un cronómetro digital que apreciaba centésimas de segundo, que la precisión del instrumento de medida limita a un número finito y por consiguiente discreto los valores que puede tomar dicha variable. Esto implica que cada número registrado no supone la aparición de ese valor exacto, sino la existencia de un valor dentro de un determinado intervalo.

Por ejemplo, el registro de un tiempo de 56 centésimas, en realidad nos señala que el sujeto en cuestión ha reaccionado en un tiempo comprendido entre 555 y 565 milésimas.

Lo anterior no tiene ninguna incidencia, si consideramos en nuestra tabla estadística intervalos semiabiertos, pero existe una tradición que en estos casos, construye los intervalos de tal forma que ambos extremos estén incluidos en él. Por ejemplo, si los tiempos de reacción posibles están comprendidos entre 50 y 61 centésimas, siguiendo la construcción habitual los intervalos podrían ser:

50 a 52,   53 a 55,   56 a 58 y 59 a 61

Estos serían los límites aparentes de los intervalos, pero si tomamos en consideración el carácter continuo de esta variable, los límites reales de los intervalos serían:

49,5 a 52,5 52,5 a 55,5 55,5 a 58,5 y 58,5 a 61,5

La distinción entre límites reales y aparentes tendrá importancia en aquellas operaciones estadísticas en las que intervenga la amplitud del intervalo.

Representaciones gráficas

Las tablas estadísticas presentan toda la información respecto de las modalidades y sus respectivas frecuencias, sin embargo, suele ser útil completarlas con una representación gráfica.

El objeto de las representaciones gráficas es proporcionar una información visual de los datos numéricos que permita poner en evidencia algunas características esenciales de la distribución. En especial, las gráficas permiten una rápida comparación del valor de las frecuencias de las distintas modalidades.

El principio general de estas representaciones es asignar a cada modalidad una imagen simple cuya longitud, área o volumen sea proporcional a la frecuencia correspondiente. A continuación describiremos algunas de las representaciones gráficas más habituales, indicando para que tipos de variables son adecuadas.

Diagrama de rectángulos o barras

Se construye a partir de unos ejes cartesianos. En el eje de abscisas se colocan los valores de la variable y sobre ellos se levantan barras o rectángulos de alturas iguales o proporcionales a las frecuencias correspondientes.

Este tipo de gráficos es adecuado para variables nominales, ordinales y variables cuantitativas discretas. Para estos dos últimos tipos de variables, pueden construirse también utilizando las frecuencias acumuladas.

Ejemplo: La distribución del número de intentos necesarios, hasta recitar correctamente un párrafo, en 120 escolares fue la siguiente:

Intentos	Nº sujetos
1	32
2	54
3	22
4	8
5	4

Y el diagrama de barras correspondiente sería:

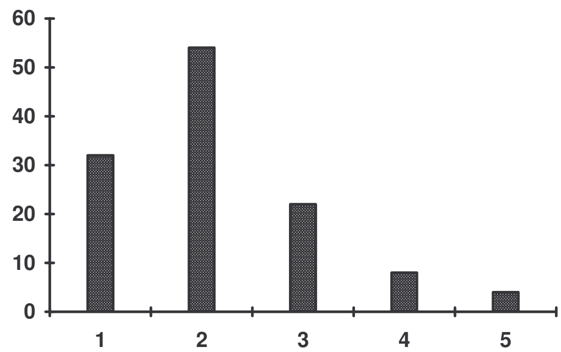


Diagrama de sectores

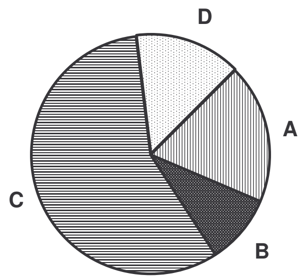
Consiste en un círculo que se divide en tantos sectores como modalidades tenga la variable, de tal forma que el área de cada sector, o lo que es equivalente su amplitud, sea proporcional a la frecuencia de la modalidad que representa.

Al igual que el anterior, este gráfico es adecuado para representar variables nominales, ordinales y cuantitativas discretas, aunque su empleo más habitual es en las variables nominales.

Ejemplo: La distribución de 84 alumnos de primero de Psicología, según la opción de Bachillerato que habían cursado, se da en la tabla siguiente:

Opción	Nº alumnos
A	16
B	8
C	48
D	12

El diagrama de sectores correspondiente a esta distribución sería:



Histograma

El histograma se utiliza exclusivamente para representar variables cuantitativas continuas. Está constituido por tantos rectángulos yuxtapuestos como clases se consideren para la variable. Estos rectángulos tienen una base igual o proporcional a la amplitud del intervalo y altura igual o proporcional a la frecuencia rectificada de la clase.

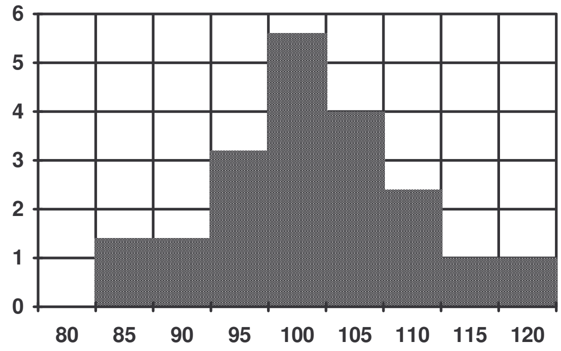
Las frecuencias rectificadas son el resultado de dividir las frecuencias por las amplitudes de las clases correspondientes. El objeto de esta operación es conseguir que las áreas de los rectángulos sean proporcionales a las frecuencias. En el caso de que todos los intervalos tengan igual amplitud, esta operación no es necesaria y pueden tomarse directamente como alturas las frecuencias de los intervalos.

