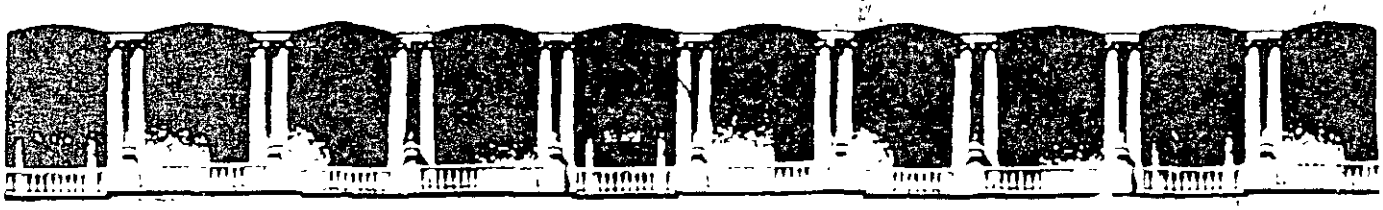


FACULTAD DE INGENIERIA U.N.A.M.
DIVISION DE EDUCACION CONTINUA

CURSOS INSTITUCIONALES

MODULO BÁSICO DE TECNOLOGÍA DE LA INFORMACIÓN Y BIOESTADÍSTICA

Bioestadística, Metodología y aplicaciones

Del 27 de Agosto al 17 de Septiembre del 2003

APUNTES GENERALES

CI - 120

Instructor: Ing. Rodolfo González Maldonado
I S S S T E
AGOSTO/SEPTIEMBRE DEL 2003

Palacio de Mineria Calle de Tacuba 5 Primer piso Deleg Cuauhtémoc 06000 México, D.F. APDO Postal M-2285
Teléfonos 5512-8955 5512-5121 5521-7335 5521-1987 Fax 5510-0573 5521-4021 AL 25

CURSO:

ESTADISTICA, METODOS Y APLICACIONES

OBJETIVO:

CONOCER COMO RECOPILAR, ORGANIZAR, ANALIZAR E INTERPRETAR LOS DATOS NUMERICOS A PARTIR DE HERRAMIENTAS ESTADISTICAS COMO: LA MEDIA, MODA, VARIANZA, GRAFICOS, ETC.

ING. RODOLFO GONZALEZ
AGOSTO 2003

INDICE

Introducción

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.

El propósito de este libro es el de dar conceptos y explicar técnicas que permitan realizar ambos procesos, a los cuales de forma conjunto se les suele denominar *Análisis de Datos*

Conceptos Previos

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

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 poseén 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. Pués 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 posilidades que éste puede presentar y además deben ser disjuntas un individiio no puede presentar a la vez má.s 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 individiios 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	*	*	. .	*

Matriz de Datos

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, en donde se recogen las observaciones de los caracteres "edad", "sexo", "cáncer", "caida anormal del cabello" y "profesión" en los 100 individuos seleccionados en la muestra.

	edad	sexo	cáncer	caida cabello	profesion
individuo 1	32	masculino	no	no	agricultor
individuo 2	29	femenino	no	no	maestra
. . .	. . .	. . .	. . .	. . .	. . .
individuo 100	61	masculino	si	si	agricultor

Estudio de Contaminación Radioactiva

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.

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 cálculode 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

Agrupamiento en intervalos

Si tenemos la opción de poder agrupar los datos en intervalos, lo primero que debemos plantearnos (independientemente de lo que más arriba comentábamos) es la cuestión de cuantos y cuales intervalos elegir.

Previamente daremos algunas definiciones. Si los intervalos o *clases*, como a veces se denominan, son:

$$[x_0 , x_1) , [x_1 , x_2) , \ldots , [x_{j-1} , x_j) , \ldots , [x_{k-1} , x_k]$$

llamaremos amplitud del intervalo j-ésimo a $x_j - x_{j-1}$, j=1.....k, hablando de intervalos de amplitud constante o variable según tengan o no todos la misma amplitud.

Llamaremos extremos de la clase j-ésima a x_{j-1} y a x_j, y por último, llamaremos centro o *marca de clase* correspondiente al intervalo *j-ésimo* al punto medio del intervalo, es decir, a $c_j = (x_j + x_{j-1})/2$.

A lo largo de la página consideraremos que el dato x_j pertenece al intervalo j+1, j=1, ... k-1 , siendo el x_k del k-ésimo. Hacemos notar también que el primer y

último intervalo generalmente tienen, respectivamente, el extremo inferior y superior indeterminados con objeto de incluir observaciones poco frecuentes.

Respecto a la cuestión que nos planteábamos al comienzo de este apartado, podemos considerar como regla general la de construir, siempre que sea posible, intervalos de amplitud constante, sugiriendo sobre el número k de intervalos a considerar el propuesto por Sturges

$$k = 1 + 3'322 \log_{10} n$$

siendo n el número total de datos.

Una vez determinado el número k de intervalos a considerar, y si es posible tomarlos de igual amplitud, esta será:

$$a = \frac{x_{(n)} - x_{(1)}}{k}$$

en donde $x_{(n)}$ es el dato mayor y $x_{(1)}$ el menor.

Ejemplo 1: "Niveles de Colinesterasa"

Se midieron los niveles de colinesterasa en un recuento de eritrocitos en *µmol/min/ml* de 34 agricultores expuestos a insecticidas agrícolas, obteniéndose los siguientes datos:

Individuo	Nivel	Individuo	Nivel	Individuo	Nivel
1	10.6	13	12.2	25	11.8
2	12.5	14	10.8	26	12.7
3	11.1	15	16.5	27	11.4
4	9.2	16	15.0	28	9.3
5	11.5	17	10.3	29	8.6
6	9.9	18	12.4	30	8.5
7	11.9	19	9.1	31	10.1
8	11.6	20	7.8	32	12.4
9	14.9	21	11.3	33	11.1
10	12.5	22	12.3	34	10.2

11	12.5	23	9.7
12	12.3	24	12.0

Niveles de Colinesterasa

Aplicando la fórmula de Sturges obtenemos:

$k = 1 + 3'322 \log_{10} 34 = 1 + 3'322 \cdot 1'53148 = 6'08757$

es decir, una sugerencia de 6 intervalos. Como el mayor valor es $x_{(34)} = 16'5$ y el menor $x_{(1)} = 7'8$, la longitud sugerida es

$$a = \frac{16'5 - 7'8}{6} = 1'45$$

Parece, por tanto, razonable tomar como amplitud 1'5, obteniendo como intervalos en los que clasificar los datos

[7'5 - 9) , [9 - 10'5) , [10'5 - 12) , [12 - 13'5) , [13'5 - 15) , [15 - 16'5]

Distribuciones Unidimensionales de Frecuencia

En este apartado consideraremos que tenemos datos correspondientes a un solo carácter, el cual, como antes dijimos llamaremos ***variable estadística*** y representaremos por X.

Llamaremos *frecuencia total* al número de datos n. Llamaremos *frecuencia absoluta n_i de la modalidad M_i* (valor x_i o intervalo I_i) de la variable X al número de datos que presentan la modalidad M_i (valor x_i o valor del intervalo I_i). Si existen ***k modalidades posibles***, se verificará

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

Llamaremos *frecuencia relativa f_i* de la modalidad M_i (valor x_i o intervalo I_i) de la variable X al cociente $f_i = n_i/n$, verificándose:

$$f_i = f_1 + f_2 + \ldots + f_k = 1$$

Llamaremos *frecuencia absoluta acumulada N_i hasta la modalidad M_i* (valor x_i o intervalo I_i) a la suma

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

Claramente es $N_k=\sum_{j=1}^{k} n_j=n$

Llamaremos *frecuencia relativa acumulada F_i hasta la modalidad M_i* (valor x_i o intervalo I_i) al cociente $F_i=N_i/N$, o lo que es lo mismo, a

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

siendo $F_k=\sum_{j=1}^{k} f_j=1$

Distribuciones unidimensionales de frecuencias

La tabla formada por las distintas modalidades (valores o intervalos) del carácter X y por las frecuencias absolutas (relativas, absolutas acumuladas o relativas acumuladas) recibe el nombre de *distribución de frecuencias absolutas (relativas, absolutas acumuladas o relativas acumuladas* respectivamente).

Tenemos, por tanto, para cada tipo de datos, cuatro distribuciones de frecuencias, obteniéndose a partir de una cualquiera de ellas las tres restantes, supuesto que se conoce la frecuencia total.

Las cuatro distribuciones de frecuencias se expresan en tablas como siguientes dependiendo del tipo de datos que sean:

1. Carácter cualitativo:

M_i	n_i	f_i	N_i	F_i
Cualidad$_1$	n_1	f_1	N_1	F_1
Cualidad$_2$	n_2	f_2	N_2	F_2

...	...	...	...	...
Cualidad$_i$	n_i	f_i	N_i	F_i
...	...	...	...	...
Cualidad$_k$	n_k	f_k	$N_k=n$	$F_k=1$
	n	**1**		

Carácter Cualitativo

2. Carácter cuantitativo sin agrupar:

X_i	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_i	n_i	f_i	N_i	F_i
...	...	...	...	...
x_k	n_k	f_k	$N_k=n$	$F_k=1$
	n	**1**		

Carácter Cuantitativo sin Agrupar

3. Carácter cuantitativo agrupado en intervalos:

I_i	n_i	f_i	N_i	F_i
I_1	n_1	f_1	N_1	F_1
I_2	n_2	f_2	N_2	F_2
...	...	...	...	...
I_i	n_i	f_i	N_i	F_i
...	...	...	...	...
I_k	n_k	f_k	$N_k=n$	$F_k=1$
	n	**1**		

Carácter Cuantitativo Agrupado en Intervalos

Ejemplo: "Tratamiento de Radiación y Cirugía"

En un estudio sobre las razones por las que no fue completado un tratamiento de radiación seguido de cirugía en pacientes de cáncer de cabeza y cuello se obtuvieron los datos dados por la siguiente distribució de frecuencias absolutas.

Causas	n_i
Rehusaron Cirugía	26
Rehusaron Radiación	3
Empeoraron por una enfermedad ajena al cáncer	10
Otras causas	1
	40

Datos

Las cuatro distribuciones de frecuencia serán.

Causas	n_i	f_i	N_i	F_i
Rehusaron Cirugía	26	0'650	26	0'650
Rehusaron Radiación	3	0'075	29	0'725
Empeoraron por una enfermedad ajena al cáncer	10	0'250	39	0'975
Otras causas	1	0'025	40	1
	40	1		

Distribución de Frecuencias

Ejemplo: "N° de Hijos"

Tras encuestar a 25 familias sobre el número de hijos que tenían, se obtuvieron los siguientes datos.

N° de hijos(X_i)	0	1	2	3	4	
N° de familias(n_i)	5	6	8	4	2	25

Datos

Las cuatro distribuciones de frecuencia serán:

X_i	n_i	f_i	N_i	F_i
0	5	0'20	5	0'20
1	6	0'24	11	0'44
2	8	0'32	19	0'76
3	4	0'16	23	0'92
4	2	0'08	25	1
	25	**1**		

Distribución de Frecuencias

Ejemplo:

Los datos del de los *Niveles de Colinesterasa*, agrupados en los intervalos allí obtenidos, proporcionan las cuatro siguientes distribuciones de frecuencias

I_i	n_i	f_i	N_i	F_i
7'5-9	3	0'088	3	0'088
9-10'5	8	0'236	11	0'324
10'5-12	10	0'294	21	0'618
12-13'5	10	0'294	31	0'912
13'5-15	1	0'029	32	0'941
15-16'5	2	0'059	34	1
	34	**1**		

Distribución de Fecuencias

Representación Gráfica de las Distribuciones Unidimensionales de Frecuencias

La representación gráfica de una distribución de frecuencias depende del tipo de datos que la constituya.

a. **Datos correspondientes a un carácter cualitativo**

La representación gráfica de este tipo de datos está basada en la proporcionalidad de las áreas a las frecuencias absolutas o relativas. Veremos dos tipos de representaciones:

1 *Diagrama de sectores*

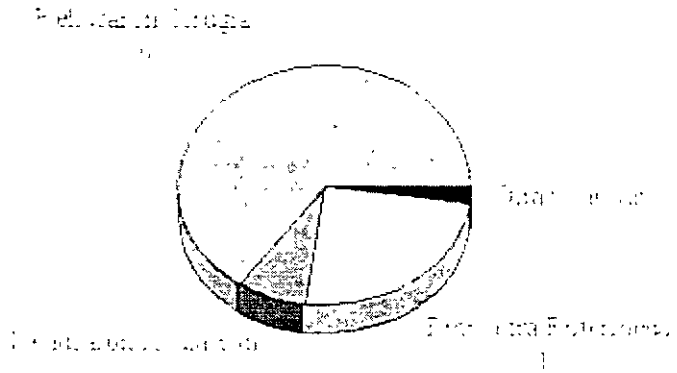
Está representación gráfica consiste en dividir un círculo en tantos sectores circulares como modalidades presente el carácter cualitativo, asignando un ángulo central a cada sector circular proporcional a la frecuencia absoluta n_i, consiguiendo de esta manera un sector con área proporcional también a n_i.

Ejemplo:

Así, los ángulos que corresponden a las cuatro modalidades de la tabla adjunta serán:

	Número de casos	Ángulo(grados)
Rehusaron cirugía	26	234°
Rehusaron radiación	3	27°
Empeoraron por una enfermedad ajena al cáncer	10	90°
Otras causas	1	9°

Y su representación en un diagrama de sectores será:

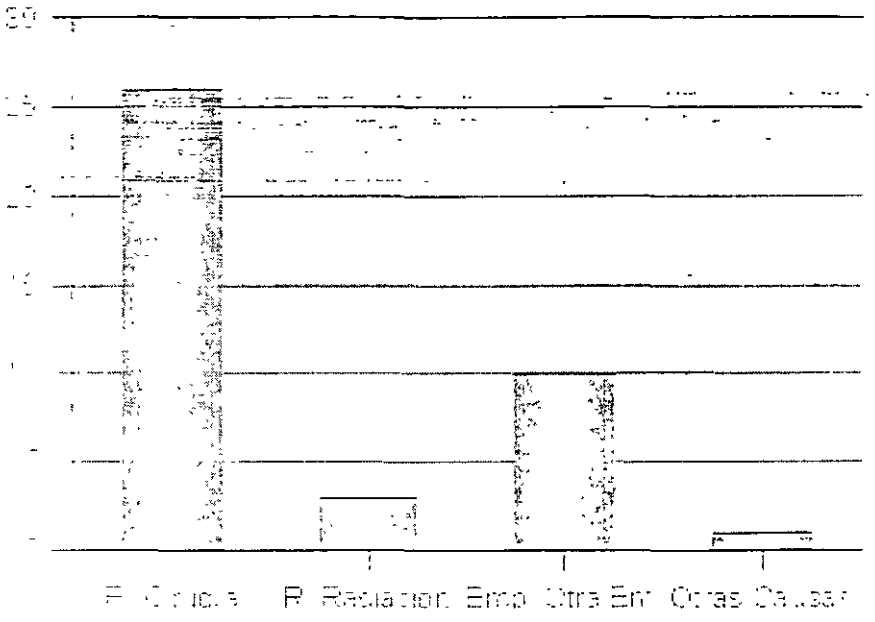


2. *Diagrama de rectángulos*

Esta representacion gráfica consiste en construir tantos rectángulos como modalidades presente el carácter cualitativo en estudio. todos ellos con base de igual amplitud. La altura se toma igual a la frecuencia absolua o relativa (según la distribución de frecuencias que estemos representando). consiguiendo de esta manera rectángulos con áreas proporcionales a las frecuencias que se quieren representar.

Ejemplo:

La representación gráfica de la distribución de frecuencias absolutas del ejemplo anterior será de la forma:



b. **Datos sin agrupar correspondientes a un carácter cuantitativo**

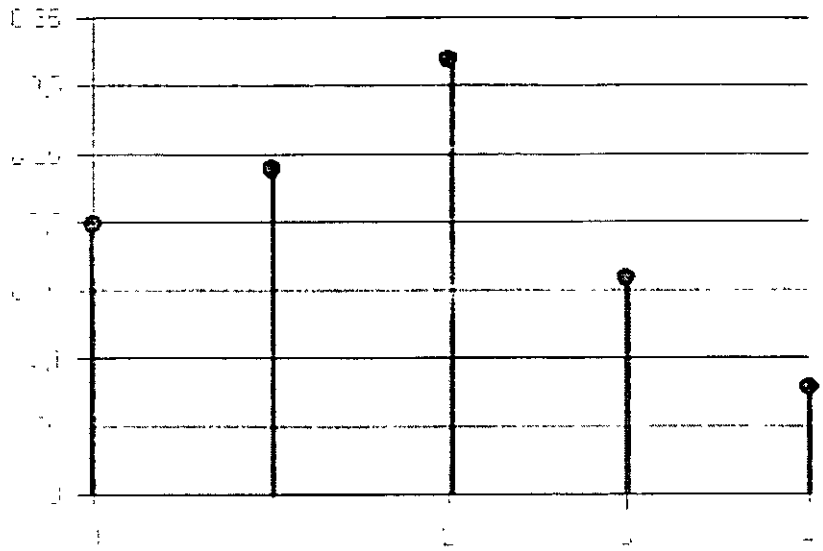
Estudiaremos dos tipos de representaciones gráficas, correspondientes a distribuciones de frecuencias (absolutas o relativas) no acumuladas y acumuladas.

1. *Diagrama de barras.*

Consiste en levantar, para cada valor de la variable, una barra cuya altura sea su frecuencia absoluta o relativa, dependiendo de la distribución de frecuencias que estemos representando.

Ejemplo:

Así, la representación gráfica de la distribución de frecuencias del *ejemplo del nº de hijos* será:

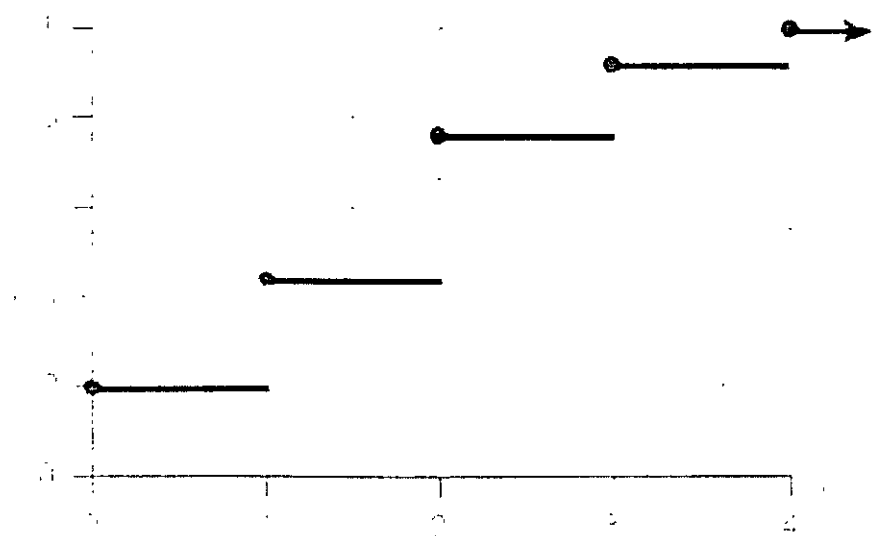


2 *Diagrama de frecuencias acumuladas.*

Esta representación gráfica se corresponde con la de una función constante entre cada dos valores de la variable a representar, e igual en cada tramo a la frecuencia relativa acumulada (o absoluta acumulada si se trata de representar una distribución de frecuencias absolutas) hasta el menor de los dos valores de la variable que construyen el tramo en el que es constante.

Ejemplo:

También para el ejemplo del *Número de Hijos*, se tendrá un *diagrama de frecuencias acumuladas* como el del siguiente gráfico:



c. **Datos agrupados en intervalos correspondientes a un carácter cuantitativo**

Al igual que antes, existen también dos tipos de representaciones gráficas dependiendo de si la distribución de frecuencias en estudio es de datos acumulados o de datos sin acumular.

1. *Histograma.*

Al ser esta representación una representación por áreas, hay que distinguir si los intervalos en los que aparecen agrupados los datos son de igualamplitud o no.

Si la amplitud de los intervalos es constante, dicha amplitud puede tomarse como unidad y al ser

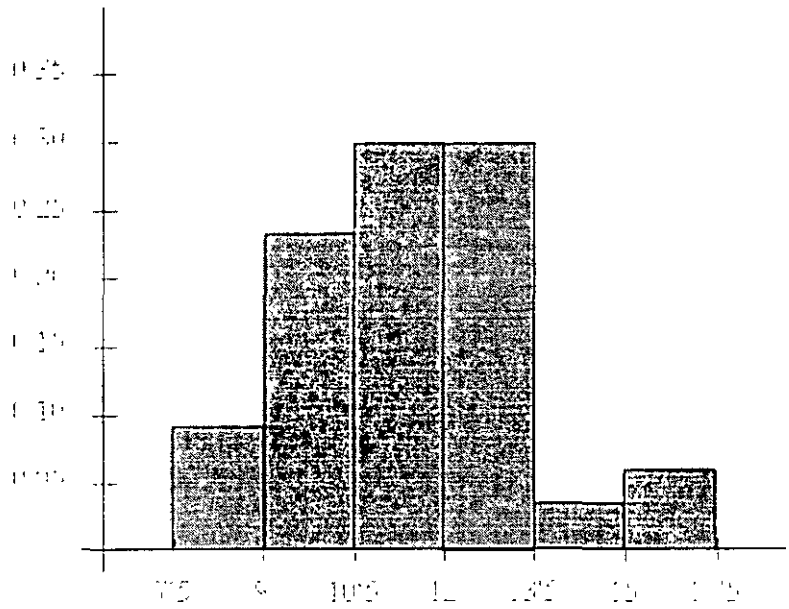
Frecuencia (área) = amplitud del intervalo · altura

la altura correspondiente a cada intervalo puede tomarse igual a la frecuencia.

Si los intervalos tienen diferente amplitud, se toma alguna de ellas como unidad (generalmente la menor) y se levantan alturas para cada intervalo de forma que la ecuación anterior se cumpla.

Ejemplo:

En el ejemplo de los *Niveles de Colinesterasa*, al tener los intervalos igual amplitud, la representación gráfica será:

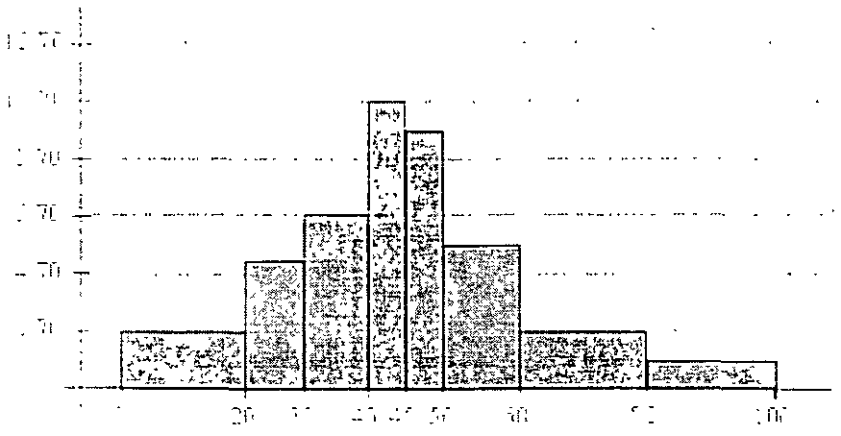


Ejemplo:

Si tuviéramos una distribución de frecuencias como la siguiente, correspondiente a puntuaciones obtenidas en un test psicológico y en la que los intervalos son de diferente amplitud

I_i	n_i	f_i
0-20	8	8/70
20-30	9	9/70
30-40	12	12/70
40-45	10	10/70
45-50	9	9/70
50-60	10	10/70
60-80	8	8/70
80-100	4	4/70
	$?n_i$= 70	$?f_i$=1

Tomando la amplitud 5 como unidad, deberemos levantar para el primer intervalo una altura de 2/70 para que el área sea la frecuencia relativa 8/70. Procediendo de la misma manera con el resto de los intervalos obtendríamos como representación gráfica la figura siguiente:



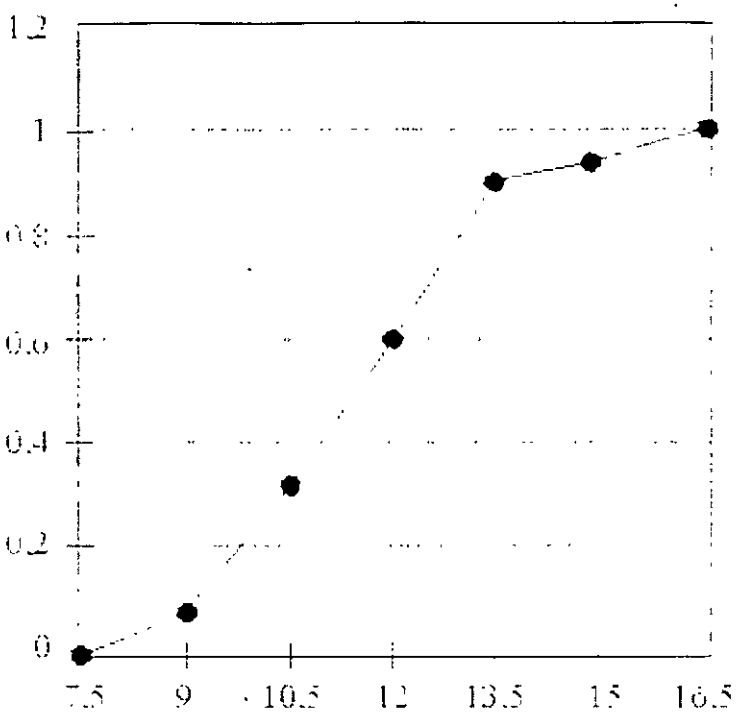
Obsérvese que la suma de todas las áreas debe ser 1, tanto si los intervalos de la distribución de frecuencias relativas son o no de igual amplitud.

2. *Polígono de frecuencias acumuladas:*

Se utiliza para representar distribuciones de frecuencias (relativas o absolutas) acumuladas. Consiste en representar la gráfica de una función que una por segmentos las alturas correspondientes a los extremos superiores de cada intervalo, tengan o no todos igual amplitud, siendo dicha altura igual a la frecuencia acumulada, dando una altura cero al extremo inferior del primer intervalo y siendo constante a partir del extremo superior del último.

Ejemplo:

Así, para el ejemplo de los *Niveles de Colinesterasa*, el *polígono de frecuencias relativas acumuladas* tendrá una representación gráfica de la forma:


1.2
1
0.8
0.6
0.4
0.2
0
7.5
9
10.5
12
13.5
15
16.5

MEDIDAS DESCRIPTIVAS

Medidas de Tendencia Central

En esta sección definiremos una serie de medidas o valores que tratan de representar o resumir a una distribución de frecuencias dada, sirviendo la cual además para realizar comparaciones entre distintas distribuciones de frecuencias. Estas medidas reciben el nombre de *promedios, medidas de posición o medidas de tendencia central*.

a. **Media aritmética**

Llamando $x_1, \ldots, x_k$ a los datos distintos de un carácter en estudio, o las marcas de clase de los intervalos en los que se han agrupado dichos datos, y $n_1, \ldots, n_k$ a las correspondientes frecuencias absolutas de dichos valores o marcas de clase, llamaremos *media aritmética* de la distribución de frecuencias a

$$\bar{a} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

en donde n es la frecuencia total.

Ejemplo 1:

La media aritmética de las veinticinco familias encuestadas será:

$$\bar{a} = \frac{\sum_{i=1}^{5} x_i \cdot n_i}{n} = \frac{0 \cdot 5 + 1 \cdot 6 + 2 \cdot 8 + 3 \cdot 4 + 4 \cdot 2}{25} = \frac{42}{25} = 1'68$$

es decir, las familias encuestadas tienen un número medio de hijos de 1'68.

Ejemplo 2:

Se midieron los niveles de colinesterasa en un recuento de eritrocitos en μ*mol/min/ml* de 34 agricultores expuestos a insecticidas agrícolas, obteniéndose los siguientes datos:

Individuo	Nivel	Individuo	Nivel	Individuo	Nivel
1	10.6	13	12.2	25	11.8
2	12.5	14	10.8	26	12.7
3	11.1	15	16.5	27	11.4
4	9.2	16	15.0	28	9.3
5	11.5	17	10.3	29	8.6
6	9.9	18	12.4	30	8.5
7	11.9	19	9.1	31	10.1
8	11.6	20	7.8	32	12.4
9	14.9	21	11.3	33	11.1
10	12.5	22	12.3	34	10.2
11	12.5	23	9.7		
12	12.3	24	12.0		

La distribución de frecuencias las marcas de clase será:

Intervalo	I_i	7'5-9	9-10'5	10'5-12	12-13'5	13'5-15	15-16'5	
Marca de Clase	x_i	8'25	9'75	11'25	12'75	14'25	15'75	
Frecuencia	n_i	3	8	10	10	1	2	$\sum n_i=25$

la cuál proporciona una media aritmética de

$$\bar{a} = \frac{\sum_{i=1}^{6} x_i \cdot n_i}{n} = \frac{388'5}{34} = 11'426$$

b. **Mediana**

La mediana es otra medida de posición, la cual se define como aquel valor de la variable tal que, supuestos ordenados los valores de ésta en orden creciente, la mitad son menores o iguales y la otra mitad mayores o iguales

Así, si en la siguiente distribución de frecuencias,

x_i	n_i	N_i
0	3	3
1	2	5
2	2	7
	7	

ordenamos los valores en orden creciente,

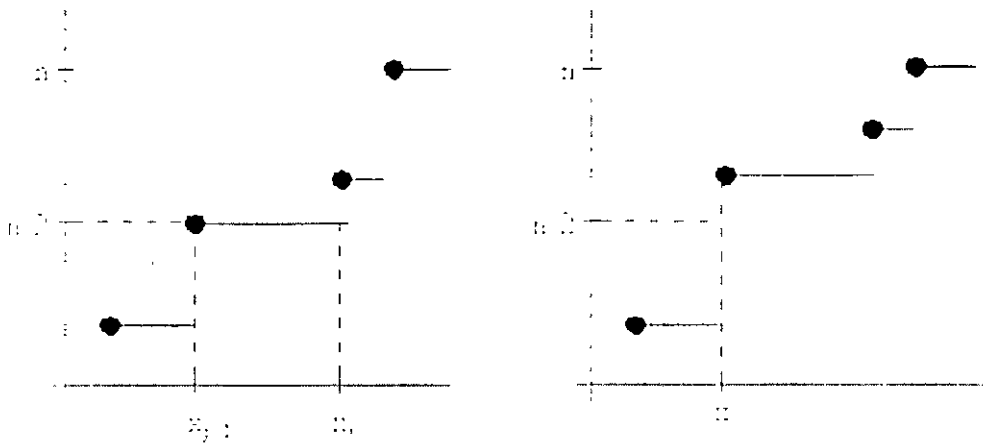
0 0 0 1 1 2 2

el 1 será el valor que cumple la definición de mediana.

Lógicamente, en cuanto el valor de la frecuencia total sea ligeramente mayor, este procedimiento resulta inviable. Por esta razón, daremos a continuación una fórmula que permita calcularla. No obstante, será necesario distinguir los casos en los que los datos vengan agrupados de aquellos en los que vengan sin agrupar.

- **Datos sin agrupar:**

Las gráficas siguientes, correspondientes a un diagrama de frecuencias absolutas acumuladas, recogen las dos situaciones que se pueden presentar:



Si la situación es como la de la figura de la derecha, es decir, si

Si la situación que se presenta es como la de la figura de la izquierda, entonces la mediana queda indeterminada, aunque en este caso se toma como mediana la media aritmética de los dos valores entre los que se produce la indeterminación: así pues, si

$N_{j-1} = n/2 < Nj$

entonces la mediana es

$$\frac{x_{j-1} + x_j}{2}$$

Ejemplo 1:

La distribución de frecuencias acumuladas del ejemplo del *número de hijos* era

Nº de hijos(x_i)	0	1	2	3	4
Frecuencias Acumuladas(N_i)	5	11	19	23	25

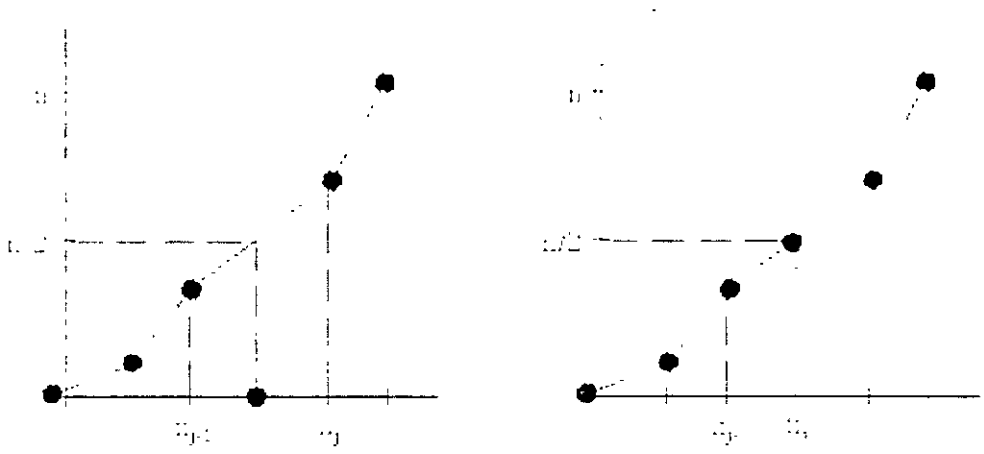
y como es n/2=12'5 y en consecuencia

11 < 12'5 < 19

la mediana será M_e= 2.

c. **Datos Agrupados**

Las gráficas siguientes, correspondientes a *polígonos de frecuencias absolutas acumuladas*, nos plantea de nuevo dos situaciones diferentes a considerar:



El más sencillo, el de la derecha, en el que existe una frecuencia absoluta acumulada N_i tal que $n/2 = N_i$, la mediana es $M_e = x_i$.

Si la situación es como la que se representa en la figura de la izquierda, en la que

$$N_{i-1} < n/2 < N_i$$

entonces, la mediana, está en el intervalo $[x_{i-1}, x_i)$, es decir entre x_{i-1} y x_i, tomándose en ese caso, por razonamientos de proporcionalidad, como mediana el valor

$$M_e = x_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot c_i$$

siendo c_i la amplitud del intervalo $[x_{i-1}, x_i)$.

Ejemplo:

La distribución de frecuencias del ejemplo de los *niveles de colinesterasa* es:

Intervalo	I_i	7'5-9	9-10'5	10'5-12	12-13'5	13'5-15	15-16'5

Frecuencia	n_i	3	8	10	10	1	2
Frecuencia Acumulada	N_i	3	11	21	31	32	34

Al ser n/2 = 17 y estar

11 < 17 < 21

la mediana estará en el intervalo [10'5 , 12). y aplicando la fórmula anterior. será

$$M_e = 10'5 + \frac{\frac{34}{2} - 11}{10} \cdot 1'5 = 11'4$$

c **Moda**

La *moda* se define como aquel valor de la variable al que corresponde máxima frecuencia (absoluta o relativa). Para calcularla. también será necesario distinguir si los datos están o no agrupados.

- Datos sin agrupar:

Para datos sin agrupar. la determinación del valor o valores (ya que puede haber más de uno) modales es muy sencilla. Basta observar a que valor le corresponde una mayor n_i. Ese será la moda.

Así en el ejemplo del número de hijos. la simple inspección de la tabla siguiente proporciona como valor para la moda el $M_d = 2$.

Nº de hijos(x_i)	0	1	2	3	4	
Nº de familias(n_i)	5	6	**8**	4	2	$?n_i=25$

- Datos agrupados:

Si los datos se presentan agrupados en intervalos es necesario, a su vez, distinguir si éstos tienen o no igual amplitud.

Si tienen amplitud constante *c*, una vez identificado el intervalo modal $[x_{i-1}, x_i)$, es decir el intervalo al que corresponde mayor frecuencia absoluta $n_i = \max\{n_1, \ldots, n_k\}$, la moda se define, también por razones geométricas, como

$$M_d = x_{i-1} + \frac{n_{i-1}}{n_{i-1} + n_{i+1}} \cdot c$$

Ejemplo:

Este ejemplo presenta un caso de distribución bimodal, ya que tanto el intervalo [10'5 - 12) como el [12 - 13'5) tienen frecuencia absoluta máxima. Deberíamos aplicar, por tanto, para cada uno de los dos intervalos la fórmula anterior, determinando así las dos modas de la distribución. No obstante, este ejemplo presenta además la peculiaridad adicional de ser ambos intervalos modales contiguos. En esta situación se considera la distribución unimodal, eligiendo como moda el extremo común, $M_d = 12$.

Si los intervalos tuvieran distinta amplitud c_i, primeros debemos normalizar las frecuencias absolutas n_i, determinando los cocientes

$$h_i = \frac{n_i}{c_i}, \quad i = 1, \ldots, k$$

y luego aplicar la regla definida para el caso de intervalos de amplitud constante a los l_i. Es decir, primero calcular el $l_j = \max\{l_1, \ldots, l_k\}$ para determinar el intervalo modal $[x_{i-1}, x_i)$ y luego aplicar la fórmula

$$M_d = x_{i-1} + \frac{l_{i-1}}{l_{i-1} + l_{i+1}} \cdot c_j$$

siendo c_j la amplitud del intervalo modal $[x_{i-1}, x_i)$.

Ejemplo:

Las frecuencias *normalizadas* correspondientes al ejemplo de intervalos con distinta amplitud serán.

I_i	n_i	l_i
0-20	8	0'4
20-30	9	0'9
30-40	12	1'2
40-45	10	2
45-50	9	1'8
50-60	10	1
60-80	8	0'4
80-100	4	0'2

con lo que el intervalo modal es el [40 - 45) y la moda

$$M_d = 40 + \frac{[illegible]}{[illegible]} \cdot 5 = 43$$

A diferencia de lo que ocurre con la media o con la mediana, sí es posible determinar la moda en el caso de datos cualitativos. Así, en el ejemplo del *tratamiento de radiación seguido de cirugía* puede afirmarse que la causa modal por la que no fue completado el tratamiento es M_d = rehusaron cirugía.

d. **Cuantiles**

Los cuantiles o cuantilas son las últimas medidas de posición que veremos. De hecho algunos autores las incluyen dentro de las medidas de dispersión al ser medidas de posición no centrales.

El cuantil $p_{r/k}$ r= 1,2,..... k - 1 se define como aquel valor de la variable que divide la distribución de frecuencias, previamente ordenada de forma creciente, en dos partes, estando el (100·r/k)% de ésta formado por valores menores que $p_{r/k}$.

Si k = 4 los (tres) cuantiles reciben el nombre de *cuartiles*. Si k = 10 los (nueve) cuantiles reciben, en este caso, el nombre de *deciles*. Por último, si k = 100 los (noventa y nueve) cuantiles reciben el nombre de *centiles*.

Obsérvese que siempre que r y k mantengan la misma proporción (r/k) obtendremos el mismo valor. Es decir, por ejemplo, el primer cuartil es igual al vigésimo quinto centil.

En este sentido, la mediana M_e es el segundo cuartil, o el quinto decil, etc.

Para el cálculo de los cuantiles de nuevo hay que considerar si los datos vienen o no agrupados en intervalos.

- Datos sin agrupar:

  Si los datos vienen sin agrupar y es

  $$N_{j-1} < \frac{r}{k} \cdot n < N_j$$

  el r-ésimo cuantil de orden k será $p_{r/k} = x_j$, valor al que corresponde la frecuencia absoluta acumulada N_j.

  Si la situación fuera de la forma

  $$N_{j-1} = \frac{r}{k} \cdot n < N_j$$

  tomaríamos, en esta situación indeterminada,

  $$p_{r/k} = \frac{x_{j-1} + x_j}{2}$$

- Datos agrupados:

  Si los datos se presentan agrupados y, para algún j, fuera

  $$\frac{r}{k} \cdot n < N_j$$

  el *résimo cuantil de orden k* sería $p_{r/k} = x_j$.

Por último, si fuera

$$N_{j-1} < \frac{\cdot}{\cdot} < N_j$$

el intervalo a considerar sería el $[x_{j-1}, x_j)$, al que corresponde frecuencia absoluta n_j y absoluta acumulada N_j, siendo entonces el cuantil el dado por la expresión.

$$\frac{\cdot}{\cdot} = x_{j-1} + \frac{\frac{\cdot}{\cdot} - N_{j-1}}{n_j} c_j$$

en donde c_j es la amplitud del intervalo $[x_{j-1}, x_j)$.

Si el intervalo a considerar fuera el $[x_0, x_1)$, se tomaría en la expresión anterior $N_{j-1} = 0$.

Ejemplo:

Vamos a determinar la tercera cuartila del ejemplo del número de hijos.

Nº de hijos(x_i)	0	1	2	3	4	
Nº de familias(n_i)	5	6	8	4	2	$?n_i=25$
Nº de familias(N_i)	5	**11**	**19**	23	25	

Como es

$$\frac{\cdot}{\cdot} \cdot 25 = \frac{\cdot}{\cdot} \cdot 25 = 18'75$$

y $11 < 18'75 < 19$, será $\mathbf{p_{3/4}=2}$.

Ejemplo:

Vamos a determinar la séptima decila del ejemplo de los *niveles de colinesterasa*.

Intervalo	I_i	7'5-	9-	10'5-	12-	13'5-	15-

		9	10'5	12	13'5	15	16'5
Frecuencia	n_i	3	8	10	10	1	2
Frecuencia Acumulada	N_i	3	11	21	31	32	34

Como es:

$$\frac{n}{2} = \frac{34}{2} = 17 < 23'8$$

$21 < 23'8 < 31$, el intervalo a considerar será el $[12 , 13'5)$, siendo

$$Me = 12 + \frac{23'8 - 21}{10} \cdot 1'5 = 12'42$$

Las *medidas de posición* estudiadas en la sección anterior servían para resumir la distribución de frecuencias en un solo valor. Las *medidas* de dispersión, a las cuales dedicaremos esta sección, tienen como propósito estudiar lo concentrada que está la distribución en torno a algún promedio.

Estudiaremos las cuatro medidas de dispersión más utilizadas: *recorrido, varianza, desviación típica y coeficiente de variación de Pearson*, estando definidas las tres primeras medidas en unidades concretas y estándolo la cuarta en unidades abstractas.

a. **Recorrido**

Si x_{max} es el dato mayor o la última marca de clase, si es que los datos vienen agrupados en intervalos, y x_{min} el dato menor o primera marca de clase, llamaremos recorrido a

$$R = x_{max} - x_{min}$$

Así, en el ejemplo de los *niveles de colinesterasa* el recorrido es

$R = 15'75 - 8'25 = 7'5$

y en el ejemplo del *número de hijos* será $R = 4 - 0 = 4$.

La principal ventaja del recorrido es la de proporcionar una medida de la dispersión de los datos fácil y rápida de calcular.

b. **Varianza**

Denotando por $x_1,\ldots,x_k$ los datos o las marcas de clase, llamaremos varianza a

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - a)^2 n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - a^2$$

siendo a la media de la distribución.

Así en el ejemplo del *número de hijos* la varianza es

$s^2 = 4'24 - (1'68)^2 = 1'4176$.

y en ejemplo de los *niveles de colinesterasa*

$s^2 = 133'97 - (11'426)^2 = 3'42.$

Al valor

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - a)^2 \cdot n_i = \frac{n \cdot s^2}{n-1}$$

se le denomina *cuasivarzanza.*

c. **Desviación Típica**

La varianza tiene un problema, y es que está expresada en unidades al cuadrado. Esto puede producir una falsa imagen de la dispersión de la distribución. En su lugar suele utilizarse su raíz cuadrada, denominada *desviación típica.* Así, la desviación típica de la distribución de frecuencias del ejemplo de los *niveles de colinesterasa* es $s = 1'1906$ y la del ejemplo del *número de hijos* $s = 1'85$.

d. **Coeficiente de Variación de Pearson**

La desviación típica sirve para medir de forma eficaz la dispersión de un conjunto de datos entorno a su media. Desgraciadamente esta medida puede resultar engañosa cuando tratamos de comparar la dispersión de dos conjuntos de datos. Así, si por ejemplo tenemos dos grupos de mujeres de 11 y 25 años con medias y desviaciones típicas dadas por la tabla siguiente:

	Peso Medio	Desviación Típica
11 años	40 Kgr.	2 Kgr.
25 años	50 Kgr.	2 Kgr.

puede parecernos, al observar en ambos grupos una desviación típica igual, que ambos grupos de datos tienen la misma dispersión. No obstante, como parece lógico, no es lo mismo una variación de dos kilos en un grupo de elefantes que en uno de conejos. El *coeficiente de Variación de Pearson* elimina esa posible confusión al ser una medida de la variación de los datos pero en relación con su media. Se defide como

$$C_p = \frac{s}{a} \cdot 100$$

siendo s y a respectivamente la desviación típica y la media de la distribución en estudio y en donde el factor 100 tiene como único objetivo el evitar operar con valores decimales.

De la defición de V_p se deduce fácilmente que aquella distribución a la que corresponda mayor coeficiente tendrá mayor dispersión.

En el ejemplo anterior, al grupo de mujeres de 11 años le corresponde un coeficiente de variación de Pearson igual a

$$V_p = \frac{\quad}{4} \cdot 100 = 5$$

y al grupo de las mujeres de 25 años

$$V_p = \frac{\quad}{5} \cdot 100 = 4$$

lo que indica una mayor dispersión en el grupo de mujeres de 11 años.

Para el ejemplo del *número de hijos* V_p toma el valor

$$V_p = \frac{\quad}{\quad} \cdot 100 = \quad$$

y en el de los *niveles de colinesterasa*

$$V_p = \frac{\quad}{\quad} \cdot 100 = \quad$$

Medidas de Asimetría

Diremos que una distribución es simétrica cuando su mediana, su moda y su media aritmética coincidan. Claramente las distribuciones de los ejemplos de los *niveles de colinesterasa* y del *n° de hijos* no son por tanto, simétricas.

Diremos que una distribución es *asimétrica a la derecha* si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda.

Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es *asimétrica a la izquierda.*

Existen varias medidas de la asimetría de una distribución de frecuencias. Aquí estudiaremos dos de ellas.

a. **Coeficiente de Asimetría de Pearson**

Se define como:

$$A_p = \frac{\bar{x} - Mo}{s}$$

siendo cero cuando la distribución es simétrica, positivo cuando existe asimetría a la derecha y negativo cuando existe asimetría a la izquierda.

En el ejemplo del *número de hijos* A_p es igual a

$$A_p = \frac{2,07}{1,39} = 1,4892$$

indicando una ligera asimetría a la izquierda en la distribución de frecuencias correspondiente.

De la misma manera, para el ejemplo de los *niveles de colinesterasa* también se observa una ligera asimetría a la izquierda, al ser

$$A_p = \frac{1,9425 - 1,2}{1,155} = 0,64$$

De la definición se observa que este coeficiente solo se podrá utilizar cuando la distribución sea unimodal. La otra medida de asimetría que veremos no presenta este inconveniente.

b. **Coeficiente de Asimetría de Fisher**

Se define como

$$A_F = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 \cdot n_i}{n \cdot S^3}$$

siendo x_i los valores de la variable o las marcas de clase y $S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot n_i}{n}}$. Llamada a veces *cuasidesviación típica.*

La interpretación del coeficiente de Fisher es la misma que la del coeficiente de Pearson, si la distribución es simétrica vale cero, siendo positivo o negativo cuando exista *asimetría a la derecha o izquierda respectivamente.*

Toda probabilidad cumple una serie de propiedades. las cuales se obtienen como consecuencia de los axiomas que debe de cumplir. A continuación vamos a.demostrar las más importantes:

1. **P( ) = 0.**

   En efecto Si consideramos la sucesión infinita

   $$A_1, \emptyset, \emptyset, \ldots$$

   es

   $$\bigcup_{i=1}^{\infty} A_i = A$$

   por lo que. por el axioma 3. deberá ser:

   $$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P(A_i)$$

   es decir.

   $$P(A)=P(A) + \sum_{i=2}^{\infty} P(A_i)$$

   de donde se deduce que $P(A_i)= P(\emptyset)$, para todo i=2..... no debe sumar nada. es decir. debe ser

   $$P(\emptyset) = 0.$$

2. **Se cumple la *aditividad* finita para sucesos incompatibles. Es decir,**

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} A_i$$

si $A_i \cap A_j = \emptyset$, $i \neq j$

En efecto: Basta considerar la sucesión

$$\{A_i\}_{i=1}^{\infty} = \{A_1, \ldots, A_n, \emptyset, \emptyset, \ldots\}$$

y aplicar de nuevo el axioma 3 y luego la propiedad anterior, quedando

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) + 0$$

es decir, la propiedad deseada.

3. **La probabilidad del complementario de un suceso A es**

$$P(A') = 1 - P(A)$$

En efecto: Aplicando primero el axioma 2 y luego la aditividad finita acabada de demostrar, será

$$P(A \cup A') = P(\Omega) = 1$$

y

$$P(A) + P(A') = 1$$

de donde se obtiene la propiedad propuesta.

4. **Si dos sucesos son tales que $A \subseteq B$, entonces $P(A) \leq P(B)$.**

En efecto: B se puede poner de la forma

$$B = A \cup (B - A)$$

con lo que, por la aditividad finita de la probabilidad, será

$$P(B) = P(A) + P(B - A)$$

La propiedad enunciada se tendrá ahora como consecuencia de ser $P(B - A) \geq 0$ por el axioma 1.

5. **La probabilidad de todo suceso A es un número entre 0 y 1:**

$$0 \leq P(A) \leq 1.$$

En efecto De hecho, el que sea mayor que cero es una de las exigencias requeridas para que sea probabilidad (axioma 1).

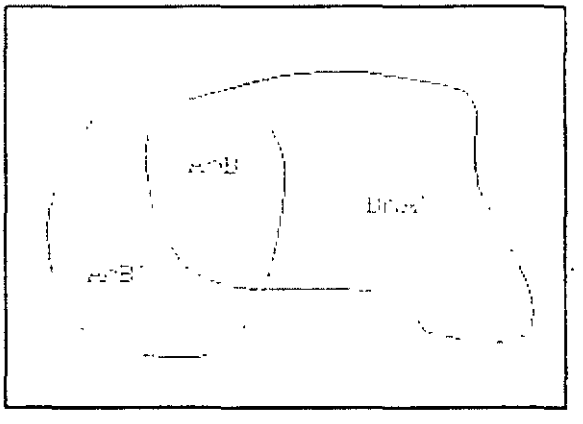
El que sea menor que 1 se obtiene de la propiedad anterior observando que todo suceso A está contenido en el suceso seguro, $A \subset \Omega$.

6. **Si dos sucesos no son incompatibles, la probabilidad de su unión debe calcularse por la siguiente regla:**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

En *efecto:* Los sucesos A y *B* se pueden escribir como unión de sucesos disjuntos de la forma.

$$A = (A \cap B) \cup (A \cap B'), \quad B = (A \cap B) \cup (A' \cap B)$$



con lo que, por la propiedad de aditividad finita antes demostrada, será

$$P(A) = P(A \cap B) + P(A \cap B') \text{ y } P(B) = P(A \cap B) + P(A' \cap B)$$

es decir,

$P(A \cap B') = P(A) - P(A \cap B)$ y $P(A' \cap B) = P(B) - P(A \cap B)$.

Como, por otro lado, A U B se puede expresar como unión disjunta de la forma

$$A \cup B = (A \cap B) \cup (A \cap B') \cup (A' \cap B)$$

su probabilidad será

$$P(A \cup B) = P(A \cap B) + P(A \cap B') + P(A' \cap B)$$

y, sustituyendo los valores antes calculados para los dos últimos sumandos, quedará

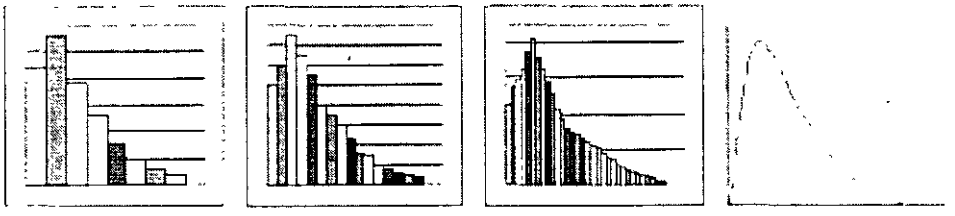
$$P(A \cup B) = P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B)$$

o en definitiva,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

LEYES DE DISTRIBUCION DE VARIABLES ALEATORIAS

Las distribuciones de probabilidad son idealizaciones de los polígonos de frecuencias. En el caso de una variable estadística continua consideramos el histograma de frecuencias relativas, y se comprueba que al aumentar el número de datos y el número de clases el histograma tiende a estabilizarse llegando a convertirse su perfil en la gráfica de una función.



Las distribuciones de probabilidad de variable continua se definen mediante una función y=f(x) llamada **función de probabilidad** o **función de densidad**.

Así como en el histograma la frecuencia viene dada por el área, en la función de densidad la probabilidad viene dada por el área bajo la curva, por lo que:

- El área encerrada bajo la totalidad de la curva es 1.
- Para obtener la probabilidad P(a□X□b) obtenemos la proporción de área que hay bajo la curva desde a hasta b.
- La probabilidad de sucesos puntuales es 0. P(X=a)=0

Función de densidad y función de distribución

Llamaremos función de densidad de una variable aleatoria continua X a una función f que cumple:

- es positiva
- el área total bajo la curva, es decir entre f(x) y el eje de abscisas, es 1

- el área determinada por f(x), el eje de abscisas y las rectas x=a, x=b, es la probabilidad de que la variable continua X esté en el intervalo [a,b], P(a≤X≤b)

Considera la función:

$$f(x) = \begin{cases} 0 & si \quad x < 0 \\ 0{,}5x & si \quad 0 \le x \le 2 \\ 0 & si \quad x > 2 \end{cases}$$

Comprueba que se trata de una función de densidad

- En efecto f(x)≥0 y el área total bajo la curva, en este caso un triángulo es 1

Si X es una variable aleatoria cuya función de densidad es f calcula

- P(X ≤ 0.75)=
- P(X ≤ 1,25)=
- P(0,75≤X ≤ 1.25)=

Cambia el valor de a y b y calcula el área

Como has visto en el ejercicio anterior la p(a≤X≤ b) viene dada por el área entre la curva y=f(x) y el eje de abscisas desde a hasta b, si has estudiado ya cálculo integral sabrás que este área es:

$$p(a \le X \le b) = \text{"área bajo la curva desde a hasta b"} = \int_a^b f(x)dx$$

Dada una variable aleatoria X la función que asigna a cada número real la probabilidad p(X≤ x) se llama función de distribución. Viene dada por

$$F(x) = p(X \le x) = \int_{-\infty}^{x} f(t)dt$$

La función de distribución correspondiente a la función de densidad anterior es

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ 0{,}25x^2 & si \quad 0 \le x \le 2 \\ 0 & si \quad x > 2 \end{cases}$$

- Observa que el área bajo la curva de 0 a x es la del triángulo de base x y altura 0,5x.

Calcula en la gráfica y compara los resultados con los obtenidos anteriormente

- $P(X \le 0{,}75)=F(0{,}75)=$
- $P(X \le 1{,}25)=F(1{,}25)=$
- $P(0{,}75 \le X \le 1{,}25)=F(1{,}25)-F(0{,}75)=$

Parámetros en una distribución de probabilidad

Por analogía con las variables estadísticas podemos definir también aquí la media μ y la desviación típica σ de la variable aleatoria.

- La media μ, también llamada esperanza matemática, es un valor representativo de todos los valores que toma la variable aleatoria X, lo podemos imaginar como el punto sobre el eje de abscisas donde al poner una cuña la figura plana definida por la función de densidad quedará en equilibrio. Para calcularla hemos de hacer:

$$\mu = \int_{-\infty}^{\infty} x f(x)dx$$

- La desviación típica σ es una medida de la dispersión de los valores que toma la variable aleatoria de la media. Como ocurría con las variables estadísticas la desviación típica será más pequeña o más grande según la gráfica de la función de densidad sea más estrecha o más ancha en torno a la media. En este caso se calcula:

$$\sigma = \sqrt{\int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2}$$

Comprueba, si sabes integrar, que en ejemplo anterior la media es 4/3 y la desviación típica 0,47

Considera la función de la escena:

$$f(x) = \begin{cases} 0 & si \quad x < 1 \\ x - 1 & si \quad 1 \leq x \leq 2 \\ -x + 3 & si \quad 2 < x \leq 3 \\ 0 & si \quad x > 3 \end{cases}$$

Comprueba que se trata de una función de densidad

Calcula, mediante las áreas de las figuras correspondientes:

- $P(X \leq 1,5)$=
- $P(1,75 \leq X \leq 2,25)$=

¿Por simetría, cuál crees que es la media en este caso?

EJEMPLOS DE DISTRIBUCIONES DE VARIABLE CONTINUA

La distribución uniforme

Las distribuciones uniformes corresponden al experimento de elegir dos puntos al azar entre dos fijos **m** y **n**. Como la probabilidad de elegir cualquier punto es la misma, la función de densidad tendrá la misma altura en todos los puntos entre **m** y **n**, es decir se trata de una función constante desde **m** a **n**, de altura 1/(m-n).

Supón que tenemos una cuerda de 2 m de longitud que queremos cortar por

un punto al azar a distancia x de uno de los extremos.

¿Cuál es la función de densidad?

- Se trata de elegir un punto al azar entre 0 y 2, como el área debe ser 1, la altura del rectángulo será 1/2

Calcula:

- $P(X \leq 0,5)$
- $P(0,5 \leq X \leq 1,25)$

Cambia el valor de a y b y calcula el área ahora se reduce a la de un rectángulo, fíjate que b debe ser mayor que a

Si la cuerda mide 3 m ¿cuál sería ahora la función de densidad?, ¿y la probabilidad de cortar la cuerda de forma que uno de los trozos mida como máximo 1 m? *(Utiliza la escena anterior cambiando m y n)*

- En esta distribuciones la media coincide con el punto medio del segmento [a,b], $\mu = \frac{a + b}{2}$
- La desviación típica es $\sigma = \frac{b - a}{\sqrt{12}}$

Sea X el momento elegido al azar en que una persona llega a una cita entre la 1 y las 2 de la tarde.

¿Cuál es en este caso la función de densidad?

¿Cuál es el valor medio esperado?, ¿con qué desviación típica?

Calcula la probabilidad de que llegue en la primera media hora $P(X \leq 1,5)$

Calcula la probabilidad de que aparezca en los últimos 15 minutos $P(1,75 \leq X \leq 2)$

La distribución exponencial

Las distribuciones exponenciales se utilizan como modelo para representar tiempos de funcionamiento o tiempos de espera. Su función de densidad que depende de un parámetro k es de la forma $f(x)=ke^{-kx}$

- La media de esta distribución es 1/k y la desviación típica también es 1/k

La variable X representa el tiempo en horas que una persona tarda en realizar determinado trabajo y sigue una distribución exponencial con parámetro 2

¿Cuál es el tiempo medio en que se espera realice dicho trabajo?

¿Cuál es la probabilidad de que lo realice en menos de 30 minutos?, ¿y en más de 1 hora?

En este caso para calcular el área hay que recurrir a calcular la integral, pero puedes ver el resultado dando los valores adecuados a a, b

El tiempo, en meses, de espera en determinado servicio sanitario se distribuye exponencialmente con media 2; ¿cuál es la función de densidad?, ¿y la probabilidad de que un paciente espere menos de 1 mes, ¿y entre 2 y 4 meses? *(Utiliza la escena anterior cambiando k)*

Introducción a la inferencia estadística

La construcción de modelos probabilísticos presentada en el capítulo anterior es el caso típico de razonamiento deductivo: se establecen hipótesis respecto al mecanismo generador de los datos y con ellas se deducen las probabilidades de los valores posibles. La Inferencia Estadística realiza el proceso inverso: dadas las frecuencias observadas de una variable, inferir el modelo probabilístico que ha generado los datos. Para ello debemos calcular los parámetros que definen las distintas distribuciones, pero esto requiere conocer los valores de la variable que estemos estudiando para todos y cada uno de los elementos de la población (conjunto de homogéneo de elementos en los que se estudia una variable dada), lo cual no es posible por varias razones:

Imposibilidad física de acceder a toda la población, por ejemplo para calcular la probabilidad de cara de una moneda requiere su lanzamiento infinitas veces.

Imposibilidad económica de acceder a toda la población, p. e. no se podrían pagar los análisis para determinar el nivel medio de colesterol en un país.

Imposibilidad por destrucción del individuo, p. e. el estudio de la duración media de un modelo de marcapasos implicaría esperar la destrucción de toda la producción.

Sea cual sea el caso, con poblaciones de un tamaño N suficientemente grande la única alternativa factible es su determinación aproximada a través de una muestra (subconjunto representativo de la población).

La Inferencia Estadística es el conjunto de métodos que permiten obtener una conclusión a cerca de una población a través de la información proporcionada por una muestra, un procedimiento inductivo que va de lo particular (muestra) a lo general (población). Cuando la información deseada de la población es el valor de alguno de sus parámetros, la técnica a utilizar es la estimación.

La estimación puede ser de dos tipos. Mediante *estimación puntual* se persigue dar un único valor aproximado del parámetro desconocido, quedando sin especificar cómo de buena es tal aproximación. Mediante la *estimación por intervalo* se persigue dar un intervalo de valores, alguno de los cuales es el verdadero valor del parámetro desconocido, con una cierta seguridad de que la afirmación sea cierta. En el primer caso se afirmaría " la proporción de varones en España es aproximadamente el 49%", en el segundo, "la proporción de varones en España es algún número entre el 48% y el 50% caso con seguridad". El valor 49% se dice que es una estimación puntual de p(la verdadera proporción de varones en España); el intervalo (48%-50%) se dice que es un *intervalo de confianza* para p.

Muestreo aleatorio

Ya que el conocimiento de la población lo va a proporcionar la muestra, es lógico que la misma no se deba tomar de un modo arbitrario, sino que debe representar adecuadamente a toda la población. Si la muestra no es representativa, nada de lo que se concluya a partir de ella será válido para la población de interés, sino que lo será para la subpoblación que representa. Así, para determinar el nivel medio de colesterol de todos los españoles, la muestra no puede tomarse sólo de personas de edad avanzada, ni sólo de individuos que aparezcan en la guía telefónica, ni sólo de individuos que acuden a un hospital, etc. Para que la muestra sea representativa de la población, es preciso que sea extraída de ella de modo que:

1º Todos los individuos de la población tengan la misma probabilidad de ser seleccionados e incluidos en la muestra (*igual probabilidad*)

2º La selección de un individuo no influya para nada en la selección o no de otro individuo cualquiera (*independencia*).

Cuando ello se verifica diremos que la muestra es una *muestra aleatoria*. La obtención de una muestra aleatoria requiere en primer lugar la identificación completa de la población en estudio; a continuación se numeran los individuos de la población y, por medios similares a un sorteo, se extrae al azar un conjunto de números, los individuos correspondientes a ellos forman una muestra aleatoria de tal población. Para hacer esta selección podemos utilizar también las tablas de números aleatorios.

ESTIMACIONES

Estimación puntual

Supongamos que se desea conocer la estatura media μ de todos los españoles. Si tomamos una muestra de $n = 100$ españoles ¿qué valor elegiremos como el más aproximado, presuntamente, a μ? Parece razonable que si 170 cm es la estatura media de dicha muestra, debemos afirmar que $\mu=170$ es inexacto (pues la media muestral no coincide en general con μ), convengamos en indicar lo anterior así: $\mu_1 = 170$, indicando el subíndice en el parámetro que la cantidad es una estimación puntual del mismo. De un modo general, una estimación puntual es un valor que se propone para el parámetro desconocido, valor que se obtiene determinando en la muestra el parámetro muestral paralelo al poblacional. Así, una estimación puntual para la media μ de una v.a es la media muestral $\mu_1 = \overline{x}$, para la varianza

σ^2 de una v.a. es la varianza muestral $q^2=s^2$ ó para la proporción de una Binomial p es la proporción muestral p_1.

Estimación por intervalo de confianza

Los estimadores puntuales sólo dan una idea aproximada del valor del parámetro a estimar, no conociéndose cómo de buena es la aproximación; ellos simplemente proporcionan el mejor número que pueda proponerse como valor del parámetro. Por ejemplo decir que $\mu_1=170$ cm significa que la estatura media de todos los españoles es aproximadamente 170 cm, pero el término "aproximado" no se sabe si alude a 1 cm arriba o abajo, o a 1 metro arriba o abajo. De hecho no puede esperarse gran cosa de un estimador.

Los problemas anteriores eran de esperar pues realmente es demasiado pedir que a partir de una muestra pueda calcularse el valor del parámetro tan exactamente como si se tomara toda la población. En realidad lo que importa es que el valor de la media muestral $\bar{x}$,por ejemplo, no esté demasiado alejado de μ, y esto se comprueba con los intervalos de confianza.

El objetivo es realizar afirmaciones del tipo: "la estatura media (de los españoles no sé exactamente cuanto es, pero es casi seguro alguno de los valores $169 \leq \mu \leq 172$, con una cierta seguridad. La seguridad alude a la probabilidad de que la afirmación sea cierta, con lo que el problema de obtener intervalos de confianza para un parámetro ϖ radica en encontrar dos valores a y b tales que $P(a \leq \varpi \leq b)=1-\alpha$,donde (a , b) es el intervalo de confianza para ϖ , 1 - α el nivel de confianza del intervalo (usualmente próximo a 1) y α el *nivel de error* del intervalo (usualmente próximo a 0).

Intervalo de Confianza para una media

Variables Normales.

Supongamos una v. a. x con distribución $N(\mu : \sigma)$ en donde la media μ es desconocida y la varianza σ^2, la suponemos por ahora conocida. Con el fin de estimar μ (colesterol medio, nivel medio de glucosa, altura media de los varones mayores de edad, etc.) se va a tomar una muestra aleatoria $x_1, x_2, \ldots, x_n$ que proporciona una media que será una estimación puntual de μ. Aceptaremos sin demostrarlo que:

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}} \quad \textbf{(4.1)}$$

con probabilidad del 95%, y así tenemos el intervalo buscado. Esta expresión debe interpretarse adecuadamente. Ella indica que el 95% de las muestras de tamaño n tendrán una media que, al sustituirla en la expresión, da lugar a un intervalo que contiene en su

interior a μ, en tanto que otro 5% no sucederá esto. Nótese que se ha dicho que "el intervalo contiene en su interior a μ, y no que "μ cae en el interior del intervalo": la primera afirmación es cierta pues los extremos del intervalo son v. a. por depender de $\overline{x}$ que también lo es, la segunda afirmación es falsa pues μ es un parámetro (valor fijo aunque desconocido), no una v.a., no pudiendo variar. Así pues debe decirse que hay una probabilidad del 95% de que el intervalo contenga al parámetro.

En el ejemplo de la estatura media μ de los españoles, si se tiene que $169 \leq \mu \leq 172$, dado que el 95% de los intervalos contienen a μ, diremos que "tenemos la esperanza de que este sea uno de los 95 intervalos de cada 100 que dejan en su interior a μ, esperando no haber tenido la mala suerte de que el intervalo obtenido sea uno de los 5 de cada 100 intervalos erróneos". Más abreviadamente, diremos que μ está entre (169 ; 172) "con una confianza del 95%": de ahí el nombre de intervalo de confianza. Conviene notar que ahora se habla de "confianza" , y no de "probabilidad" como antes, pues los extremos del intervalo ya son números fijos y μ o está o no está dentro.

El intervalo (4.1) podemos expresarlo abreviadamente como $\mu \in \overline{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$, debiéndose el valor 1.96 al 5% de error tomado, es decir $z_{\alpha/2} = 1.96$ en la tabla de la Distribución Normal.. De un modo general, si en lugar de una confianza del 95% tomamos una de (1 - α), (o en lugar de un error del 5% se toma uno de α), entonces el intervalo será:

$$\mu \in \overline{x} \pm z_{\alpha} \frac{\sigma}{\sqrt{n}} \quad \textbf{(4.2)}$$

con z_{α} en la tabla de la D. N..

> Ejemplo 1: Para determinar la estatura media de los varones adultos españoles, se tomó una muestra al azar de 10 de ellos en la que se obtuvo los valores 162, 176, 169, 165, 171, 169, 172, 168, 167 y 175 cm. Determinar el valor de la estatura media, suponiendo que $\sigma^2 = 16$.
>
> Un estimador puntual para la estatura media μ es la $\overline{x}$ que en este caso es 169,4. Para dar un intervalo de confianza hemos de suponer que es una v. a. normal. Como n=10, $\overline{x}$ = 169,4 y σ = 4, para el intervalo de confianza al 95%, la expresión (4.1) indica que $\mu \in 169{,}4 \pm 1{,}96\frac{4}{\sqrt{10}} = 169{,}4 \pm 2{,}48 = (166{,}92\ ,\ 171{,}88)$.
>
> Así pues, esperamos que este intervalo sea un de los 95 de cada 100 que contienen a μ, o, más brevemente, la estatura media de los españoles varones adultos es algún valor entre 166,92 cm y 171,88 cm con una confianza del 95%.

Es evidente que un intervalo de confianza para un α dado será tanto más preciso cuanto más estrecho sea. Así, será preferible afirmar que la estatura media está entre 170 y 171 cm al 95% de confianza, que afirmar que la estatura está entre 165 y 175 con igual confianza. Como la longitud del intervalo es dos veces su radio, el mismo puede disminuirse aumentando el valor del tamaño de la muestra (pues n aparece dividiendo). Ello responde a una regla que será general en toda la Estadística: cuanto más grande sea una muestra, más información da y más precisas son las conclusiones que se obtengan a partir de ella.

La otra forma de estrechar el intervalo es disminuyendo la confianza (es decir, aumentando el error). Así $z_{.05} = 1.96$, pero $z_{.15} = 1.44$, que por ser menor da un intervalo más estrecho. Sin embargo ahora la anchura del intervalo ha disminuido a costa de la seguridad (confianza) del mismo, y ello no es deseable. Lo usual es considerar errores α del 5%, aunque en ocasiones se utilizan otros como los del 1% o del 10%. Nos podemos preguntar ¿se puede dar un intervalo al 100% de confianza?: la respuesta es que esto exigiría una $z_{\infty} = \infty$, con lo que el intervalo sería $(-\infty, \infty)$ que en el caso del ejemplo daría lugar a la afirmación "la estatura media de los españoles está entre $-\infty$ y ∞", que es absolutamente cierta y absolutamente inútil también.

Hasta este momento hemos supuesto que la varianza de la población era conocida, lo que no suele ser real. Cuando σ es desconocida, lo lógico es sustituirla por su estimador s, obteniendo así que $\mu \in \bar{x} \pm z_{\alpha} \frac{s}{\sqrt{n}}$. Sin embargo s es una v. a. y unas veces será más grande que σ y otras más pequeña, lo que da una cierta imprecisión al intervalo. Conviene ensanchar un poco el intervalo para que la confianza del mismo permanezca. El modo de hacerlo consiste en aumentar el valor de z_{α}, localizándolo en una tabla distinta. Ahora tendremos:

$$\mu \in \bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n}} \quad \textbf{(4.3)}$$

con t_{α} en la tabla de la distribución t de Student con (n-1) grados de libertad, tabla que presenta los valores de t_{α} en un formato similar al de la distribución normal, excepto en que la nueva variable depende de un nuevo parámetro llamado grados de libertad.

> Ejemplo 2: Resolver el ejemplo anterior sin suponer conocido el valor de σ^2.
>
> De antes se conoce que n =10 y = 169.4. Ahora es preciso calcular la varianza muestral por la fórmula correspondiente lo que da s = 4.3. Como $t_{.05}(9\ g.l.) = 2.262$ en la tabla , entonces es el intervalo de confianza para μ al 95% de confianza.

La interpretación del nuevo intervalo es idéntica del que resultaba cuando la varianza era conocida, la única diferencia es que ahora no sólo el centro del intervalo es variable, sino que también lo es su radio.

Tamaño de la muestra.

En la fase de diseño de una experiencia suele plantearse cuál debe ser el tamaño mínimo de la muestra para lograr una precisión dada en la estimación de la media. Así, ¿cuántos españoles debo tomar para determinar su estatura media con una precisión de 1 cm? Con ello se quiere indicar que si concluyo que debo tomar n = 100 españoles y tomo una muestra de 100 de ellos, la estatura media en la muestra ($\overline{x}$) distará de la media de la población (μ) en menos de 1 cm (en general d cm), es decir que $|\overline{x} - \mu| \leq d$ con una cierta confianza. Otro modo de decir lo mismo es afirmar que si es $\overline{x}$ =170 en la muestra de 100 que se ha decidido como idónea, entonces sé que μ va a estar entre 169 y 171 (es decir entre $\overline{x}$ -d y $\overline{x}$ +d). . Como además se tiene $\mu \in \overline{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$ habrá de ser $d = z_\alpha \frac{\sigma}{\sqrt{n}}$, y despejando n queda:

$$n = \frac{z_\alpha^2 \cdot \sigma^2}{d^2} \quad \textbf{(4.4)}$$

La expresión (4.4) tiene la desventaja de depender de σ^2, valor desconocido usualmente.

Tenemos varias alternativas para resolver este inconveniente:

1º)Sustituir σ^2 por el valor máximo que se piense pueda tomar, según nuestras experiencias previas. En el peor de los casos n será mayor de lo necesario. Quedaría:

$$n = \frac{z_\alpha^2 \cdot \{Max\ \sigma^2\}}{d^2} \quad \textbf{(4.5)}$$

2º) Tomar una muestra piloto de tamaño n' pequeño, obtener en ella su varianza s^2 y entonces:

$$n = \frac{t_\alpha^2 \cdot s^2}{d^2} \quad \textbf{(4.6)}$$

con t_α en la Tabla de la t de Student con n'-1 g.l.

3º) Enunciar la precisión en términos de fracciones de σ. Así, si deseamos ocurra que $|\overline{x} - \mu| \leq K\sigma$ con una confianza 1-α , cambiando d por $K\sigma$ en la (4.4) queda:

$$n = \frac{z_\alpha^2}{K^2} \quad \textbf{(4.7)}$$

Ejemplo 3: Determinar el tamaño de muestra requerido para obtener la estatura media de la población, con una precisión de 1 cm, si la varianza poblacional es $\sigma^2 = 25$.

Tomando n=97 individuos, según la fórmula (4.4) la media de ellos estará en el intervalo $x \pm 1$ al 95% de confianza. El redondeo se hace siempre por exceso pasa asegurar la precisión.

Ejemplo 4: Determinar el tamaño de la muestra para obtener la estatura media de una población con una precisión de 0.3σ.

Ahora n=43, según la expresión (4.7),y, entonces la media está en $\overline{x} \pm 0.3\sigma$

Ejemplo 5: Con datos del Ejemplo 1 como muestra piloto, determinar n con precisión d=4cm

Ahora n´=10 y $n = \left(\frac{2{,}262 \cdot 4{,}3}{4}\right)^2 = 5{,}9 \cong 6$. Como $6 < 10 = n'$, ello indica que con la muestra piloto nos basta para la precisión deseada.

Ejemplo 6: Igual que el anterior pero exigiendo un d = 1 cm.

De nuevo n´= 10 y ahora $n = \left(\frac{2{,}262 \cdot 4{,}3}{1}\right)^2 = 94{,}6 \cong 95$, con lo que son precisos 85 individuos más que antes.

Intervalo de confianza para una proporción.

Vamos a empezar este apartado planteando un ejemplo.

Ejemplo 7: Si de 100 personas encuestadas, 30 se manifiestan a favor de un determinado partido político, ¿qué porcentaje de votos obtendría dicho partido de celebrarse en ese momento las elecciones" (confianza del 95%)

Obsérvese que x="nº de individuos, entre los 100 encuestados, que votarán al candidato" es una Binomial de parámetro n = 100 y p desconocido. El objetivo es determinar p teniendo en cuenta que x sigue una B(n,p), con n = 100 y x = 30 el valor obtenido experimentalmente de esa Binomial. Conviene expresar que todo lo que sigue contiene las fórmulas para p expresadas en tantos por uno, no en %.

Intervalo.

La distribución Binomial, bajo ciertas circunstancias, se aproxima a una Normal. Los resultados siguientes se basan en esta aproximación. La expresión más tradicional del intervalo de confianza para una proporción p es la siguiente:

$$p \in \left\{ \frac{x \pm z_\alpha \sqrt{\frac{x(n-x)}{n}} \pm 0{,}5}{n} \right\} \quad \textbf{(4.8)}$$

Esta expresión es válida si x > 20 y n-x >20.Tiene la ventaja de ser cómoda, pero a cambio es más imprecisa y tiene unas condiciones de validez más exigentes. La siguiente expresión es más exacta (pero más incómoda) y para su validez basta con que sean x > 5 y n - x > 5:

$$p \in \frac{1}{n + z_\alpha^2} \left\{ (x \pm 0{,}5) + \frac{z_\alpha^2}{2} \pm z_\alpha \sqrt{\frac{z_\alpha^2}{4} + \frac{(x \pm 0{,}5)(n - x \pm 0{,}5)}{n}} \right\} \quad \textbf{(4.9)}$$

Ejemplo 7(continuación):

Aquí n = 100 y x = 30. Como x > 20 y n - x = 70 > 20, se puede utilizar (4.8):

$p \in \frac{1}{100}\left\{30 \pm 1{,}96\sqrt{\frac{30 \cdot 70}{100}} \pm 0{,}5\right\} = (0.2052\ ,\ 0{,}3948)$, es decir que piensan votar al partido entre un 20.52% y un 39.48% de la población. Si usamos la (4.9) que es más exacta:

$$p \in \frac{1}{100 + 1{,}96^2}\left\{(30 \pm 0{,}5) + \frac{1.96^2}{2} \pm 1.96\sqrt{\frac{1.96^2}{4} + \frac{(30 \pm 0{,}5)(70 \pm 0{,}5)}{100}}\right\} = (0.2145\ ;\ 0{,}4011)$$

para obtener este intervalo, se han considerado en primer lugar todos los signos (-) y después todos los signos (+).

Tamaño de la muestra

Ejemplo 8: En relación con el ejemplo anterior, el partido político desea realizar una encuesta con el fin de determinar el porcentaje de votantes con una precisión del 3% ¿A cuántos individuos hay que encuestar (confianza del 95%).

El objetivo es decidir a qué número n de individuos hay que preguntar para que el porcentaje de votos favorables entre ellos difiera del porcentaje nacional en menos de d = 3%.

Esto garantiza que, tomada la muestra, si el porcentaje en ella es de 30% el porcentaje nacional será 27% < p < 33%, es decir que p está en 30% ±3% con una confianza del 95%

De un modo general, si d es la precisión (máxima diferencia a admitir entre la estimación y p), hay una fórmula paralela a la (4.4):

$$n = \frac{z_\alpha^2 pq}{d^2} \quad \textbf{(4.10)}$$

La idea es tener garantías de que tomando una muestra de tamaño n, la proporción poblacional p de individuos que verifican la característica es, con una confianza de $(1 - \alpha)$, alguno de los valores entre $p_i \pm d$, con p_i la proporción en la muestra y d un número dado de antemano.

El problema, una vez más, es que la expresión anterior depende de p (que es desconocido). Puede demostrarse que pq es tanto mayor cuanto más se aproxime p a 0,5 alcanzando el máximo cuando p = 0,5, o sea,

$$\underset{0 \le p \le 1}{Max}\ pq = \frac{1}{4} \quad \textbf{(4.11).}$$

Como sucede en todas las fórmulas de tamaño de muestra, n es tanto más grande cuanto mayor sea la confianza del intervalo y cuanto menor sea d (cuanta mayor precisión se desee). La (4.11) aporta una novedad: el tamaño de la muestra es más grande cuanto más se aproxime p al valor 0,5, disminuyendo cuando nos enfrentemos a caracteres raros (p pequeño) o muy frecuentes (p grande). Igual sucede con la anchura de los intervalos de confianza para p: son más anchos cuanto más se acerque p a 0,5. Volviendo al problema del desconocimiento de p, la aplicación de (4.10) puede hacerse de dos modos:

1º) Si no se tiene idea alguna acerca de su posible valor, sustituir pq por 1/4, quedando:

$$n = \frac{z_\alpha^2}{4d^2} \quad \textbf{(4.12)}$$

2º) Si se tiene alguna información, sustituir p por el valor más cercano posible (y compatible con la información) a 0,5.

Ejemplo 8 (continuación):

Si el partido es nuevo y no se tiene idea acerca del porcentaje posible de votos favorables, sería $\mu \in 169{,}4 \pm 2{,}262\frac{4{,}3}{\sqrt{10}} = 169 \pm 3{,}08 = (166{,}32\ ,\ 172{,}48)$

Si el partido sabe que nunca en elecciones anteriores ha obtenido más del 30% de los votos y le sorprendería que esto no siguiera siendo así, sería

$$n = \frac{1.96^2 \cdot 0,3 \cdot 0,7}{(0.03)^2} \approx 897$$

BIBLIOGRAFIA

FUNDAMENTOS DE BIOESTADISTICA

Pagano Gauvreau

Thomas Learning

ESTADISTICA PARA ADMINISTRACION Y ECONOMIA

Anderson Sweeney Williams

Thomson Editores