El histograma puede construirse tanto a partir de las frecuencias absolutas como de las frecuencias relativas. En el primer caso la suma de las áreas de todos los rectángulos será igual al número total de individuos y en el segundo caso a la unidad.

Ejemplo: Las puntuaciones directas, obtenidas en el test de Weschler por un grupo de 100 escolares, presentan la siguiente distribución:

Puntuaciones	Nº sujetos
80 - 90	14
90 - 95	16
95 - 100	28
100 - 105	20
105 - 110	12
110 - 120	10

El histograma correspondiente a esta distribución sería:



Polígono de frecuencias

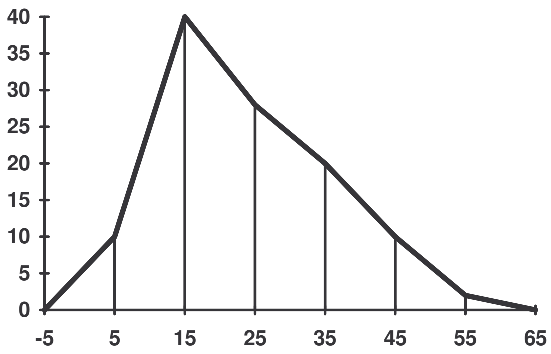
Esta forma de representación gráfica se utiliza únicamente para variables cuantitativas continuas, cuyos intervalos tengan igual amplitud, aunque existen gráficos de aspecto similar, los perfiles ortogonales que se utilizan en variables nominales, ordinales y cuantitativas discretas.

Para construir el polígono de frecuencias se procede de la siguiente forma: En el eje de abscisas se colocan las marcas de clase de los intervalos que se hayan considerado para la variable, añadiendo una más a la izquierda que correspondería al intervalo que tendría como extremo superior el extremo inferior de la primera clase, y otra a la derecha perteneciente al intervalo cuyo extremo inferior coincidiría con el superior de la última clase. En el eje de ordenadas se toman los valores correspondientes a las frecuencias de las distintas clases. Se marcan los puntos que representan a estas coordenadas y se traza la poligonal que los une.

Ejemplo: Las puntuaciones en la escala de Ansiedad Estado del cuestionario de Ansiedad Estado-Rasgo (STAI) de un grupo de pacientes de un hospital fue la siguiente:

Puntuación	Nº de pacientes
0 - 10	10
10 - 20	40
20 - 30	28
30 - 40	20
40 - 50	10
50 - 60	2

El polígono de frecuencias que obtendríamos como representación de esta distribución sería:



El área encerrada entre el polígono de frecuencias y el eje de abscisas es igual al total de individuos, si se construye utilizando las frecuencias absolutas e igual a la unidad si se emplean las frecuencias relativas.

También puede construirse el polígono de frecuencias sobre el histograma, cuando los intervalos son de la misma amplitud, uniendo los puntos medios de las bases superiores de los rectángulos y también, para tener un polígono cerrado sobre el eje de abscisas, con los centros de los intervalos adicionales descritos anteriormente.

Curva acumulativa o de distribución

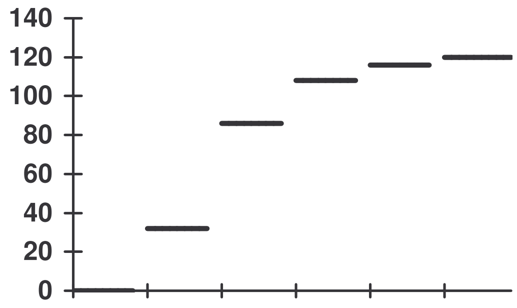
Es una curva representada en unos ejes cartesianos que para cada valor de la variable en abscisas x_i nos da en ordenadas el número de individuos, cuyo valor de la variable es inferior o igual a x_i, si la construimos a partir de las frecuencias absolutas. Si partimos de las frecuencias relativas obtendremos en ordenadas, para cada valor x_i, la proporción de individuos cuyo valor es inferior o igual a x_i.

Esta representación puede emplearse para variables ordinales y para variables cuantitativas, tanto discretas como continuas. No obstante, en el caso de variables ordinales o cuantitativas discretas al no existir valores intermedios, entre dos consecutivos, el gráfico presenta una forma característica denominada curva en escalera.

Ejemplo: Volviendo a la distribución del número de intentos necesarios, para recitar correctamente un párrafo, si calculamos las frecuencias acumuladas, tendremos la tabla siguiente:

Intentos	Nº sujetos	Frec. acumulad.
1	32	32
2	54	86
3	22	108
4	8	116
5	4	120

La curva de distribución correspondiente sería:



En el caso de variables continuas, aunque únicamente se conocen los valores de la curva para los extremos de los intervalos, se aproxima a una curva continua trazando la poligonal que une estos puntos.

Ejemplo: Si a partir de la distribución de las puntuaciones en la Escala de Ansiedad Estado, que vimos en un ejemplo anterior, calculamos las frecuencias acumuladas, tendremos la siguiente tabla:

Puntuación	Nº de pacientes	Frec. acumulad.
0 - 10	10	10
10 - 20	40	50
20 - 30	28	78
30 - 40	20	98
40 - 50	10	108
50 - 60	2	110

construimos la curva de distribución, obtendremos la siguiente gráfica: