Documento 3/1
Santiago de Chile

NOCIONES DE ESTADISTICA*/

*/ Tomado de Temas de Estadística, preparado por un grupo de Profesores Licenciados en Ciencias Exactas y Estadísticos Facultativos del Instituto Nacional de Estadística, Madrid, 1957, y reproducido para uso exclusivo del Curso de Capacitación en Planificación de los Recursos Humanos, 1968, organizado por la Oficina Internacional del Trabajo (OIT) y el Instituto Latinoamericano de Planificación Económica y Social (ILPES).

A. FRECUENCIAS Y PROBABILIDADES. HISTOGRAMA Y DISTRIBUCION DE PROBABILIDADES

Variables Estadísticas Cuantitativas

El conjunto de valores numéricos que resultan de una observación estadística constituye una variable estadística cuantitativa.

Por ejemplo, si en un examen de los 50 alumnos de una clase las puntuaciones obtenidas han sido:

2	1	5	7	6	5	6	4	2	4
0	6	1	10	7	2	9	5	4	6
3	5	8	1	7	8	3	7	6	5
4	5	6	9	8	4	7	5	1	0
5	10	5	6	5	4	2	3	5	6

la variable estadística cuantitativa es, en este caso, la puntuación, y toma valores comprendidos entre 0 y 10.

Son ejemplos de variables estadísticas cuantitativas la edad de los habitantes, la talla de los soldados de un regimiento, el peso de los alumnos de un curso, el precio de un artículo en los distintos mercados de una ciudad, el jornal de los obreros de una fábrica, la cotización de una moneda en la Bolsa durante un mes, la cuantía de la renta de los contribuyentes de una provincia, el número de horas de duración de las lámparas de un lote de fabricación, etc.

Las variables cuantitativas pueden ser continuas y discretas; son contínuas aquellas en que las unidades en las que vienen expresados los datos admiten subdivisión y pueden, por tanto, teóricamente, tomar todos los valores de un cierto intervalo, como la talla, cuya unidad de expresión - el metro - puede dividirse en centímetros, milímetros, etc.; son variables discretas las que no admiten aquella subdivisión de unidades, como la fecundidad de las mujeres, que se expresa por número de hijos habidos y cuya unidad no admite divisiones.

Llamaremos frecuencia absoluta, o simplemente frecuencia de un dato, al número de veces que se presenta en el total de observaciones. En el cuadro anterior de puntuaciones, la frecuencia de la puntuación 5 es 12, pues es el número de alumnos que figura en ese cuadro con 5 puntos.

/Asimismo se

Asimismo se denomina frecuencia relativa el cociente entre la frecuencia absoluta de un dato y el total de ellos u observaciones.

En el ejemplo citado, la frecuencia relativa de la puntuación 5 es $\frac{12}{50}$

Cuando los datos llegan al estadístico, éste se encuentra con una serie de números desordenados, como en el cuadro de puntuaciones a que nos estamos refiriendo. La primera operación lógica debe ser, entonces, la ordenación del conjunto de datos de la variable conforme a un criterio, que puede ser, por ejemplo, en orden creciente.

Las puntuaciones anteriores ordenadas con ese criterio quedarían así, en sentido horizontal:

0	0	1	1	1	1	2	2	2	2
3	3	3	4	4	4	4	4	4	5
5	5	5	5	5	5	5	5	5	5
5	6	6	6	6	6	6	6	6	7
7	7	7	7	8	8	8	9	9	10

Este conjunto de valores tiene ya una estructura más apta para cualquier clasificación y estudio; pero, sin embargo, todavía es susceptible de una mejor exposición.

Tablas de Frecuencias

Dentro del proceso de elaboración de una estadística, la operación que consiste en expresar los resultados de la observación y recuento por medio de tablas o cuadros se denomina tabulación. Los cuadros o tablas numéricas comenzaron a usarse en el año 1740, atribuyéndose su introducción al danés Anchersen. Su uso desde entonces ha sido constante, constituyendo un medio insustituible para la exposición de los resultados de las investigaciones sobre fenómenos colectivos o aleatorios.

Uno de los tipos de tablas que más se presentan en Estadística son las de distribuciones de frecuencia o simplemente tablas de frecuencias.

Se entiende por distribución de frecuencia la exposición de la correspondencia entre los valores de la variable y sus repsectivas frecuencias. Se forma para ello una tabla de simple entrada con dos columnas; en

./la primera

la primera se colocan las diferentes modalidades numéricas de la variable, generalmente en orden creciente, y en la segunda, las frecuencias o número de veces de cada una de aquéllas.

Así, en el cuadro de las 50 puntuaciones formaríamos la siguiente tabla de frecuencias:

Puntuación x_i	Número de alumnos n_i
0	2
1	4
2	4
3	3
4	6
5	12
6	8
7	5
8	3
9	2
10	1
TOTAL	50

Los distintos valores de la variable los designaremos por x_1, x_2, x_3..., x_i..., y los de las frecuencias por n_1, n_2, n_3..., n_i...

Tablas de frecuencias agrupadas.

Cuando el número de valores de la variable es grande o ésta es contínua, se forma la tabla de frecuencias, agrupándolas en intervalos, como vamos a explicar a continuación.

Supongamos que las cotizaciones de los títulos, de una sociedad eléctrica han sido las siguientes durante los 40 días hábiles de un trimestre:

187	156	194	172	181	168	199	204
161	159	188	190	174	201	170	183
177	144	197	178	183	141	196	184
191	181	151	167	179	151	188	173
186	173	172	147	180	175	186	164

/Si en

Si en vez de colocar ordenados, uno por uno, esos valores, los agrupamos clasificándolos en clases, que constituyen intervalos, con sus límites inferior y superior, y contamos cuantos valores hay en cada intervalo, la tabla de frecuencias agrupadas quedará así, eligiendo, por ejemplo, intervalos de 10 valores de amplitud:

Cotización x_i	Número de días n_i
De 140 a 150	3
" 150 a 160	4
" 160 a 170	4
" 170 a 180	10
" 180 a 190	11
" 190 a 200	6
" 200 a 210	2
TOTAL	40

A la diferencia entre dos límites inferiores consecutivos la denominaremos amplitud de la clase o del intervalo.

Si bien la igualdad de amplitud de todos los intervalos es muy útil, por favorecerse notablemente la aplicación de algunos procedimientos estadísticos fundamentales, a veces la irregularidad con que se presentan las frecuencias aconseja introducir variaciones de amplitud de unas clases a otras. En el estudio de la mortalidad por grupos de edad, aunque suelen considerarse clases de cinco años, el de menores de un año presenta características tan típicas y tan distintas respecto a las restanted edades, que la tabla de frecuencias suele disponerse así:

Grupos de edad x_i	Número de fallecidos n_i
Menores de 1 año	
De 1 a 5 años	
De 5 a 10 años	
.	
De 60 a 70 años	
De 70 a 80 años	
De 80 y más	

/siendo el

siendo el primer intervalo de un año de amplitud y de cinco los siguientes; al llegar, en cambio, a los 60, suelen agruparse de diez en diez años, por no interesar ya tanto la discriminación.

Generalmente, y por ser muy pequeñas las frecuencias, en los intervalos extremos, a fin de evitar el tener que escribir clases cuya frecuencia sea nula, suelen ampliarse hasta los límites de variabilidad total del fenómeno las dos clases extremas colocando en la primer la letra α como límite inferior, y en la última, ω como límite superior, quedando así, por ejemplo:

a	———	100
100	———	150
150	———	200
200	———	250
250	———	300
300	———	ω

Tabla de frecuencias relativas.

Si la frecuencia de cada clase se divide por el total de frecuencias, se forma una distribución de frecuencias relativas, cuya suma es 1. Para evitar cifras decimales en las frecuencias, suelen multiplicarse todos los cocientes por la unidad seguida de los ceros necesarios, y el total, en vez de 1, pasará a ser esa potencia de 10 por la que se ha multiplicado, obteniéndose así frecuencias por 100, 1.000, etc. Las frecuencias relativas las designaremos por h_1, h_2,..., h_i,...

Vamos a formar la tabla de frecuencias relativas deducida de la distribución de las cotizaciones, anteriormente expuesta. Dividiendo cada una de las frecuencias absolutas n_i, 3, 4, 4, 10, etc., por el total 40, los cocientes $\frac{3}{40}$, $\frac{4}{40}$, $\frac{4}{40}$, $\frac{10}{40}$, etc., serán las frecuencias relativas de cada clase o intervalo, y su suma, naturalmente, 1. Suelen escribirse en forma decimal en vez de fraccionaria. La tabla quedaría así, con otra columna más, resultando de multiplicar las frecuencias relativas por 1.000.

/Cotización

Cotización x_i	Días n_i	Días Total h_i	Días / Total X 1.000
De 140 a 150	3	0,075	75
" 150 a 160	4	0,100	100
" 160 a 170	4	0,100	100
" 170 a 180	10	0,250	250
" 180 a 190	11	0,275	275
" 190 a 200	6	0,150	150
" 200 a 210	2	0,050	50
TOTALES	40	1,000	1,000

Tabla de frecuencias acumuladas.

Es también muy conveniente formar clases cada vez mayores y con uno de los límites, el inferior o el superior, común a todas ellas. Es decir, en el repetido ejemplo anterior, la primera clase sería hasta 150, la segunda abarcaría todas las frecuencias de las dos primeras clases y llegaría hasta 160, y así sucesivamente. Tales frecuencias se denominan acumuladas y las representaremos por N_i.

De análogo modo al explicado para formar las frecuencias relativas puede formarse otra columna con las frecuencias relativas acumuladas.

Sea la siguiente distribución de frecuencias, que se refiere a las fábricas de cemento clasificadas por su capacidad anual de producción:

Miles de toneladas x_i	Número de fábricas n_i
De 0 a 20	2
" 20 a 40	2
" 40 a 60	1
" 60 a 80	4
" 80 a 100	4

/Miles de

(cont.)

Miles de toneladas x_i	Número de fábricas n_i
De 100 a 120	7
" 120 a 140	5
" 140 a 160	9
" 160 a 180	5
" 180 a 200	2
" 200 a 220	1
TOTAL	42

La distribución de frecuencias acumuladas la formaremos, con las clases, hasta 20, hasta 40, hasta 60, etc.; es decir, que cada una abarca desde 0 hasta el límite que indica, y entonces a la primera clase le corresponderá la frecuencia 2; a la segunda, la suma de 2 + 2, o sea 4; a la tercera, la suma de 4 + 1, es decir, 5, etc., quedando, pues, así:

Miles de toneladas	Número de fábricas
Hasta 20	2
" 40	4
" 60	5
" 80	9
" 100	13
" 120	20
" 140	25
" 160	34
" 180	39
" 200	41
" 220	42

Consideraciones generales en la formación de tablas de frecuencia.

La distribución de los datos de un colectivo observado, en tabla de frecuencias, es una operación fundamental que no puede, por ello realizarse en forma arbitraria.

Para conseguir formar una tabla de frecuencias con un buen criterio, debemos considerar:

/a) El

a) El recorrido u oscilación extrema.

b) El número de clases.

c) Los límites de las clases.

d) El punto medio de cada clase.

El recorrido u oscilación extrema es la diferencia entre los datos de mayor y menor valor. En el primer ejemplo de las puntuaciones sería 10 - 0 = 10, y en el de las cotizaciones de títulos, como la cotización más alta es 204 y la más baja 141, la oscilación es 63.

El conocimiento de este dato está íntimamente ligado al número de clases.

Para establecer en cada distribución el número de clases no pueden darse reglas fijas, pues hay que estudiar dicha distribución e incluso hacer ensayos hasta conseguir hacer resaltar las características del fenómeno observado. La mayoría de los estadísticos creen que el número de clases o intervalos no debe ser inferior a 7 ni pasar de 20.

Una vez fijado el número de intervalos, la amplitud de ellos, si ésta es constante, será el cociente entre el recorrido y aquel número.

Para evitar dudas y confusiones, debe tenerse sumo cuidado en que un dato no pueda quedar incluído teóricamente en dos clases, porque coincida su valor con el límite superior de una y el inferior de la siguiente. En la tabla de frecuencias de las cotizaciones, tal como está presentada, no sabemos si el dato 150 hay que contarle como correspondiente al primer intervalo o al segundo.

Para que desaparezca la ambigüedad, el procedimiento correcto más generalizado consiste en hacer que el límite superior de cada clase y el inferior de la siguiente se diferencien en una unidad de orden decimal más alto que aquel en que vienen expresados los datos. Si las cotizaciones vienen expresadas en números enteros, los intervalos deben escribirse así:

/Cotización

Cotización x_i	Núm. de días n_i
De 140 a 149,9	3
" 150 a 159,9	4
" 160 a 169,9	4
" 170 a 179,9	10
" 180 a 189,9	11
" 190 a 199,9	6
" 200 a 209,9	2
TOTAL	40

con cuya forma no queda ninguna duda respecto al intervalo de inclusión de cada frecuencia.

Denominaremos centro del intervalo o marca de clase al punto medio de cada intervalo, y se halla como media entre los límites inferiores consecutivos.

En la distribución de las fábricas de cemento, la media entre 0 y 20 es 10; entre 20 y 40, 30; entre 40 y 60, 50, etc., las marcas de clase se escriben inmediatamente a la derecha de los intervalos:

Valores	Marcas de clase x_i	Frecuencias n_i
De 0 a 19,9	10	2
" 20 a 39,9	30	2
" 40 a 59,9	50	1
" 60 a 79,9	70	4
" 80 a 99,9	90	4
" 100 a 119,9	110	7
" 120 a 139,9	130	5
" 140 a 159,9	150	9
" 160 a 179,9	170	5
" 180 a 199,9	190	2
" 200 a 219,9	210	1

/La marca

La marca de clase representa en todos los cálculos estadísticos a su clase; y como se admite que dentro de cada una de ellas la frecuencia correspondiente se distribuye de manera uniforme, dicho punto central del intervalo es el valor teórico al que se atribuye la frecuencia de su clase.

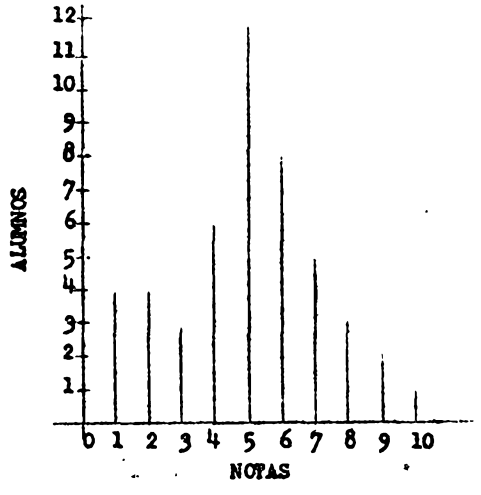
Representaciones Gráficas

Complemento necesario de toda tabla de frecuencias es su representación gráfica en un sistema de coordenadas cartesianas rectangulares.

Los tipos de gráficos característicos de las variables estadísticas cuantitativas son: diagrama de barras, histograma de frecuencias, polígono de frecuencias y curva acumulativa u ojiva.

1. Diagramas de barras. Son gráficos construídos para representar distribuciones de frecuencias cuyos valores no estén agrupados. Tomando esos valores como alscisas, en esos puntos se trazan ordenadas iguales a las frecuencias respectivas.

Para representar la distribución de las notas de los 50 alumnos (primer ejemplo del tema), el gráfico sería así:



/2. Histograma.

2. Histograma de frecuencias. Es típica esta representación gráfica para tablas de frecuencias agrupadas y se realiza de la manera siguiente:

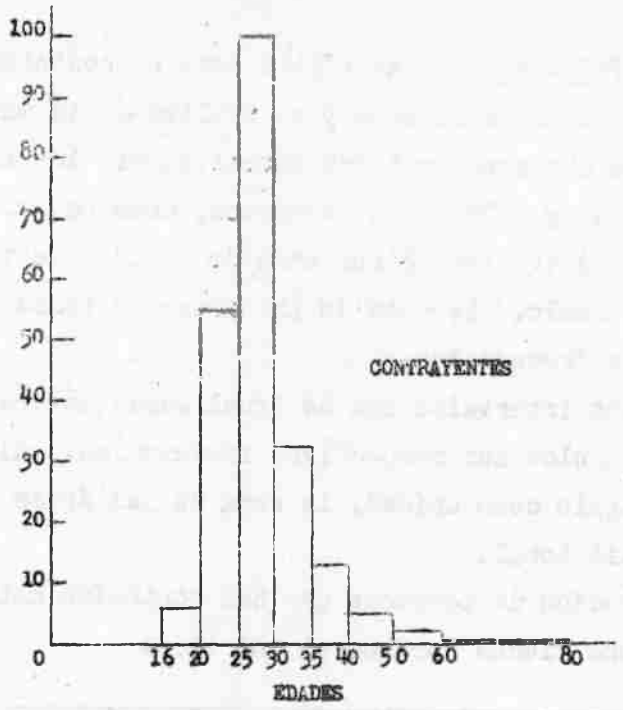
Sobre el eje de abscisas se toman sucesivamente los intervalos de clases de la variable, y sobre esos segmentos, como bases, se construyen rectángulos de alturas iguales al cociente de dividir su frecuencia por la amplitud del intervalo. La suma de las áreas de todos los rectángulos equivale al total de frecuencias.

Cuando todos los intervalos son de igual amplitud, basta tomar como altura de los rectángulos sus respectivas frecuencias. Al considerar la base de cada rectángulo como unidad, la suma de las áreas será también igual a la frecuencia total.

Sea la distribución de personas que han contraído matrimonio por grupos de edad en una ciudad durante el año 1956:

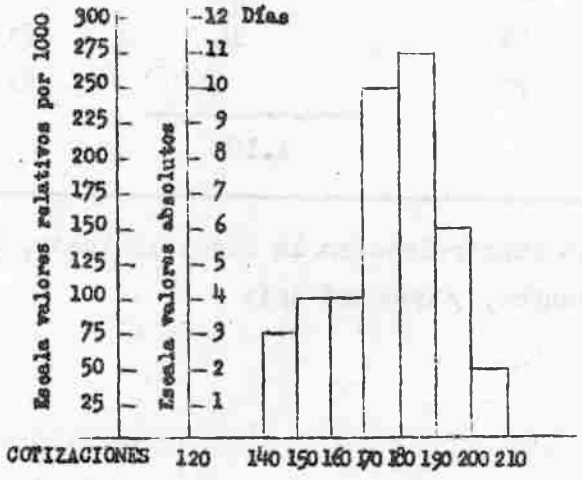
Grupos de edad	Marcas de clase x_i	Número de contrayentes n_i	Cociente de n_i por la amplitud de cada intervalo
16 a 19 años	18	22	22: 4 = 5,5
20 a 24 "	22,5	275	275: 5 = 55,0
25 a 29 "	27,5	508	508: 5 = 101,6
30 a 34 "	32,5	162	162: 5 = 32,4
35 a 39 "	37,5	64	64: 5 = 12,8
40 a 49 "	45	49	49: 10 = 4,9
50 a 59 "	55	14	17: 10 = 1,4
60 a 80 "	70	6	6: 20 = 0,3
TOTAL		1.100	

El histograma lo construímos en la forma indicada, tomando como ordenadas estos cocientes, y quedará así:



Consideremos ahora nuevamente la distribución relativa a las cotizaciones, en la cual los intervalos eran de la misma amplitud, y de la cual obtuvimos también las frecuencias relativas por 1.000. Pues bien: construyendo dos escalas verticales, una para las frecuencias absolutas y otra para las relativas, con la natural correspondencia de valores, que en el caso de que tratamos es 40 equivalente a 1.000, el mismo histograma servirá para representar las dos distribuciones, de valores absolutos y de valores relativos.

El gráfico sería:

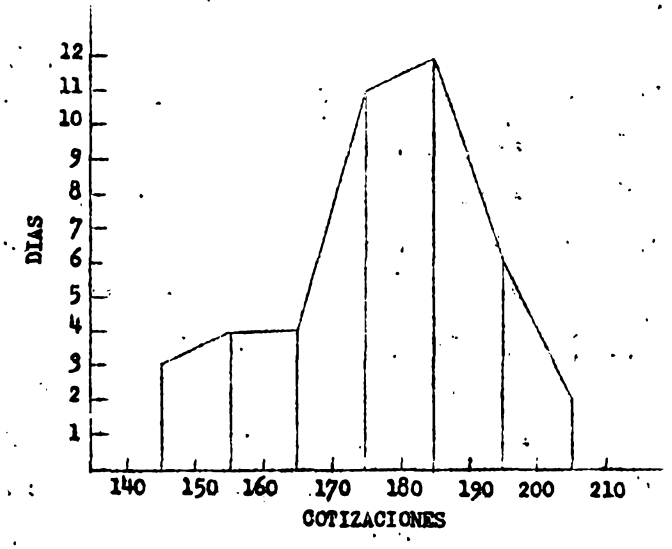


/3. Polígono

3. Polígono de frecuencias. Se obtiene levantando ordenadas en los puntos medios de las clases, de altura igual o proporcional a las frecuencias respectivas y uniendo los extremos superiores de dichas ordenadas por una línea poligonal.

Hay que advertir que los segmentos de esa línea poligonal no tienen significación estadística alguna.

La representación gráfica del polígono de frecuencias para los mismos datos anteriores sería:



Si los valores de la variable variasen de forma continua, es decir, que los intervalos fuesen infinitésimos y el número de observaciones fuese cada vez mayor, el polígono anterior se aproximaría a una línea curva ideal, que recibe el nombre de curva de frecuencia, de gran importancia en Estadística.

Aunque las formas que ofrecen las series estadísticas son infinitas en su variedad, pueden reducirse a cuatro tipos generales, a los que se aproximan todas las distribuciones de frecuencia:

Tipo 1º. Distribuciones simétricas (fig. 1). En ellas la distribución adopta la misma forma a ambos lados de la ordenada máxima. En la práctica es dificilísimo que se dé el caso de obtener una curva perfectamente simétrica.

/Suelen presentarse

Suelen presentarse estas distribuciones en forma aproximada a la simetría en la observación da datos biométricos, talla, perímetro torácico, etc.; en algunos hechos demográficos, como el estudio de la probabilidad del nacimiento de uno de los sexos; en algunas investigaciones biológicas, principalmente en Botánica; en Meteorología, en el reparto de temperaturas medias, etc. Reciben también el nombre de curvas en forma de campana.

Tipo 2º. Distribuciones moderadamente asimétricas (fig. 2). Son aquellas en que las fecuencias disminuyen más rápidamente a un lado del máximo que al otro y se presentan en muchos campos de observación.

Tipo 3º. Distribuciones fuertemente asimétricas o en forma de J (fig. 3). En ellas las frecuencias crecen hasta un máximo que se halla en uno de los extremos del campo de la variable.

Aparecen estas curvas en algunos hechos económicos, como distribución de rentas; en otros demográficos, como al estudiar la probabilidad de fallecimiento por disentería en las diferentes edades, etc.

Tipo 4º. Distribuciones en forma de U (fig. 4). Presentan dos máximos en los extremos del campo de variación y un mínimo en el centro.

En metereología hay ejemplos típicos de estas curvas al observar los grados de nubosidad del cielo, pues corresponde un máximo al valor más pequeño, cielo despejado, y otro para el mayor, cubierto totalmente. En Demografía, al representar el número de mujeres por el número de hijos tenidos, aparece un máximo para el caso 0, ningún hijo, y otro para 2.

A continuación se representan los cuatro tipos de curvas de frecuencia reseñados.

FIG. 1

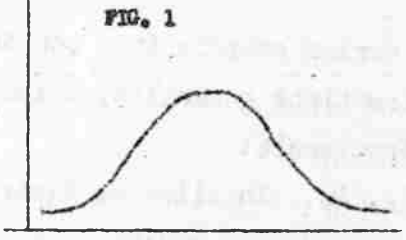
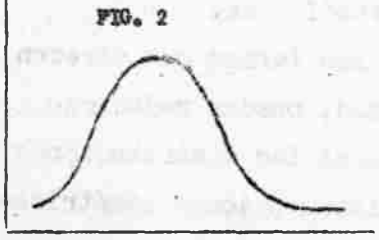


FIG. 2



/Fig. 3

FIG. 3

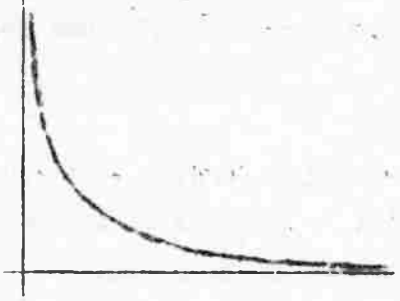
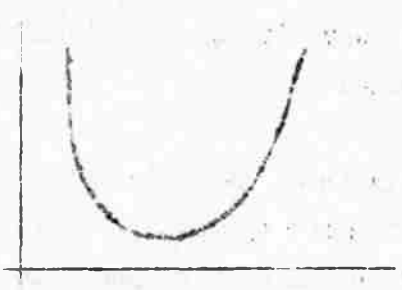


FIG. 4

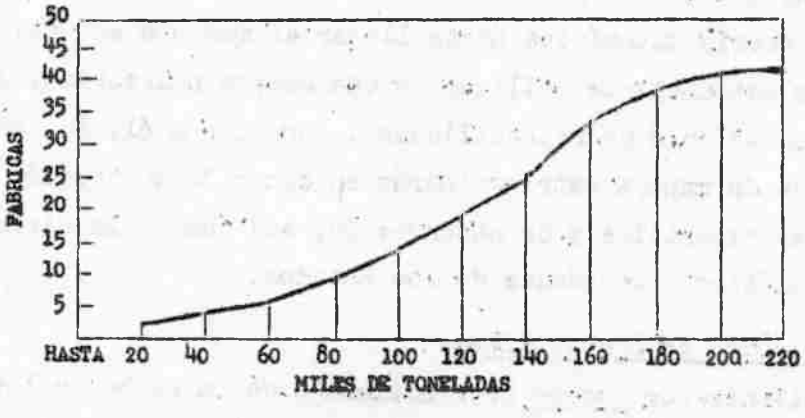


4. Curva acumulativa u ojiva. Las distribuciones acumuladas pueden representarse gráficamente de la siguiente forma:

En los valores de las abscisas correspondiente a los límites superiores de las clases se van levantando ordenadas de altura igual o proporcional a las frecuencias acumuladas. Uniendo los extremos de dichas ordenadas por una línea poligonal resulta el polígono acumulativo, cuyo límite teórico es la curva acumulativa.

Debido a la forma que generalmente adopta la curva, suele designarse con el nombre de ojiva.

En la distribución de las fábricas de cemento calculamos la columna de frecuencias acumuladas. Su representación gráfica será:



/Si la

Si la curva límite del histograma es expresable por una función analítica continua,

$$y = f(x),$$

la curva límite de la distribución acumulativa será la integral de esa función y tendrán que coincidir en la misma ordenada el máximo de la curva de frecuencias y el punto de inflexión de la ojiva.

Probabilidades

Todos los hechos, sucesos o fenómenos que se producen en el mundo dependen del concurso de diversas causas y de la proporción variable en que éstas actúan.

No obstante, muchas veces esas causas nos son desconocidas, apareciendo así los sucesos o fenómenos de azar, aleatorios o estadísticos, y de los cuales se reseñaron sus principales características en el tema 1.

Estos fenómenos aleatorios constituyen, como ya se ha explicado también, el campo de aplicación de la estadística y el fundamento matemático de ella es el llamado Cálculo de Probabilidades o Teoría de la probabilidad.

Como todas las ramas de la Matemática, el Cálculo de Probabilidades surgió como consecuencia del planteamiento de problemas concretos de carácter práctico, en este caso, los juegos de azar, principalmente los de dados, muy en boga en los siglos XVI y XVII. Primero Pascal, luego Jacobo Bernouilli, y después numerosos matemáticos, han ido perfeccionando esta teoría matemática hasta llegar al momento actual, en el que los métodos estadísticos utilizan de una manera necesaria y eficaz los teoremas del Cálculo de Probabilidades. Gracias a él, la Estadística ha extendido de manera extraordinaria su campo de aplicación a todas las ciencias experimentales y de observación, así como a la dirección administrativa, política y económica de los Estados.

Concepto clásico de probabilidad.

Se entiende por suceso la realización de un hecho cualquiera: el nacimiento de un varón, el hacer blanco un torpedo en un barco, un accidente de aviación, el obtener "cara" al lanzar al aire una moneda,

/el cometer

el cometer un error en la medida de una longitud, la extracción de una bola de una urna, la cotización de un valor industrial en la Bolsa, etc.

Aún con desconocimiento absoluto del concepto matemático, en el lenguaje vulgar, la noción de probabilidad se usa en forma más o menos precisa para indicar el mayor o menor grado de confianza que se tiene en la realización de un suceso.

Pues bien, clásicamente, se define la probabilidad de un suceso, por el número que resulta del cociente entre el número de casos en que puede presentarse (casos favorables) y el número total de casos posibles, con la condición de que todos éstos tengan igual posibilidad de presentarse.

Llamando p a la probabilidad, n_1 a los casos favorables y N a todos los casos posibles, se tiene:

$$p = \frac{n_1}{N}$$

Se deduce inmediatamente, por la propia definición, que el valor de p será siempre un número comprendido entre 0 y 1, correspondiendo los casos extremos a la imposibilidad y a la certeza, respectivamente.

Ejemplos: Si se lanza un dado al aire sobre una mesa, y está construído de forma homogénea, cualquier cara tiene la misma posibilidad de aparecer; serán, pues, seis los casos posibles. La probabilidad de obtener el número 2 será, puesto que hay una sola cara con este número de puntos:

$$p = \frac{1}{6}$$

Análogamente, la probabilidad de obtener múltiplos de 2 será $\frac{3}{6}$, pues son tres los casos posibles, las caras 2, 4 y 6.

Si en una urna hay 14 bolas blancas y 6 negras, la probabilidad de extraer una bola negra será:

$$p = \frac{6}{20} = \frac{3}{10}$$

Si con esa misma urna se quisiera averiguar la probabilidad de que la bola extraída sea roja, al no haber ningún caso favorable, el numerador de la fracción es 0 y la probabilidad nula.

Si lo que se desea es hallar la probabilidad de que la bola sea blanca o negra, como coinciden los casos favorables con los posibles,

$$p = \frac{20}{20} = 1$$

y se obtiene la certeza.

La definición de probabilidad que se acaba de dar es debida a Laplace y se apoya en el postulado de la indiferencia, que establece que si un fenómeno aleatorio puede dar lugar a la presentación de varios sucesos distintos y no hay ninguna razón para suponer que está favorecida la presentación de alguno sobre los restantes, todos tienen la misma probabilidad.

Por dar lugar la definición clásica de probabilidad a paradojas y no poderse aplicar a determinados grupos de problemas, modernamente se establece el concepto de probabilidad y las primeras propiedades, a partir de un conjunto de postulados y axiomas, tras de los cuales se desarrolla la teoría matemática del Cálculo de Probabilidades.

Suceso contrario.

Se denomina suceso contrario de uno dado a la realización de la modalidad opuesta a la que se ha tomado como favorable, es decir, el que se verifica cuando no se realiza éste.

Si un hecho puede presentar n_1 casos favorables y n_2 desfavorables, o contrarios, la probabilidad de obtener suceso favorable es:

$$p = \frac{n_1}{n_1 + n_2}$$

/puesto que

puesto que $n_1 + n_2$ es el conjunto de todos los posibles.

Análogamente, designando por q la probabilidad de realización del hecho desfavorable o suceso contrario, es:

$$q = \frac{n_2}{n_1 + n_2}$$

Sumando miembro a miembro ambas igualdades, queda:

$$p + q = \frac{n_1 + n_2}{n_1 + n_2} = 1 \; ; \quad q = 1 - p$$

es decir, la suma de probabilidades de dos sucesos contrarios es 1.

Ejemplo: La probabilidad de sacar un as en la extracción de una carta de las 40 de la baraja española, es:

$$p = \frac{4}{40} \; ,$$

y la probabilidad de sacar una carta que no sea as,

$$q = 1 - p = 1 - \frac{4}{40} = \frac{36}{40}$$

Sucesos elementales y sucesos complejos.

Se entiende por suceso elemental o simple aquel cuya realización se considera aisladamente y no está ligada a ninguna circunstancia distinta de las suyas propias.

Por analogía se denominará probabilidad simple la que se refiere a un suceso simple. Todos los ejemplos citados hasta ahora son de probabilidad simple.

En cambio, se llamará suceso complejo aquel que se compone de varios sucesos simples, es decir, cuando su realización va unida a otros de una forma indistinta o disyuntiva, o bien por el concurso simultáneo o sucesivo de ellos.

/La probabilidad

La probabilidad de presentación de uno de aquellos sucesos se denominará probabilidad compleja.

Ahora bien: dentro de los sucesos complejos distinguiremos dos clases que no se excluyen: a) Sucesos compatibles e incompatibles; b) Sucesos dependientes o independientes.

Sucesos compatibles son aquellos que pueden presentarse o realizarse simultáneamente, e incompatibles, el caso contrario.

Si en el lanzamiento de un dado consideramos un suceso el obtener un múltiplo de 3 y otro suceso el obtener un número de puntos igual o menor que 4, ambos sucesos son compatibles.

Se dirá que dos sucesos son dependientes o independientes, según que la presentación simultánea o sucesiva de cada uno de ellos influya o no sobre el otro. Si lanzamos al aire dos monedas al mismo tiempo o sucesivamente, la aparición de cara o cruz en cada una de ellas no tiene ninguna relación entre sí; serán dos sucesos independientes. Pero si de una urna con bolas de varios colores extraemos una y después otra, esta segunda extracción está condicionada a la primera, puesto que según fuera el resultado de ésta variará la proporción de colores y, por tanto, la nueva probabilidad; serán pues, sucesos dependientes.

Teoremas Fundamentales

Teorema de adición o de probabilidad total.

Vamos a estudiar cómo se calcula la probabilidad de realización de un hecho, que se verifica por la presentación indistinta o disyuntiva de otros varios.

Se consideran dos casos: a) Que los sucesos componentes del resultado, sean compatibles; b) Que sean incompatibles.

Caso a) Los sucesos pueden verificarse simultáneamente, pero para la realización total basta que se produzca uno solo de ellos.

Si los sucesos que implican el resultado son A y B, sea n_1 el número de casos en que se realiza la presentación de A y B; n_2 el número de casos en que se realiza A, pero no B; n_3 en que se realice B, pero no A, y n_4, cuando no se producen ni A ni B.

/La probabilidad

La probabilidad de realización de A, puesto que los casos en que aparece son $n_1 + n_2$, será:

$$p(A) = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4}$$

y análogamente la de B:

$$p(B) = \frac{n_1 + n_3}{n_1 + n_2 + n_3 + n_4}$$

La probabilidad de que se realicen simultáneamente A y B, puesto que la frecuencia de esta posibilidad es n_1, será:

$$p(AB) = \frac{n_1}{n_1 + n_2 + n_3 + n_4}$$

Sumando las dos primeras igualdades y restando esta última, queda

$$p(A) + p(B) - p(AB) = \frac{n_1 + n_2 + n_3}{n_1 + n_2 + n_3 + n_4}$$

pero esta fracción representa la probabilidad de presentación de A o B, puesto que $n_1 + n_2 + n_3$ son todos los casos favorables a ambos sucesos.

Queda en definitiva:

$$p(A + B) = p(A) + p(B) - p(AB) \qquad [1]$$

teorema que dice: <u>La probabilidad total de un suceso, que puede realizarse por la presentación de dos hechos A o B, o de ambos conjuntamente, es la suma de las probabilidades de cada uno de éstos, menos la probabilidad simultánea de ambos</u>.

Ejemplo: Se va a extraer una carta de una baraja de 40 cartas, y se gana si se saca una <u>figura</u> (rey, caballo, sota) o un <u>basto</u>.

/La probabilidad

La probabilidad de sacar una figura será $\frac{12}{40}$, puesto que hay 12 figuras en las 40 cartas; la probabilidad de obtener un basto es $\frac{10}{40}$, porque cada palo tiene 10 cartas, y la probabilidad de obtener figura y que sea de bastos es $\frac{3}{40}$,

Luego la probabilidad total, o sea, la se ganar por la obtención de una figura o de un basto es, según [1]:

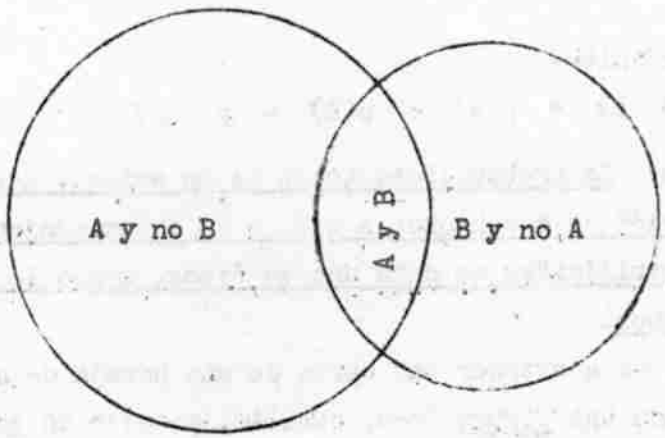
$$p(A + B) = \frac{12}{40} + \frac{10}{40} - \frac{3}{40} + \frac{19}{40}$$

Diagrama de Euler.

Consiste en representar gráficamente el caso de probabilidad total, cuando los sucesos componentes son compatibles, es decir, no se excluyen mutuamente.

Si se representan por un círculo el conjunto de casos en que aparezca el suceso A, y por otro círculo el de casos en que se presente B, al poder presentarse simultáneamente A y B, dichos círculos tendrán una parte común, es decir, el plano quedaría dividido en cuatro regiones, así:

ni A ni B



/y gráficamente

y gráficamente se ve que al sumar el área de A con la de B, habrá que restarla la común, para que quede la total A + B.

Caso b) Los sucesos son incompatibles, es decir, que se excluyen mutuamente. Al no poder producirse simultáneamente la presentación de A y B, la probabilidad conjunta p(AB) es 0 y la[1] queda:

$$p(A + B) = p(A) + p(B) \qquad [2]$$

o clásico teorema de las probabilidades totales en que los sucesos componentes son incompatibles.

Por generalización, si un suceso se verifica por la realización de uno cualquiera de los hechos A_1, A_2, A_3, ..., A_n, incompatibles entre sí, la probabilidad total del suceso es:

$$p(A_1 + A_2 + A_3 + \ldots + A_n) = p(A_1) + p(A_2) + p(A_3) + \ldots + p(A_n)$$

Ejemplos: 1°. Si en una bolsa hay 20 bolas blancas, 30 negras y 50 verdes, la probabilidad de sacar una bola blanca o verde es:

$$p(A + B) = \frac{20}{100} + \frac{50}{100} = \frac{70}{100} = 70\,\%$$

2°. Se tiene calculado, por ejemplo, que de cada 100 torpedos disparados contra un barco, 11 le alcanzarán en puntos vitales, 34, en sitios no vitales, y 55, no le alcanzarán. La probabilidad de hacer impacto en el barco es:

$$p(A + B) = \frac{11}{100} + \frac{34}{100} = \frac{45}{100} = 45\,\%$$

Teorema de composición o de probabilidad compuesta.

Consideremos ahora el problema de la realización de un suceso por el concurso simultáneo o sucesivo de otros dos sucesos. La probabilidad correspondiente se denomina compuesta.

Primer caso. Sean A y B dos sucesos dependientes, es decir, que la probabilidad del segundo está condicionada a la del primero; se la señala con la notación p(B|A), expresando con ella la probabilidad de

/que se

que se realice B en el supuesto de que se ha realizado A.

Sean, como antes, n_1, n_2, n_3 y n_4 el número de casos de presentación de A y B, de A y no de B, de B y no de A, y ni de A ni de B, respectivamente.

La probabilidad de presentación de A y B es:

$$p(AB) = \frac{n_1}{n_1 + n_2 + n_3 + n_4}$$

y la de presentación de A:

$$p(A) = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4}$$

Ahora, la probabilidad condicionada de B a A, puesto que los casos favorables son n_1, y los posibles, $n_1 + n_2$, es:

$$p(B|A) = \frac{n_1}{n_1 + n_2}$$

Multiplicando miembro a miembro las dos últimas igualdades, queda:

$$p(A) \cdot p(B) = \frac{n_1}{n_1 + n_2 + n_3 + n_4}$$

que según la primera igualdad, es la probabilidad de presentación simultánea o sucesiva de A y B, quedando en definitiva:

$$p(AB) = p(A) \cdot p(B|A) \qquad [3]$$

teorema que dice: La probabilidad compuesta de un suceso que se realiza por la presentación simultánea o sucesiva de dos hechos A y B, es el producto de la probabilidad aislada de A por la condicionada de B.

Al mismo resultado se llega comenzando por B, es decir:

$$p(AB) = p(B) \cdot p(A|B)$$

/Ejemplo: Si

Ejemplo: Si disponemos de una bolsa en la que hay 20 bolas blancas y 30 negras, y se quiere averiguar la probabilidad de extraer dos bolas blancas, simultánea o sucesivamente, sin reponer la bola extraída primeramente, se tendrá:

La probabilidad aislada de extraer bola blanca es:

$$\frac{20}{50} = \frac{2}{5}$$

Para conseguir que las dos bolas sean blancas, la segunda extracción está condicionada a que haya sido blanca la primera. Luego la probabilidad condicionada es $\frac{19}{49}$, y la pedida:

$$p(AB) = \frac{2}{5} \times \frac{19}{49} = \frac{38}{245}$$

Segundo caso. Si los sucesos componentes son independientes, esto es, la presentación de B no tiene relación con la de A, ni viceversa.

$$p(B|A) = p(B)$$

y, por tanto,

$$p(AB) = p(A) \cdot p(B) \qquad [4]$$

o sea, el clásico teorema de la probabilidad compuesta, en el caso de que los sucesos sean independientes.

Se generaliza al caso de más de dos sucesos componentes, y queda:

$$p(A_1, A_2, A_3 \ldots\ldots A_n) = p(A_1) \cdot (p(A_2) \cdot p(A_3) \ldots\ldots p(A_n)$$

Ejemplos: 1°. La probabilidad de sacar una carta de cada una de las 40 cartas de la baraja, y que sea una rey y la otra sota, es:

$$p(AB) = \frac{4}{40} \times \frac{4}{40} = \frac{1}{100}$$

2°. Si se lanzan dos dados, uno blanco y otro negro, y deseamos hallar la probabilidad de obtener un número par en el blanco y mayor que 4 en el negro, es:

/Probabilidad de

Probabilidad de obtener número par: $\frac{3}{6}$;

probabilidad de obtener número mayor que 4 (5 ó 6): $\frac{2}{6}$

$$p(AB) = \frac{3}{6} \times \frac{2}{6} = \frac{6}{36} = \frac{1}{6}$$

3º. La probabilidad de un tirador en hacer blanco es $\frac{3}{5}$. Se desea averiguar la probabilidad de hacer tres disparos y hacer blanco dos veces, fallando la tercera.

Como $\frac{3}{5}$ es la probabilidad aislada de cada tirada, $\frac{2}{5}$ será la de fallar. Por tanto,

$$p(AB) = \frac{3}{5} \times \frac{3}{5} \times \frac{2}{5} = \frac{18}{125}$$

Ley del azar

Si se considera un experimento bien definido, con una serie de pruebas sucesivas, y en las n primeras pruebas se ha presentado el suceso A, n_1 veces, sabemos que la frecuencia relativa es $\frac{n_1}{n}$. A medida que n va siendo mayor, estos cocientes toman valores cada vez más próximos unos a otros. Es una ley plenamente observada y comprobada, y se dice que esas frecuencias tienden a estabilizarse alrededor de un valor constante, que es la probabilidad, en los casos en que pueda calcularse teóricamente.

Esta ley del azar, en virtud de la cual cuando el número de pruebas en grande la frecuencia relativa del suceso difiere poco de su probabilidad, es la unión entre la práctica y la teoría, entre el concepto empírico o ley estadística y el modelo teórico o ley de probabilidad.

Entre las consecuencias importantes de esta ley del azar se encuentran el poder hallar la verdadera distribución que alcanzaría el hecho

/observado si

observado si se hiciesen experiencias ilimitadamente, y el determinar hasta qué punto la posible diferencia entre las frecuencias empírica y teórica es debida al azar o a una imperfección de los métodos empleados en el experimento.

Ejemplo: Se ha realizado una experiencia consistente en hacer 4.000 extracciones de una urna que contenía las bolas blancas y negras en partes iguales, reponiendo la bola, o sea, con probabilidad para ambos colores de 1/2. Designando por 0 la aparición de blanca y por 1, la negra, los resultados fueron:

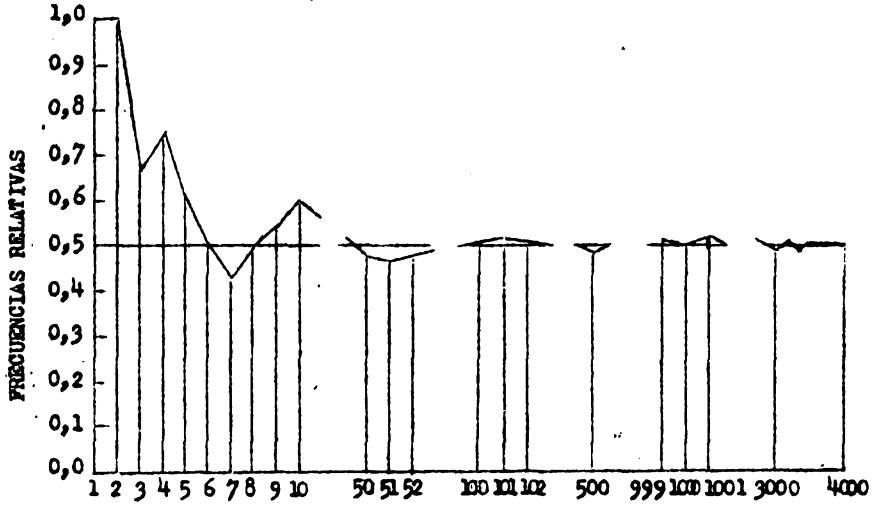
Prueba	Bola obtenida	Total acumulado de bolas negras	Frecuencias relativas de bolas negras
1	1	1	1
2	1	2	1
3	0	2	0,66
4	1	3	0,75
5	0	3	0,60
6	0	3	0,5
7	0	3	0,42
8	1	4	0,5
9	1	5	0,55
10	1	6	0,6
50	1	24	0,48
51	0	24	0,47
100	0	51	0,51
101	1	52	0,515
102	0	52	0,509
500	1	249	0,498

/Cont.

(cont.)

Pruba	Bola obtenida	Total acumulado de bolas negras	Frecuencias relativas de bolas negras
999	0	500	0,51
1.000	0	500	0,5
1.001	1	501	0,501
3.000	0	1.498	0,499
4.000	1	2.001	0,5002

Se puede representar estos resultados en el gráfico siguiente, en el que se ha cortado la escala horizontal para dar cabida a las 4.000 abscisas, y en el cual se observa claramente cómo las frecuencias tienden a estabilizarse alrededor del valor 0,5 que es la probabilidad "a priori".



Ese paso de la teoría a la práctica, por medio de la ley del azar, conduce al método <u>inductivo estadístico</u> o <u>inferencia estadística</u>, peculiar de las ciencias experimentales, pues al establecer una ley descriptiva de determinado fenómeno la única posibilidad es realizar nuevos

/experimentos que

experimentos que la comprueben y ver si puede medirse el grado de confianza de dichos experimentos.

Sustituímos las probabilidades relativas por las frecuencias de sucesos comprobados experimentalmente y si establecemos un valor ε tan pequeño como se quiera, todo suceso cuya probabilidad sea $\geq 1 - \varepsilon$ se considera como seguro con un coeficiente de confianza de $1 - \varepsilon$.

Variable aleatoria.

En los temas 2° y 3° se ha establecido el concepto de variable estadística, entendiendo por tal el conjunto de valores (resultados de observaciones o experiencias) con sus correspondientes frecuencias.

Si se consideran la infinidad de pruebas posibles hechas en las mismas condiciones, el conjunto de los valores resultantes, 0, 1, 2, 3, ..., x, ..., n, ..., cada uno con su probabilidad, constituye una variable aleatoria discreta X.

La suma de todas las probabilidades es la unidad.

Es, pues, la variable aleatoria el modelo teórico del concepto práctico de variable estadística.

El conjunto de pares de valores formados por los de la variable y sus probabilidades respectivas se denomina ley o distribución de probabilidad de la variable aleatoria.

Si se considera la extracción de cartas en la baraja española, la variable aleatoria está constituída por los valores, 1, 2, 3, 4, 5, 6, 7, 10, 11 y 12, y su ley de probabilidad es:

X	1	2	3	4	5	6	7	10	11	12
p	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1

Cuando se consideren fenómenos cuyos resultados numéricos formen una variable contínua, al representar el histograma de frecuencias relativas, hacer el número de observaciones cada vez mayor y la amplitud de los intervalos tendiendo a cero, el histograma tiende a convertirse en una curva, que se denomina de probabilidad.

/Así, pues,

Así, pues, una variable aleatoria continua queda definida por el recorrido y por una función $f(x) \geq 0$, que se denomina de densidad de la probabilidad.

El área total limitada por la curva y el eje de abscisas vale la unidad; el área limitada entre la curva y las ordenadas correspondientes a dos abscisas, a y b, nos dará la probabilidad de obtención de un valor comprendido en el intervalo (a, b).

Esperanza matemática o valor medio teórico y varianza.

Definida para la variable estadística la media aritmética

$$a = \frac{\Sigma x_i n_i}{N} = \sum x_i \frac{n_i}{N} = \Sigma x_i h_i ,$$

en donde h_i son las frecuencias relativas, en virtud de la ley del azar, al pasar del campo práctico al teórico, y sustituir dichas frecuencias relativas por probabilidades, se obtiene:

$$x = \Sigma x_i p_i = \Sigma x_i f(x_i) ,$$

expresión que recibe el nombre de valor medio teórico o esperanza matemática.

Análogamente, recordando la expresión de la varianza:

$$s^2 = \frac{\Sigma (x_i - a)^2 n_i}{N} = \sum (x_i - a)^2 \frac{n_i}{N} = \Sigma (x_i - a)^2 h_i ,$$

y haciendo el mismo paso se obtiene:

$$\sigma^2 = \Sigma (x_i - \alpha)^2 p_i = \Sigma (x_i - \alpha)^2 f(x_i)$$

En las variables aleatorias continuas, las anteriores sumas serían integrales. No nos extendemos en esta dirección por no figurar estos conocimientos en la Enseñanza Media actual.

B. PROMEDIOS Y VARIANZAS

Valores Medios: Su cálculo abreviado

En el tema anterior hemos representado diversos ejemplos de distribuciones de frecuencia, observando que presentan caracteres comunes: a) Pequeñas frecuencias en los valores extremos; b) Acumulación de frecuencias muy altas alrededor de un valor central, próximo o coincidente con el de frecuencia máxima.

Como consecuencia de esto, y para tener un resumen cuantitativo del fenómeno, en vez de manejar todos los datos - que en la mayor parte de los casos sería muy incómodo -, podremos caracterizar la distribución mediante algunos valores numéricos, eligiendo un valor central, alrededor del cual se encuentran distribuídas las frecuencias.

Este proceso de simplificación, que permite representar y comparar las distribuciones, ha sido denominado por R. A. Fisher la reducción de los datos estadísticos

El valor de la variable elegido para representar a una distribución se llama promedio, que no es otra cosa que una medida de posición, un valor "medio", que debe hallarse comprendido entre los valores extremos de la variable.

Un promedio debe reunir las siguientes condiciones, precisadas por Yule.

1) Debe estar perfectamente definido para no dejar nada a la apreciación del observador.

2) Debe estar basado en todas las observaciones hechas, para que cada una tenga su debida influencia.

3) No debe tener un carácter matemático demasiado abstracto.

4) Debe ser de fácil cálculo.

5) Ha de adaptarse con facilidad a los cálculos algebraicos.

Consideraremos los siguientes promedios: Media aritmética, mediana, moda o promedio típico, media geométrica y media armónica.

Media aritmética.

Es el más conocido y usado en la práctica.

/Si consideramos

Si consideramos N valores de la variable $x_1, x_2, \ldots, x_N$ definiremos la media aritmética, que representaremos por a, como el cociente de la suma de los valores de la variable por el número de ellos, es decir

$$a = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Si en vez de considerar frecuencias absolutas consideramos las relativas, la media artimética toma la forma

$$a = x_1h_1 + x_2h_2 + \ldots + x_kh_k ,$$

o bien

$$a = \Sigma x_i h_i$$

en la que

$$h_i = \frac{n_i}{N}$$

En efecto, la media aritmética

$$a = \frac{\Sigma x_i n_i}{N}$$

se puede escribir

$$a = \sum x_i \frac{n_i}{N} = \Sigma x_i h_i$$

Cálculo Abreviado de la Media Aritmética

Podemos calcular la media aritmética de una manera abreviada, razonando en la forma siguiente: Supongamos que transformamos las N observaciones $x_1, x_2, \ldots, x_k$ en nuevas observaciones $x'_i, x'_2, \ldots x'_k$, mediante la siguiente relación: $x_i = 0_x + x'_i$, en la que 0_x es un número fijo llamado origen de trabajo, elegido de modo que las diferencias $x_i - 0_x$ sean más fáciles de manejar que las propias medidas.

/Sustituyendo en

Sustituyendo en la media aritmética, tendremos:

$$a = \frac{\Sigma x_i n_i}{N} = \frac{\Sigma(O_x + x'_i)n_i}{N} = \frac{O_x \ \Sigma n_i}{N} + \frac{\Sigma x'_i n_i}{N}$$

y como

$$\frac{\Sigma n_i}{N} = \frac{N}{N} = 1 \quad y \quad \frac{\Sigma x'_i n_i}{N} = a'$$

(media aritmética de los valores $x'_i = x_i - O_x$), sustituyendo tendremos

$$a = O_x + a'$$

o bien

$$a = O_x + \frac{\Sigma(x_i - O_x)n_i}{N}$$

Apliquemos esta fórmula a la determinación de la media aritmética de la distribución de la tabla 1.

Procederemos en la forma siguiente:

1) Elegiremos un origen auxiliar de trabajo.

2) Dispondremos en columna las diferencias entre cada una de las observaciones y el nuevo origen.

3) Efectuaremos los productos de estas diferencias (desviaciones con respecto al origen auxiliar de trabajo) por las frecuencias respectivas.

4) Calcularemos la suma de estos productos.

5) Dividiremos esa suma por la suma de frecuencias, con lo que habremos hallado la media de los valores x'_i, o sea, a'.

6) Añadiremos a este valor hallado el valor del origen auxiliar de trabajo, con lo cual obtendremos la media aritmética pedida.

El cálculo toma la disposición práctica de la tabla 5, eligiendo como origen de trabajo 37,5.

/Tabla 1

Tabla 1

Varones clasificados según edad al contraer matrimonio en cierta ciudad

Edad en años Intervalos	Varones Frecuencias
De 20 a 25	108
De 25 a 30	406
De 30 a 35	190
De 35 a 40	72
De 40 a 50	39
De 50 a 60	9
TOTAL	824

Tabla 2

Intervalos	Marcas de clase x_i	Frecuencias n_i	Desviación respecto a 37,5 $x'_i = x_i - 0_x$	Productos $x'_i \; n_i$
(1)	(2)	(3)	(4)	(5)
De 20 a 25	25,5	108	- 15	- 1.620
De 25 a 30	27,5	406	- 10	- 4.060
De 30 a 35	32,5	190	- 5	- 950
De 35 a 40	37,5	72	0	0
De 40 a 50	45	39	7,5	292,5
De 50 a 60	55	9	17,5	157,5
TOTAL		824		- 6.180

$$a' = \frac{-6.180}{824} = -7,5; \qquad a = 37,5 - 7,5 ; \quad a = 30 \text{ años}$$

Si la diferencia entre cada dos valores consecutivos de la variable es constante, es decir, si la distribución está formada por intervalos de igual amplitud, se simplifican los cálculos mediante el siguiente razonamiento: Si para medir las desviaciones $x'_i = x_i - 0_x$ tomamos como unidad la amplitud c del intervalo de clase, tendremos que cada una de las desviaciones respecto de este origen serán

$$x''_i = \frac{x_i - 0_x}{c}$$

que se calculan muy fácilmente, ya que bastará poner un cero frente al valor 0_x elegido, y a partir de él ir aumentando una unidad en el sentido creciente de la variable y disminuyéndola en el sentido decreciente.

Las operaciones necesarias para su aplicación son las mismas que en el caso anterior, sin más que, después de hallada la media aritmética de de las $x''_i = \frac{x_i - 0_x}{c}$, que representaremos por a'', multiplicar por la amplitud c del intervalo para pasar nuevamente a unidades de serie, puesto que $x'_i = cx''_i$.

La fórmula, por tanto, será:

$$a = 0_x + a''c$$

o bien

$$a = 0_x + \frac{\sum \left(\frac{x_i - 0_x}{c} \right) n_i}{N} c.$$

Aplicación a la distribución de la tabla 2.

La disposición práctica para el cálculo figura en la tabla 3.

/Tabla 3

Tabla 3

Operaciones

Marcas de clase x_i	Frecuencias n_i	Desviaciones respecto a 1,68 en unidades de clase $x''_i = \frac{x_i - 0_x}{c}$	Productos $x_i n_i = \frac{x_i - 0_x}{c} n_i$
1,58	3	- 5	- 15
1,60	4	- 4	- 16
1,62	6	- 3	- 18
1,64	14	- 2	- 28
1,66	28	- 1	- 28
1,68	49	0	0
1,70	32	1	32
1,72	23	2	46
1,74	8	3	24
1,76	3	4	12
TOTAL	170		9

Elijamos para 0_x = 1,68; tendremos

$$a'' = \frac{9}{170}\ ;\ c = 0{,}02,$$

luego aplicando la fórmula

$$a = 1{,}68 + \frac{9}{170} \times 0{,}02$$

$$a = 1{,}68 + \frac{0{,}18}{170}\ ;\quad a = 1{,}681 \text{ metros.}$$

/Propiedades de

Propiedades de la Media Aritmética.

1a. La suma de las desviaciones de una variable respecto a la media aritmética es nula.

En efecto, si $x_1, x_2, \ldots, x_N$ son N valores de la variable, se verifica

$$\Sigma(x_i - a) = (x_1 - a) + (x_2 - a) + \ldots + (x_N - a) =$$
$$(x_1 + x_2 + \ldots + x_N) - (a + a + \ldots + a)$$

y como

$$\frac{x_1 + x_2 + \ldots + x_N}{N} = a; \quad x_1 + x_2 + \ldots + x_N = Na,$$

sustituyendo en la igualdad anterior $\Sigma(x_i - a) = Na - Na = 0$; luego

$$\Sigma(x_i - a) = 0.$$

Si $x_1, x_2, \ldots x_K$ son los valores de la variable o marcas de clase (si está agrupada en intervalos) la propiedad anterior, se enunciará: La suma de los productos de las desviaciones de una variable respecto a la media aritmética por sus frecuencias respectivas es nula.

Su demostración es análoga a la anterior

$$\Sigma(x_i - a)n_i = \Sigma x_i n_i - a \Sigma n_i,$$

y como $\Sigma n_i = N$ y de $\frac{\Sigma x_i n_i}{N} = a$ se deduce $\Sigma x_i n_i = Na$, sustituyendo en la primera igualdad, resulta $\Sigma(x_i - a)n_i = Na - Na = 0$; luego

$$\Sigma(x_i - a)n_i = 0$$

Como consecuencia de la propiedad anterior resulta que la media aritmética de dichas desviaciones es nula.

2a. La suma de los productos de los cuadrados de las desviaciones por las respectivas frecuencias es un mínimo con respecto a la media aritmética

Esta propiedad se demostrará en el tema siguiente.

/Ventajas e

Ventajas e inconvenientes de la media aritmética

Varias son las ventajas de la media aritmética; entre las más importantes citaremos: a) Es de fácil cálculo; b) Está perfectamente definida; c) Es susceptible de cálculo algebraico, y d) En su determinación intervienen todos los valores de la variable, cumpliendo las condiciones precisadas por Yule.

En el caso en que exista una gran diferencia entre unos datos y otros, la media aritmética tiene el inconveniente de que pierde gran parte de su utilidad al estar afectada excesivamente por las desviaciones extremas respecto al promedio. Por ello es muy conveniente tenerlo en cuenta en algunas de sus aplicaciones.

En estos casos es más conveniente el uso de la mediana.

Mediana.

La mediana es otra medida de posición que puede definirse como un valor de la variable, tal que, supuestos ordenados todos los valores de ésta en orden creciente (o decreciente), la mitad de dichas observaciones tienen un valor inferior o igual a él, y la otra mitad superior o igual a él. La representaremos por Me.

Para su obtención, una vez ordenados los valores de la variable, consideraremos dos casos: que el número de observaciones N sea impar o par. En el primer caso, si se verifica $N = 2p + 1$, la mediana será el que ocupe el lugar central, es decir, el lugar $p + 1$. En el segundo caso, $N = 2p$, la mediana es indeterminada, ya que todo valor comprendido entre los que ocupan los lugares p y $p + 1$, responden a la definición de mediana. En la práctica se toma como mediana la media aritmética de los valores que ocupen los lugares p y $p + 1$.

Si tenemos 11 individuos colocados por orden creciente de estaturas, la talla del que ocupa el lugar 6 será la mediana.

Sean las tallas 1,57 - 1,59 - 1,60 - 1,62 - 1,65 - 1,68 - 1,71 - 1,73 - 1,75 - 1,77 - 1,78.

La talla 1,68 es la que corresponde a la mediana, ya que la mitad de las observaciones son superiores a 1,68 y la otra mitad, inferiores.

/Si el

Si el número de individuos fuese 10, y sus tallas,

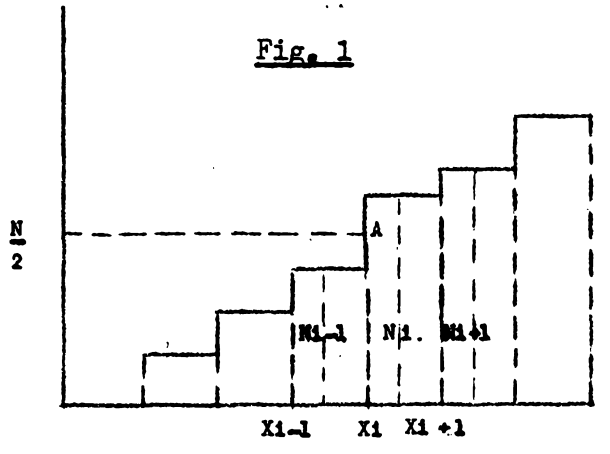
1,57 - 1,59 - 1,60 - 1,62 - 1,65 - 1,68 - 1,73 - 1,75 - 1,77 - 1,78

cualquier valor comprendido entre 1,65 y 1,68 podría tomarse para mediana. En la práctica tomaremos la media aritmética de estas tallas, es decir, de las correspondientes a los lugares quinto y sexto.

$$\text{Mediana} = \frac{1{,}65 + 1{,}68}{2} \; ; \; \text{Me} = 1{,}665$$

Para hallar la mediana en el caso de que se trate de valores agrupados, construiremos el diagrama acumulativo figura 1, y por el extremo de la observación $\frac{N}{2}$ trazaremos la paralela al eje de las abscisas, pudiendo ocurrir dos casos: a) Que esta recta corte al diagrama en un punto como el A, es decir, que esté comprendido entre las ordenadas de dos valores que la variable x_{i-1} y x_i; en este caso la mediana

Fig. 1



queda perfectamente determinada, por tanto Me = x_i , siendo x_i un valor tal que se verifique

$$N_{i-1} < \frac{N}{2} < N_i$$

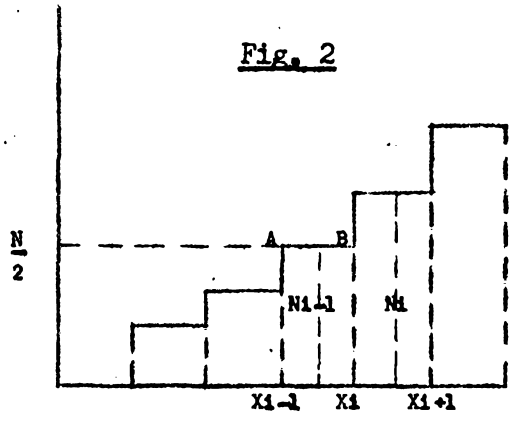
/b) Si

b) Si la recta paralela al eje de abscisas, trazada por el extremos de la ordenada correspondiente a la observación $\frac{N}{2}$, tiene un segmento AB de puntos comunes con el diagrama figura 2, la mediana aparece indeterminada, puesto que cualquier valor de la variable comprendido entre x_{i-1} y x_i tiene por ordenada $\frac{N}{2}$; en este caso suele tomarse la media aritmética de estos valores, luego

$$Me = \frac{x_{i-1} + x_i}{2}$$

siempre que $N_{i-1} = \frac{N}{2} < N_i$.

Fig. 2



Numéricamente, para hallar la mediana formaremos la distribución acumulativa. Si ninguna frecuencia acumulativa coincide con $\frac{N}{2}$ la mediana es el valor de la variable que corresponde a la primera frecuencia

/acumulativa, que

acumulativa, que supera a $\frac{N}{2}$.

La Tabla 4 representa una distribución de frecuencia, donde se ha añadido una tercera columna para las frecuencias acumuladas. Al ser

$$\frac{N}{2} = \frac{80}{2} = 40$$

la primera frecuencia acumulativa que supera a 40, es 50, que corresponde al valor 5 de la variable; luego Me = 5.

En el caso en que exista una frecuencia acumulativa que coincida con $\frac{N}{2}$ la mediana se obtendrá como la media aritmética del valor de la variable que corresponde a dicha frecuencia acumulativa, y al de la siguiente.

Sea la distribución de la tabla 5.

Como $\frac{N}{2} = \frac{70}{2} = 35$ coincide con la frecuencia acumulativa correspondiente al valor 4 de la variable, el valor de la mediana se obtendrá hallando la media aritmética de los valores 4 y 5 de la variable, es decir

$$Me = \frac{4 + 5}{2}; \quad Me = 4,5$$

Tabla 4

Notas x_i (1)	Número de alumnos n_i (2)	Frecuencias acumuladas N_i (3)
0	2	2
1	1	3
2	2	5

/Tabla 4 (cont.)

(cont).

Notas x_i (1)	Número de alumnos n_i (2)	Frecuencias acumuladas N_i (3)
3	10	15
4	20	35
5	15	50 -
6	12	62
7	9	71
8	6	77
9	2	79
10	1	80
TOTAL	80	

Tabla 5

Notas x_i	Número de alumnos n_i	Frecuencias acumuladas N_i
0	2	2
1	1	3
2	2	5
3	10	15
4	20	35
5	12	47
6	8	55
7	7	62
8	5	67
9	2	69
10	1	70
TOTAL	70	

/Para hallar

Para hallar la mediana de una distribución de valores agrupados en intervalos procederemos en la forma siguiente: Formaremos el diagrama acumulativo, donde suponemos que la frecuencia absoluta crece linealmente del punto L_{i-1} al L_i. Supongamos que el intervalo

(L_{i-1}, L_i) es tal, que $N_{i-1} < \frac{N}{2} < N_i$; en este intervalo estará la mediana. Su valor será el correspondiente a la abscisa del punto B (figura 3).

Por tanto, $Me = OB' = OA' + A'B'$; $Me = L_{i-1} + A'B'$. [1]

En los triangulos semejantes ACD y ABE se verifica

$$\frac{AD}{AE} = \frac{CD}{BE}$$

siendo

$AD = c_i$ = amplitud del intervalo,

$AE = A'B'$,

$CD = n_i$ = frecuencia absoluta del intervalo (L_{i-1}, L_i).

$BE = BB' - EB' = \frac{N}{2} - N_{i-1}$.

Sustituyendo valores

$$\frac{c_i}{A'B'} = \frac{n_i}{\frac{N}{2} - Ni - 1} \quad \text{de donde} \quad A'B' = \frac{\frac{N}{2} - Ni - 1}{n_i} \cdot c_i.$$

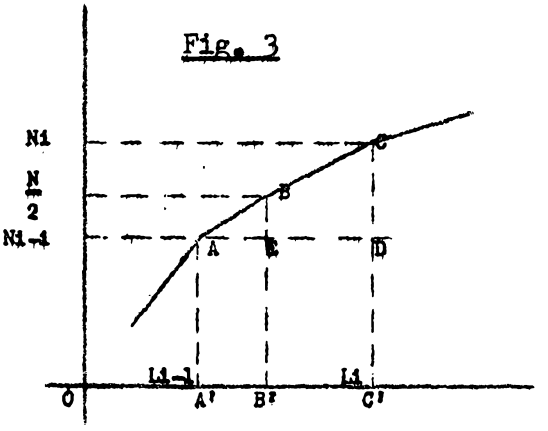
valor que, sustituído en [1], da

$$Me = Li - 1 + \frac{\frac{N}{2} - Ni - 1}{n_i} \cdot c_i.$$

/Determinar la

Determinar la mediana de la distribución de frecuencias de la Tabla 6.

Fig. 3



El proceso seguido es el siguiente; 1° Si N es el número total de observaciones, hallaremos $\frac{N}{2}$ para localizar el intervalo en que se encuentra la mediana, obteniendo así L_{i-1}. 2° Determinar la diferencia entre $\frac{N}{2}$ y las frecuencias acumuladas N_{i-1} hasta la clase inferior a la mediana y dividir esa diferencia por la frecuencia absoluta n_i, que corresponde a la clase mediana. 3° Multiplicar el cociente por la amplitud c_i del intervalo; y 4° Sumar el resultado al límite inferior del intervalo, localizado en el punto 1°.

/Tabla 6

Tabla 6

Intervalos	Frecuencias n_i	Frecuencias acumuladas N_i
1,57 - 1,59	3	3
1,59 - 1,61	4	7
1,61 - 1,63	6	13
1,63 - 1,65	14	27
1,65 - 1,67	28	55
1,67 - 1,69	49	104 -
1,69 - 1,71	32	136
1,71 - 1,73	23	159
1,73 - 1,75	8	167
1,75 - 1,77	3	170
TOTAL	170	

Teniendo en cuenta que siendo N = 170; $\frac{N}{2}$ = 85; la primera frecuencia acumulada superior a 85 es 104, luego en el intervalo 1,67 - 1,69 se encuentra la mediana, por tanto, L_{i-1} = 1,67; N_{i-1} = 55; n_i = 49; c_i = 0,02. Aplicando la fórmula

$$Me = 1,67 + \frac{85 - 55}{49} \times 0,02;$$

$$Me = 1,67 + \frac{0,60}{49}; \quad Me = 1,682 \text{ m.}$$

Para calcular la mediana no es necesario que los intervalos tengan una amplitud constante, ya que solamente inteviene en su cálculo un solo intervalo.

/Ventajas e

Ventajas e inconvenientes de la mediana. Las ventajas principales son: 1º Su facilidad de cálculo. 2º Que en ella no influyen los valores extremos de la variable; y 3º Que aunque la distribución sea irregular, es una buena medida de la tendencia central.

Tiene el inconveniente de que no se adapta al cálculo algebraico.

Promedio típico (Moda).

La moda se define como el valor de la variable, al que corresponde mayor frecuencia. La representaremos por Md.

Si n_i es la frecuencia correspondiente a la moda, se ha de verificar $n_i + 1 < n_i > n_i - 1$ también podemos decir que la moda es el valor de la variable al que corresponde un máximo relativo.

En una serie de salarios es el salario más generalizado, en una serie de precios es el precio más corriente, etc.

Si tenemos las siguientes notas de un grupo de alumnos 5 - 3 - 4 - 7 - 5 - 2 - 6 - 5. La moda será la nota 5.

$$Md = 5$$

Si existen varios máximos relativos, al mayor de todos llamaremos moda absoluta.

Las distribuciones pueden ser unimodales, bimodales o multimodales, según que exista una, dos o varias modas.

En el ejemplo anterior no hay mas que una moda.

Si consideramos los salarios de un grupo de obreros, 25 - 50 - 62 - 50 - 25 - 80 - 25 - 50 - 15, tendremos dos modas: una 25 y otra 50.

Si consideramos la distribución de la tabla 4, la moda la obtendremos tomando el valor de la variable al que corresponde mayor frecuencia. Como la mayor frecuencia es 20, que corresponde al valor 4 de la variable, tendremos

$$Md = 4$$

Por último, para hallar la moda de una distribución de valores agrupados en intervalos, determinaremos cuál es el intervalo de frecuencia máxima. Esta sería una primera determinación del promedio tipo. Pero dentro de este intervalo ¿cuál es el valor que corresponde a la ordenada

/máxima? Por

máxima? Por métodos que se explicarán más adelante se podría ajustar una curva a la distribución observada, y una vez ajustada - tendría una expresión analítica - sería fácil la determinación de la moda, ya que sería la abscisa correspondiente a la ordenada máxima.

Siempre que la distribución de los valores de la variable sea sensiblemente simétrica, se puede obtener un valor aproximado de la moda mediante la fórmula

$$Md = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \times c_i ;$$

en la que L_{i-1} es el límite inferior de la clase en cuyo intervalo está comprendida la moda; n_{i-1}, la frecuencia de la clase anterior; n_{i+1} la frecuencia de la clase posterior, y c_i, la amplitud del intervalo.

Si consideramos la distribución de la tabla 6, la máxima frecuencia 49 corresponde al intervalo 1,67 - 1,69, luego $L_{i-1} = 1,67$; $n_{i-1} = 28$; $n_{i+1} = 32$, y $c = 0,02$; por tanto,

$$Md = 1,67 + \frac{32}{28 + 32} \times 0,02; \quad Md = 1,679 \text{ m.}$$

Si los intervalos no son de amplitud constante, habrá que hallar los cocientes $k_i = \frac{n_i}{c_i}$, siendo n_i la frecuencia en el intervalo i, y c_i la amplitud de dicho intervalo.

El intervalo que tenga mayor k_i contendrá la moda, y se tomará la misma fórmula anterior, sustituyendo las n_i por las k_i. Así, la fórmula será:

$$Md = L_{i-1} + \frac{k_{i+1}}{k_{i-1} + k_{i+1}} \cdot c_i$$

Igual que para la mediana, el cálculo se basa en los intervalos y no en las marcas de clase.

/En la

En la determinación de la moda puede ocurrir que una misma distribución dé valores distintos si se varía la amplitud de los intervalos.

Supongamos, por ejemplo, las dos tablas 7 y 8 expresión del mismo hecho, pero con distintas agrupaciones de clase. En la primera el intervalo modal es 15 - 17, y en la segunda, 11 - 15. Las dos distribuciones comprenden el mismo número de casos, y únicamente hemos variado la forma de agrupación de las observaciones.

Tabla 7

Intervalos	Frecuencias
7 - 9	3
9 - 11	5
11 - 13	15
13 - 15	32
15 - 17	35
17 - 19	10

Tabla 8

Intervalos	Frecuencias
7 - 11	8
11 - 15	47
15 - 19	45

Ventajas e inconvenientes de la moda. La moda se empleará cuando los valores de la variable presenten una gran concentración hacia un valor determinado. Sólo se utilizará para las distribuciones de gran frecuencia total, es decir, de un gran número de observaciones para poder comprobar que existe efectivamente una tendencia hacia un valor.

La moda no es susceptible de cálculo algebraico, y en su determinación no intervienen todos los datos.

La media aritmética, la mediana y la moda, por ser valores correspondientes a la variable, representan puntos del eje de las abscisas.

Relación aproximada entre estos promedios.

Si la distribución es simétrica, la media aritmética, la mediana y la moda coinciden. Si es asimétrica, son diferentes, y si la asimetría es moderada existe una relación aproximada entre ellas:

$$a - Md = 3(a - Me)$$

/Esta fórmula

Esta fórmula permite calcular uno de los promedios, conocidos los otros dos; pero bien entendido que sólo tendrá validez para distribuciones ligeramente asimétricas.

Media Geométrica

Es menos utilizada que las anteriores. Tiene gran utilidad para las medidas relativas, principalmente para los problemas sobre números índices.

En el caso de N observaciones individuales, $x_1, x_2, \ldots, x_N$ se define la media geométrica como la raíz N-sima del producto de los valores de la variable. Se representa por M_o. La fórmula será:

$$M_o = \sqrt[N]{x_1 \cdot x_2 \cdots x_N} \qquad [1]$$

Si $x_1, x_2, \ldots, x_k$ son los valores de la variable, y $n_1, n_2, \ldots, n_k$, sus frecuencias respectivas, y se verifica $n_1 + n_2 + \ldots + n_k = N$, se define la media geométrica como la raíz N-zima del producto de los valores de la variable elevados a sus frecuencias respectivas.

$$M_o = \sqrt[N]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}} \qquad [2]$$

Su cálculo se hace posible mediante los logaritmos.

Tomando logaritmos en la [1], tendremos

$$\log M_o = \frac{\log x_1 + \log x_2 + \ldots + \log x_N}{N}$$

Obteniendo la siguiente propiedad: El logaritmo de la media geométrica es la media aritmética de los logaritmos de la variable.

Por esta razón este promedio recibe también el nombre de media aritmética logarítmica.

Tomando logaritmos en la 2

$$\log M_o = \frac{n_1 \log x_1 + n_2 \log x_2 + \ldots n_k \log x_k}{N}$$

/Ejemplos: Hallar

Ejemplos: Hallar la media geométrica de las siete primeras potencias de 2.

$$M_o = \sqrt[7]{2 \cdot 2^2 \cdot 2^3 \cdots 2^7} = \sqrt[7]{2^{1+2+3+\cdots+7}} = \sqrt[7]{2^{\left(\frac{1+7}{2}\right)7}} = 2^{\frac{1+7}{2}} = 2^4$$

$$M_o = 16$$

puesto que $1 + 2 + \ldots + 7$ = suma de los términos de una progresión aritmética de razón 1.

De igual manera si se trata de la distribución de la table 9.

Tabla 9

x_i	n_i
15	3
20	2
540	1

Aplicando la fórmula

$$\sqrt[3]{15^3 \cdot 20^2 \cdot 540} = 30.$$

Propiedades de la media geométrica.

El producto de los cocientes de cada valor de la variable por la media geométrica es igual a la unidad.

En efecto

$$\frac{x_1}{M_o} \cdot \frac{x_2}{M_o} \cdot \ldots \cdot \frac{x_N}{M_o} = \frac{x_1 \cdot x_2 \cdot \ldots \cdot x_N}{M_o^N}$$

y como

$$M_o = \sqrt[N]{x_1 x_2 \cdots x_N} \; ; \quad M_o^N = x_1 x_2 \cdots x_N,$$

sustituyendo en la primera

$$\frac{x_1}{M_o} \cdot \frac{x_2}{M_o} \cdot \ldots \cdot \frac{x_N}{M_o} = \frac{M_o^N}{M_o^N} = 1.$$

El cuadrado de la media geométrica es la media geométrica del cuadrado de la variable.

Se tiene $M_o^2 = \sqrt[N]{(x_1 x_2 \cdots x_N)^2} = \sqrt[N]{x_1^2 \cdot x_2^2 \cdot \ldots \cdot x_N^2} =$

/media geométrica

media geométrica de x_1^2.

Ventaja e inconvenientes de la media geométrica.

Su principal ventaja es que reduce la influencia de los valores extremos de la variable. Es un promedio perfectamente definido y puede someterse al cálculo algebraico.

El mayor inconveniente es su cálculo complicado; la gran influencia que ejercen los números pequeños y que se anula cuando hay un valor de la variable igual a cero.

Media armónica.

Es menos utilizada que los promedios anteriores, y se define en una serie de N valores: $x_1, x_2, \ldots, x_N$, como el recíproco de la media aritmética de los recíprocos de sus valores. Se representa por M - 1

Su valor se obtendrá de la forma siguiente: Como $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_N}$ son los recíprocos de $x_1, x_2, \ldots, x_N$; su media aritmética será

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}{N} \quad \text{luego} \quad M - 1 = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}$$

o bien

$$M - 1 = \frac{N}{\Sigma \frac{1}{x_i}}$$

Si $x_1, x_2, \ldots, x_k$ son los valores de una variable, y $n_1, n_2, \ldots, n_k$, sus correspondientes frecuencias, se tendrá

/M - 1

$$m - 1 = \frac{N}{\frac{1}{x_1} n_1 + \frac{1}{x_2} n_2 + \ldots + \frac{1}{x_k} n_k}$$

que puede escribirse

$$\frac{1}{M-1} = \frac{\frac{1}{x_1} n_1 + \frac{1}{x_2} n_2 + \ldots + \frac{1}{x_k} n_k}{N} \quad \text{o bien} \quad \frac{1}{M-1} = \frac{\Sigma \frac{1}{x_i} n_i}{N}$$

que nos dice que la inversa de la media armónica es igual a la media aritmética de los inversos de la variable.

Es aconsejable para promediar velocidades y tiempos.

Se adapta al cálculo algebraico.

El nombre de media armónica proviene de la aplicación que hizo Pitágoras a los acordes musicales.

Hallar la media armónica de los números 6, 12 y 24.

Sus recíprocos son:

$$\frac{1}{6} \; \frac{1}{12} \; \frac{1}{24} \; ; \quad \frac{1}{6} + \frac{1}{12} + \frac{1}{24} = \frac{4 + 2 + 1}{24} = \frac{7}{24}$$

Aplicando la fórmula

$$M_{-1} = \frac{3}{\frac{7}{24}} = \frac{72}{7} \; ; \qquad M_{-1} = 10{,}28$$

<u>Cuartiles, deciles, percentiles</u>. Son medidas de posición.

Así como la mediana la definíamos diciendo que es el valor de la variable, elegido de tal forma que si los valores están colocados en orden creciente, la mitad de las observaciones son inferiores o iguales a él, y la otra mitad son superiores o iguales al mismo, definiremos el primer cuartil, que representaremos por Q_1, como <u>el valor de la variable, elegido de tal forma, que la cuarta parte de las observaciones sean menores o iguales a él y las tres cuartas partes restantes son superiores o iguales al mismo</u>.

El segundo cuartil será la mediana.

/El tercer

El tercer cuartil, que representaremos por Q_3, es <u>el valor de la variable correspondiente a la observación que deja tres cuartas partes de las observaciones menores que él y el cuarto restantes superiores a él.</u>

El cálculo es análogo al de la mediana, sustituyendo $\frac{N}{2}$ por $\frac{N}{4}$ o por $\frac{3N}{4}$.

Así, tendremos las fórmulas

$$Q_1 = Li-1 + \frac{\frac{N}{4} - Ni-1}{n_i}.c_i \quad ; \quad Q_3 = L_{i-1} + \frac{\frac{3N}{4} - Ni-1}{n_i}.c_i$$

Sea la distribución de la tabla 6. Vemos que

$$\frac{N}{4} = \frac{170}{4} = 42,5.$$

La primera frecuencia que supera a 42,5 es 55, que corresponde al intervalo 1,65 - 1,67, aplicando la fórmula, siendo $L_{i-1} = 1,65$; $Ni - 1 = 27$; $n_i = 28$; $c_i = 0,02$.

$$Q_1 = 1,65 + \frac{42,5 - 27}{28} \times 0,02 \; ; \quad Q_1 = 1,661 \text{ m.}$$

De igual forma obtendremos: Q_3. Por ser $\frac{3N}{4} = 127,5$; la primera frecuencia superior a 127,5 es 136, que corresponde al intervalo 169 - 1,71, luego $L_{i-1} = 1,69$; $N_{i-1} = 104$; $n_i = 32$; $c_i = 0,02$; luego

$$Q_3 = 1,69 + \frac{127,5 - 104}{32}.0,02 \; ; \quad Q_3 = 1,705 \text{ m.}$$

/Análogamente, se

Análogamente, se definen los deciles y los percentiles. El primer decil sería el valor de la variable correspondiente a la observación que tuviera delante una décima parte del número de observaciones y detrás las nueve décimas restantes. Se definirían de igual modo los demás deciles. El cálculo del decil de orden K (K no puede exceder de 9) se hará con arreglo a la fórmula

$$D_k = L_{i-1} + \frac{\frac{KN}{10} - N_{i-1}}{n_i} \cdot c_i$$

Los deciles dividen a la serie en diez partes, cada una de las cuales contiene una décima parte de las observaciones.

El quinto decil es la mediana.

El K-ésimo percentil, representado por P_k, es el valor de la variable, tal que K % de las observaciones le son inferiores (K no puede exceder de 99). La fórmula es la misma que la de la mediana.

$$P_k = L_{i-1} + \frac{\frac{KN}{100} - N_{i-1}}{n_i} \cdot c_i$$

Como resumen de todo lo anterior, vamos a hallar la media artimética, la mediana, la moda, el primer cuartil, el tercer cuartil y algún decil, por ejemplo, el séptimo y el percentil doce.

Con objeto de ver la simplificación que el método abreviado para la obtención de la media aritmética proporciona, la hallaremos por el procedimiento ordinario y por el abreviado, aplicándolo a la tabla 10, en la que aparecen 1.000 individuos clasificados por sus pesos.

/Tabla 10

Tabla 10

Intervalos	Marcas de clase x_i	Frecuencias n_i	Productos $x_i n_i$	Frecuencias acumuladas N_i	Desviación con respuesta a 58,5 $x'_i = \frac{x'_i - O_x}{c_i}$	Productos $x'_i n_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
45 a 48	46,5	4	186	4	- 4	- 16
48 a 51	49,5	20	990	24	- 3	- 60
51 a 54	52,5	70	3.675	94	- 2	- 140
54 a 57	55,5	180	9.990	274	- 1	- 180
57 a 60	58,5	428	25.038	702	0	0
60 a 63	61,5	204	12.546	906	1	204
63 a 66	64,5	64	4.128	970	2	128
66 a 69	67,5	24	1.620	994	3	72
69 a 72	70,5	6	423	1.000	4	24
	TOTAL	1.000	58.596			32

$$a = \frac{58.596}{1.000} = 58,596 \text{ Kg.}$$

$O_x = 58,5$ $\quad a'' = \frac{32}{1.000} = 0,032$;

$a''c = 0,032 \times 3 = 0,096$

$a = 58,5 + 0,096 = 58,596$ Kg. $\quad Me = \frac{N}{2} = 500$;

la primera frecuencia ≥ 500 es 702; intervalo de la mediana, 57 - 60;
luego $L_{i-1} = 57$; $N_{i-1} = 274$; $n_i = 428$; $c_i = 3$;

/Me =

$$Me = 57 + \frac{500 - 274}{428} \times 3; \quad Me = 57 + 1{,}584; \quad Me = 58{,}584 \text{ Kg.}$$

Moda

Máxima frecuencia, 428; intervalo, 57 - 60; luego $L_{i-1} = 57$;

$n_{i+1} = 204$; $n_{i-1} = 180$; $c_i = 3$.

$$Md = 57 + \frac{204}{180 + 204}, 3; \quad Md = 57 + 1{,}593;$$

$Md = 58{,}593$ Kg.

Medidas de dispersión. Varianza. Desviación típica. Cálculo abreviado de ambas

Dispersión

Para el estudio o manejo del conjunto de valores que toma una variable estadística, o correspondiente a una serie estadística de observaciones, es preciso reducirlo a unos pocos valores representativos que la mente pueda retener fácilmente. Esta reducción, necesaria ya para una serie única, es indispensable para la comparación de dos o más series entre sí.

Entre esos valores están los promedios, estudiados en el tema anterior, que representan la tendencia central o dominante de los valores de la variable. Pero esta característica es insuficiente para describir abreviadamente una serie de valores. Supongamos, por ejemplo dos grupos de empleados cuyos sueldos en pesetas se distribuyen según las tablas estadísticas números 11 y 12,y que se calculan las medias aritméticas respectivas como representativas del sueldo medio en cada grupo:

Tabla 11

Sueldos x_i	Número empleados n_i	$x_i n_i$
7.200	2	14.400
8.000	2	16.000
10.000	2	20.000
12.000	2	24.000
12.800	2	25.600
	$10 = N'$	$100.000 = \Sigma x_i n_i$

$$a = \frac{\Sigma x_i n_i}{N} = 10.000 \text{ ptas.}$$

/Tabla 12

Tabla 12

Sueldos x_i	Número empleados n_i	$x_i n_i$
3.000	1	3.000
4.000	1	12.000
19.000	3	45.000
40.000	5	40.000
	10 = N	100.000 = $= \Sigma x_i n_i$

$$a = \frac{\Sigma x_i n_i}{10} = 10.000 \text{ ptas.}$$

En ambos grupos la media aritmética es 10.000 pesetas, pero es más representativa del conjunto de sueldos del primer grupo que del segundo, porque los sueldos están más cercanos a la media, es decir, fluctúan menos alrededor de las 10.000 pesetas.

Esa fluctuación o dispersión de los diversos valores alrededor del valor central es otra de las características principales de la serie de valores de una variable estadística. Su medición se realiza por las llamadas medidas de dispersión.

Noción contraria a la dispersión es la de concentración de los valores alrededor del central o promedio.

Medidas de Disperción

Entre las medidas de dispersión se estudian primero las expresadas en las mismas unidades que la variable. Las consideradas en este tema son: recorrido, intervalo intercuartílico, desviación media y desviación típica.

/1. Recorrido

1. Recorrido.

Se denomina también intervalo de variación, y es la diferencia entre los dos valores extremos de la variable. Designándole por l,

l = valor máxima de x - valor mínimo de x.

Ejemplo: Hallar el recorrido en la siguiente serie de datos:

Precios por mayor del Qm. de garbanzos en 1956.

Meses	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Sept.	Octubre	Nov.	Dic.
Pesetas	635	683	666	702	714	725	716	722	837	849	848	847

El mayor precio es 849 y el menor 635, por tanto, el recorrido es:

$$l = 849 - 635 = 214 \text{ pesetas.}$$

El recorrido no es muy usado, pues basado solamente en los dos valores extremos, es poco adecuado para caracterizar la dispersión. Así, si en un grupo de 100 hombres existe uno muy bajo, de 1,50 m. de altura, y un gigante de 2,20 m., mientras los 98 restantes tienen sus tallas comprendidas entre 1,60 m. y 1,75 m., el recorrido es 2,20 - 1,50 = 0,70 m. En cambio, se reduce a 1,75 - 1,60 = 0,15 m. si los dos primeros individuos fueran de estatura corriente.

No obstante, por la simplicidad de su cálculo, el recorrido ha sido muy usado por los anglosajones en las aplicaciones del método estadístico al control de la fabricación en serie.

2. Intervalo intercuartílico.

Para impedir el exceso de influencia de los valores extremos, se ha propuesto considerar, como medida de la dispersión, la diferencia entre los cuartiles tercero y primero: $Q_3 - Q_1$. De esta forma se deja fuera un mismo porcentaje de valores del lado de cada cuartil (un 25 por 100).

En el anterior ejemplo de los sueldos de empleados, los respectivos intervalos intercuatílicos son:

En el primer grupo, $Q_1 = 8.000$ $Q_3 = 12.000$ $Q_3 - Q_1 = 4.000$ pts.

/En el

En el segundo grupo, $Q_1 = 4.000$ $Q_3 = 9.000$ $Q_3 - Q_1 = 5.000$ pts.

En el ejemplo del tema anterior (Tabla 6)

$$Q_1 = 1,661 \quad Q_3 = 1,705 \quad Q_3 - Q_1 = 0,044 \text{ m.}$$

A veces se considera el semi-intercuartil, o sea, la mitad:

$$\frac{Q_3 - Q_1}{2}.$$

Aunque su cálculo es sencillo, el intervalo intercuartílico tiene el inconveniente de no basarse en todas las observaciones.

3. Desviación media.

Es la media aritmética de los valores absolutos de las desviaciones de los valores de la variable respecto al promedio elegido.

Teóricamente debiera elegirse la mediana como promedio, pues, según una propiedad no demostrada aquí, la suma de los valores absolutos de las desviaciones es mínima en tal caso. Sin embargo, en la práctica es la media aritmética el promedio o valor central a partir del cual se miden las desviaciones. Designando por D_m la desviación media, la fórmula es

$$D_m = \frac{1}{N} \sum |x_i - a|,$$

si los valores no están dados en distribución de frecuencias; y es

$$D_m = \frac{\sum |x_i - a| n_i}{\sum n_i}$$

cuando sí lo están.

Ejemplo: Las entradas de butaca para siete salas de espectáculos valen, respectivamente, 7, 11, 14, 25, 30, 45 y 50 pesetas. Hallar la desviación media.

$$a = \frac{\sum x_i}{N} = \frac{7 + 11 + 14 + 25 + 30 + 45 + 50}{7} = \frac{182}{7} = 26 \text{ pts.}$$

/D_m

$$D_m = \frac{1}{N} \quad x_i - a =$$

$$= \frac{|9 - 26| + |11 - 26| + |14 - 26| + |25 - 26| + |30 - 26| + |45 - 26| + |50 - 26|}{7} =$$

$$= \frac{19 + 15 + 12 + 1 + 4 + 19 + 24}{7} = \frac{94}{7} = 13{,}43 \text{ ptas.}$$

Ejemplo: Hallar la desviación media del número de pares de calzado por 100 consumidores, según la tabla 13.

Tabla 13

N° pares x_i	Número consumidores n_i	$x_i n_i$	$\lvert x - a_i \rvert$	Operaciones $\lvert x_i - a \rvert \; n_i$
1	8	8	1,76	14,08
2	40	80	0,76	30,40
3	32	96	0,24	7,68
4	10	40	1,24	12,40
5	8	40	2,24	17,92
6	2	12	3,24	6,48
	$100 = \sum x_i = N$	$276 = \sum x_i n_i$		$88{,}96 = \sum x_i - a_i \; n_i$

$$a = \frac{\sum x_i n_i}{\sum x_i} = \frac{276}{100} = 2{,}76 \qquad D_m \; \frac{\sum |x_i - a| \; n_i}{\sum x_i} = \frac{88{,}96}{100} = 0{,}89 \text{ pares}$$

En este ejemplo se emplea la segunda fórmula, por venir los datos en distribución de frecuencias.

Si bien la desviación media no es de cálculo difícil, y se basa en todos los datos, tiene el inconveniente de emplear valores absolutos, lo que impide su fácil utilización en los cálculos algebraicos.

/Otra forma

Otra forma de tener en cuenta el valor absoluto de las desviaciones a la media aritmética es considerar sus cuadrados, lo que conduce al concepto de varianza y desviación típica.

4. Varianza y desviación típica.

La varianza de una variable es la media de los cuadrados de las desviaciones respecto de su media. Designándola por s^2, la fórmula es

$$s^2 = \frac{\Sigma (x_i - a)^2}{N}$$

La desviación típica es la raíz cuadrada de la varianza:

$$s = \sqrt{\frac{\Sigma (x_i - a)^2}{N}}$$

y se suele denominar también desviación cuadrática media, siendo igual, por otra parte, a la media cuadrática de las desviaciones a la media. Es innecesario el anglicismo "standard" y peor aún castellanizarlo.

El cálculo de la desviación típica se reduce al de la varianza. Y antes de dar procedimientos para calcular ésta, es conveniente demostrar la siguiente propiedad de la media aritmética:

"La suma de los cuadrados de las desviaciones es mínima cuando se consideran a partir de la media aritmética".

Sea la variable x y un valor cualquiera x_o. Se verifica

$$x_i - x_o = x_i - a + a - x_o ,$$

y elevando al cuadrado

$$(x_i - x_o)^2 = (x_i - a)^2 + (a - x_o)^2 + 2(a - x_o)(x_t - a)$$

Sumando para los N valores, x_t resulta

$$\Sigma (x_i - x_o)^2 = \Sigma (x_i - a)^2 + N(a - x_o)^2 + 2(a - x_o)\ \Sigma(x_t - a)$$

donde el último término es nulo por serlo $\sum (x_t - a)$, según se demostró en la propiedades de la media aritmética. Luego

$$\Sigma(x_t - a)^2 = \Sigma (x_i - x_o)^2 - N(a - x_o)^2,$$

/y como

y como $N(a - x_o)^2$ es positivo, queda demostrada la propiedad enunciada, razón que justifica considerar la media como origen de las desviaciones.

La fórmula empleada en el cálculo de la varianza adquiere formas distintas según se presenten los valores de la variable.

a) Los datos no están preparados en forma de distribución de frecuencias.

La fórmula a emplear se obtiene así:

$$s^2 = \frac{\Sigma (x_i - a)^2}{N} = \frac{\Sigma(x_i^2 - 2ax_i + a^2)}{N} = \frac{\Sigma x_i^2}{N} - 2a\frac{\Sigma x_i}{N} + \frac{\Sigma a^2}{N} =$$

$$= \frac{\Sigma x_i^2}{N} - 2\frac{\Sigma x_i}{N}\frac{\Sigma x_i}{N} + \frac{Na^2}{N} = \frac{\Sigma x_i^2}{N} - 2\left(\frac{\Sigma x_i}{N}\right)^2 + \left(\frac{\Sigma x_i}{N}\right)^2 =$$

$$= \frac{\Sigma x_i^2}{N} - \left(\frac{\Sigma x_i}{N}\right)^2$$

En la práctica, los cálculos se realizan, como se hace en el ejemplo siguiente, por este orden:

1º. Se forma una columna (II) con los cuadrados de los valores x_i de la variable.

2º. Se obtienen los totales de las columnas (I) y (II), que son, respectivamente, Σx_i y Σx_i^2.

3º. Se aplica la fórmula obtenida.

Ejemplo: Hallar la desviación típica de la serie de precios en pesetas de seis libros de texto correspondientes a un curso de estudios (tabla 14).

/Tabla 14

Tabla 14

Precios x_i (I)	x^2_i (II)
80	6.400
75	5.625
62	3.844
60	3.600
92	8.464
86	7.396
455 = Σx_i	35.329 = Σx^2_i

Varianza:

$$s^2 = \frac{\Sigma x_i^2}{N} - \left(\frac{\Sigma x_i}{N}\right)^2 = \frac{35.329}{6} - \left(\frac{455}{6}\right)^2 = 5.888,17 - (75,83)^2 = 137,98$$

Desviación típica:

$$s = \sqrt{137,98} = 11,7 \text{ ptas.}$$

b) Los datos están preparados en forma de distribución de frecuencias y los valores de la variable son números sencillos o, si no lo son, se dispone de máquina de calcular.

Como en este caso a cada valor x_i de la variable corresponde una frecuencia n_t, la fórmula de la varianza es

$$s^2 = \frac{\Sigma (x_i - a)^2 n_i}{N}; \text{ donde } N = \Sigma n_t.$$

/Desarrollando como

Desarrollando como en el caso anterior, se obtiene:

$$s^2 = \frac{\Sigma x_i^2 n_i}{N} - \left(\frac{\Sigma x_i n_i}{N} \right)^2$$

En la práctica, los cálculos se realizan, como en los ejemplos siguientes, por este orden:

1º. Se forma una columna (III) con los productos $x_i n_i$.

2º. Se forma otra columna (IV) multiplicando los valores de la (II) por los correspondientes de la (III), con lo que se obtienen los productos $x_i^2 n_i$.

3º. Se obtienen los totales de las columnas (II), (III) y (IV), que son respectivamente, N, $\Sigma x_i n_i$ y $\Sigma x_i^2 n_i$.

4º. Se aplica la fórmula obtenida.

Ejemplo: En una investigación botánica se precisa hallar la desviación típica de la distribución de 1.000 cápsulas de adormidera, clasificadas por el número de pistilos (tabla 15).

Tabla 15

		Operaciones	
Pistilos x_i (I)	Cápsulas n_i (II)	$x_i n_i$ (III) = (II).(I)	$x_i^2 n_i$ (IV) = (I).(III)
8	26	208	1.664
9	55	495	4.455
10	79	790	7.900
11	122	1.342	14.762
12	158	1.896	22.752
13	167	2.171	28.223
14	156	2.184	30.576

/cont.

(cont.)

Pistilos x_i (I)	Cápsulas n_i (II)	$x_i n_i$ (III) = (II).(I)	$x_i^2 n_i$ (IV) = (I).(III)
15	122	1.830	27.450
16	67	1.072	17.152
17	28	476	8.092
18	20	360	6.480
TOTALES	1.000 = N	12.824 = $= \Sigma x_i n_i$	169.506 = $= \Sigma x^2_i n_i$

$$s^2 = \frac{\Sigma x_i^2 n_i}{N} - \left(\frac{\Sigma x_i n_i}{N}\right)^2 = \frac{169.506}{1.000} - \left(\frac{12.824}{1.000}\right)^2 =$$

$$= 169{,}506 - 12{,}824^2 = 5{,}051.$$

$$s = \sqrt{5{,}051} = 2{,}25 \text{ pistilos.}$$

Si los valores de la variable estuviesen agrupados en intervalos, puede aplicarse la misma fórmula, considerando las marcas de clase como x_i.

Ejemplo: Hallar la desviación típica de la serie de altitudes, en metros, de la capitales de provincia cuya distribución figura en la Tabla 16.

Tabla 16

Altitud en metros Intervalos	Marcas de clase x_i (I)	Número de capitales n_i (II)	Operaciones $x_i n_i$ (III) = (I).(II)	Operaciones $x^2_i n_i$ (IV) = (I).(III)
0 a 200	100	25	2.500	250.000
200 a 400	300	2	600	180.000
400 a 600	500	7	3.500	1.750.000

/(cont.)

(cont.)

Altitud en metros Intervalos	Marcas de clase x_i (I)	Número de capitales n_i (II)	$x_i n_i$ (III) = (I).(II)	$x^2_i n_i$ (IV) = (I).(III)
600 a 800	700	9	6.300	4.410.000
800 a 1.000	900	4	3.600	3.240.000
1.000 a 1.200	1.100	3	3.300	3.630.000
TOTALES		50 = N	19.800 = $= \Sigma x_i n_i$	13.460.000 = $= \Sigma x^2_i n_i$

$$s^2 = \frac{13.460.000}{50} - \left(\frac{19.800}{50}\right)^2 = 269.200 - (396)^2 = 112.384$$

$$s = \sqrt{112.384} = 335 \text{ m.}$$

c) <u>En la distribución de frecuencias los valores de la variable no son números sencillos y se diferencian cada dos consecutivos en una unidad, o bien están agrupados en intervalos de diferente amplitud.</u>

En este caso es conveniente, para abreviar el cálculo, tomar un origen auxiliar de trabajo, 0_x, arbitrario; y, como en el cálculo de la media aritmética, designando por x'_i los valores de la variable referidos al nuevo origen, se verifica:

$$x'_i = x_i - 0_x = x_i - a + a - 0_x$$

$$\frac{1}{N} \Sigma x'^2_i n_i = \frac{1}{N} \Sigma (x_i - a + a - 0_x)^2 n_i =$$

$$= \frac{1}{N} \Sigma (x_i - a)^2 n_i + 2(a - 0_x) \frac{1}{N} \Sigma (x_i - a) n_i + \frac{1}{N} \Sigma (a - 0_x)^2 n_i =$$

$$/= \frac{1}{N}$$

$$= \frac{1}{N} \Sigma (x_i - a)^2 n_i + (a - 0_x)^2 \frac{1}{N} \Sigma n_i = s^2 + (a - 0_x)^2$$

pues el doble producto es cero por serlo $\Sigma (x_t - a)n_i$, según la propiedad de la media aritmética "la suma de las desviaciones a la media es nula".

Además, al estudiar la media, se vió que

$$a = 0_x + \frac{1}{N} \Sigma (x_i - 0_x)n_i ,$$

luego, sustituyendo este valor de a en la igualdad obtenida

$$\frac{1}{N} \Sigma x'_i{}^2 n_i = s^2 + (a - 0_x)^2$$

se deduce

$$\frac{1}{N} = \Sigma x'_i{}^2 n_i = s^2 + \left[\frac{1}{N} \Sigma (x_i - 0_x)n_i \right]^2$$

o sea

$$s^2 = \frac{\Sigma (x_i - 0_x)^2 n_i}{N} - \left(\frac{\Sigma (x_i - 0_x)n_i}{N} \right)^2$$

En la práctica los cálculos se realizan, como en el ejemplo siguiente, por este orden:

1°. Se elige un valor conveniente como origen auxiliar.

2°. Se forma una columna (IV) con las diferencias $x_i - 0_x$.

3°. Se forma otra columna (V) con los productos $(x_i - 0_x)n_t$, que se obtienen multiplicando los valores correspondientes de las columnas (III) y (IV).

4°. Se forma otra columna (VI) con los productos $(x_i - 0_x)n_t$, que se obtienen multiplicando los valores correspondientes de la columnas (IV) y (V).

5°. Se obtienen los totales de las columnas (III), (V) y (VI), que son, respectivamente, N, $\Sigma (x_i - 0_x)n_t$ y $\Sigma (x_i - 0_x)^2 n_i$.

/6°. Se

6°. Se aplica la fórmula obtenida.

Ejemplo: Hallar la desviación típica en la distribución de mujeres, por grupos de edad, que se casaron durante el año 1951 en España, tal como se registra en la Tabla 17.

Tabla 17

			Operaciones		
Años de edad Intervalos	Marcas de clase	Miles de mujeres	$(x_i - 0_x)$	$(x_i - 0_x)n_i$	$(x_i - 0_x)^2 n_i$
(I)	x_i (II)	n_i (III)	(IV)=(II)-37,5	(V)=(III) . (IV)	(VI)=(IV) . (V)
20-25	22,5	89	- 15	- 1.335	20.025
25-30	27,5	71	- 10	- 710	7.100
30-35	32,5	19	- 5	- 95	475
35-40	37,5	8	0	0	0
40-50	45	5	7,5	37,5	281,25
50-60	55	1	17,5	17,5	306,25
	TOTALES	193 = N		- 2.085 = $\Sigma (x_i - 0_x) n_i$	28.187,5 = $\Sigma (x_i - 0_x)^2 n_i$

$$0_x = 37,5$$

$$s^2 = \frac{28.187,5}{193} - \left(\frac{-2.085}{193}\right)^2 = 29,41$$

$$s = \sqrt{29,41} = 5,4 \text{ años.}$$

d) <u>En la distribución de frecuencias los valores de la variable están agrupados en intervalos de igual amplitud</u>.

Denominando por c a la amplitud común de los intervalos, los valores de la variable referidos al origen auxiliar 0_x y medidos en unidades iguales a c, son:

$$/x_i - 0_x$$

$$\frac{x_i - 0_x}{c} = \frac{x'_i}{c} = x''_i \; ; \quad x'_i = cx''_i,$$

y sustituyendo esta expresión de x'_i en la fórmula anterior de la varianza

$$s^2 = \frac{1}{N} \Sigma c^2 x''_i{}^2 n_i - \left(\frac{1}{N} \Sigma c x''_i n_i\right)^2 = c^2 \frac{\Sigma x''_i{}^2 n_i}{N} - c^2 \left(\frac{\Sigma x''_i n_i}{N}\right)^2 =$$

$$= c^2 \left[\frac{\Sigma \left(\frac{x_i - 0_x}{c}\right)^2 n_i}{N} - \left(\frac{\Sigma \left(\frac{x_i - 0_x}{c}\right) n_i}{N} \right)^2 \right]$$

En la práctica, la aplicación de esta fórmula se realiza, como en el ejemplo siguiente, por este orden:

1°. Se elige un origen auxiliar de trabajo 0_x.

2°. Se forma una columna (IV) con los valores $\frac{x_i - 0_x}{c}$, que se obtienen con facilidad, porque correspondiendo cero frente a 0_x, basta aumentar de unidad en unidad en el sentido creciente de la variable y disminuir en el sentido decreciente.

3°. Se forma una columna (V) con los productos $\left(\frac{x_i - 0_x}{c}\right) n_i$. que se obtienen multiplicando los valores correspondientes de la columnas (III) y (IV).

4°. Se forma otra columna (VI) con los productos $\left(\frac{x_i - 0_x}{c}\right)^2 n_i$ que se obtienen multiplicando los valores correspondientes de las columnas (IV) y (V).

5°. Se obtienen los totales de la columnas (III), (V) y (VI), que son respectivamente N, $\Sigma \frac{x_i - 0_x}{c} n_i$ y $\Sigma \left(\frac{x_i - 0_x}{c}\right)^2 n_i$.

/6°. Se

6°. Se aplica la fórmula obtenida.

Ejemplo: Durante varios años fueron examinados de matemáticas, en una Universidad, 570 alumnos en total. En los diversos ejercicios, que debían sufrir, las calificaciones podían sumar de 0 a 99 puntos, sin existir la fracción de punto. Las calificaciones resultantes se agruparon en 20 clases, con amplitud de intervalo de 5 puntos, como figuran en la tabla 18. Las marcas de clase y las frecuencias aparecen en las columnas (II) y (III). Se pide calcular la desviación típica.

Para hallar la varianza, de acuerdo con el orden señalado, se ha elegido un origen auxiliar $0_x = 52$, por ser una marca de clase de lugar central; luego se han hallado los valores $\frac{x_i - 0_x}{c}$, teniendo en cuenta que $c = 5$ en este caso. Así, el primer valor de la columna (IV) es $\frac{2 - 52}{8} = -10$. A continuación se realizan los pasos tercero, cuarto y quinto, aplicándose después la fórmula de s^2.

Tabla 18

Puntos Intervalos	Marcas de clase x_i	Nº de calificaciones n_i	$\frac{x_i - 0_x}{c}$	$\frac{x_i - 0_x}{c} n_i$	$\left(\frac{x_i - 0_x}{c}\right)^2 n_i$
(I)	(II)	(III)	(IV)	(V)	(VI)
0 a 4	2	12	- 10	- 120	1.200
5 a 9	7	13	- 9	- 117	1.053
10 a 14	12	13	- 8	- 104	832
15 a 19	17	14	- 7	- 98	686
20 a 24	22	23	- 6	- 138	828
25 a 29	27	23	- 5	- 115	575
30 a 34	32	29	- 4	- 116	464
35 a 39	37	34	- 3	- 102	306
40 a 44	42	44	- 2	- 88	176

/(cont.)

(cont.)

(I)	(II)	(III)	(IV)	(V)	(VI)
45 a 49	47	44	- 1	- 44	44
50 a 54	52	50	0	- 1.042	
55 a 59	57	52	1	52	52
60 a 64	62	61	2	122	244
65 a 69	67	41	3	123	369
70 a 74	72	32	4	128	512
75 a 79	77	27	5	135	675
80 a 84	82	23	6	138	828
85 a 89	87	17	7	119	833
90 a 94	92	13	8	104	832
95 a 99	97	5	9	45	405
				966	
	TOTALES	570		- 76	10.914

$$0_x = 52 \; ; \quad c = 5$$

$$s^2 = 25 \left[\frac{10.914}{570} - \left(\frac{-76}{570} \right)^2 \right] = \left[19{,}15 - 0{,}13^2 \right] = 478{,}25$$

$$s = \sqrt{478{,}25} = 21{,}8 \text{ puntos.}$$

El lugar de la columna (V) correspondiente al origen elegido se suele aprovechar para colocar la suma de las desviaciones negativas (-1.042), así como en la parte alta de los totales se coloca la suma de las positivas (966), sumando luego ambas sumas debajo (-1.042 + 966 = 76).

/La desviación

La desviación típica cumple bastante bien las llamadas condiciones de Yule, descritas en el tema anterior; y en especial éstas dos: se calcula con todos los valores u observaciones y se expresa por una fórmula algebraica sencilla.

La desviación típica y la media aritmética son los dos parámetros estadísticos más usados y que mejor caracterizan una serie, en general. Además, facilitan la comparación entre varias series y entre la distribución de un conjunto y de una parte de él. En la inmensa mayoría de los casos, la media aritmética es el promedio empleado para conocer la tendencia central y la varianza o la desviación típica para medir la dispersión. El llamado "Análisis de la varianza" es, posiblemente, la parte más fundamental de la ciencia estadística, base de la técnica de la investigación moderna.

Ejemplo de cálculo de la desviación típica en una distribución teórica.

Hallar la media y la desviación típica de la distribución en que los valores de x son los números naturales 1, 2, ..., N, todos ellos con frecuencia igual a la unidad.

$$a = \frac{\Sigma x_i}{N} = \frac{1 + 2 + \ldots + N}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2}$$

$$s^2 = \frac{\Sigma x_i^2}{N} - a^2 = \frac{1^2 + 2^2 + \ldots + N^2}{N} - \left(\frac{N+1}{2}\right)^2 =$$

$$= \frac{N(N+1)(2N+1)}{6N} - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12}$$

$$s = \sqrt{\frac{N^2-1}{12}}$$

Medidas abstractas de dispersión.

Todas la medidas de dispersión anteriormente definidas se expresan en unidades de la misma especie que la variable: si la variable es una longitud, la medida calculada es una longitud. En los ejemplos aquí

/expuestos las

expuestos las diversas medidas han sido: un número de pesetas, de pares de calzado, de pistilos, de metros, de años, de puntos de calificación. Esto es un inconveniente grave al comparar la dispersión de dos o más series. Puede ocurrir que dos variables de la misma especie se expresen en diferentes unidades, tal como ocurre al comparar las tallas de un grupo de españoles (que vendrá expresada en centímetros) y las de un grupo de británicos (que vendrá expresada en pulgadas). O que, aun expresadas en la misma unidad, los niveles medios de las series sean muy diferentes, como ocurrirá con las tallas de un grupo de niños de seis años y las de un grupo de adultos. Por otra parte, la significación de la magnitud de la medida de dispersión dependerá de la de los valores de la variable: una desviación típica de 15 centímetros en las medidas de la altura de un edificio es mucho menos significativa que esa misma desviación en las medidas de la altura de un hombre.

Para eliminar esos inconvenientes se calculan los <u>coeficientes de dispersión</u>, obtenidos por cociente entre alguna de las medidas de dispersión y un valor promedio expresado en las mismas unidades. Así resulta un valor abstracto, por lo cual tales coeficientes se denominan medidas <u>abstractas</u>. También se llaman medidas <u>absolutas</u> en el sentido de que son independientes de la unidad empleada. Algunos tratadistas emplean la denominación de <u>coeficientes de dispersión relativa</u>, llamando por contrario, absolutas a las medidas que se expresan en unidades específicas.

Por ser independientes de unidad de expresión, estos coeficientes permiten comparar la variabilidad de distribuciones de distinta especie; por ejemplo, una de altura con otra de pesos.

<u>Coeficientes de variación</u>.

Es el coeficiente de dispersión más usado y se define por la fórmula

$$d = \frac{s}{a}$$

o sea, el cociente de la desviación típica y la media aritmética.

Se suele emplear multiplicado por 100:

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

/y entonces

y entonces es la desviación típica expresada en porcentaje de la media aritmética.

El valor respectivo del coeficiente de variación en cada ejemplo anterior es:

a) Tabla 14:

$$a = \frac{455}{7} = 75{,}83 \text{ ptas.} \; ; \quad s = 11{,}7 \text{ ptas.} \; ; \quad d = \frac{11{,}7}{75{,}83} = 0{,}15.$$

b) Tabla 15:

$$a = \frac{12.824}{1.000} = 12{,}824 \text{ pistilos}; \quad s = 2{,}25 \text{ pistilos}; \quad d = \frac{2{,}25}{12{,}824} = 0{,}18$$

Tabla 16:

$$a = \frac{19.800}{50} = 396 \text{ m.}; \quad s = 335 \text{ m.}; \quad d = \frac{335}{396} = 0{,}85.$$

c) Tabla 17:

$$a = \frac{-\ 2.085}{193} + 37{,}5 = 26{,}7 \text{ años}; \quad s = 5{,}4 \text{ años}; \quad d = \frac{5{,}4}{26{,}7} = 0{,}20.$$

d) Tabla 18:

$$a = 0{,}133.5 \div 52 = 51{,}3 \text{ puntos}; \quad s = 21{,}8 \text{ puntos}; \quad d = \frac{21{,}8}{51{,}3} = 0{,}42.$$

/C. Principales

C. PRINCIPALES DISTRIBUCIONES DE PROBABILIDAD: BINOMIAL Y NORMAL

Distribuciones de probabilidad

En el cálculo de probabilidades interesa conocer la probabilidad de obtener cierto valor de una variable aleatoria X. Cuando ésta sólo puede tomar valores enteros, 0, 1, 2, 3, ..., x, ... n, cada uno de ellos tiene cierta probabilidad que, en general, se desconoce. Hay, sin embargo, casos teóricos en los que dicha probabilidad, que se designa por f (x), se puede conocer, y al conjunto de estos valores de f (x) es al que se llama distribución de probabilidad de la variable X.

Las características del fenómeno teórico a considerar son las que determinan la forma o tipo de dichas distribuciones, y de éstas las más importantes, consideradas ya como clásicas en los estudios estadísticos, son tres: la distribución binomial o de Bernouilli, la distribución normal o de Laplace y Gauss y la distribución de Poisson.

En la práctica se suelen tener distribuciones de frecuencias, más teniendo en cuenta que la probabilidad de un suceso es, según la ley del azar, el valor alrededor del cual se estabiliza su frecuencia relativa, al aumentar el número de pruebas, una distribución de frecuencias relativas será análoga a una de probabilidad, y comparándola con los tipos teóricos conocidos de éstas, se puede ver a cuál se aproxima más y deducir de ello las características del fenómeno observado o bien por su desemejanza con dichas distribuciones rechazar hipótesis previamente formuladas, caso también muy frecuente en el análisis estadístico.

La distribución binomial

Esta distribución fue estudiada primeremente por Bernouilli y corresponde al caso denominado también de las pruebas repetidas con probabilidad constante.

Sea un suceso A, cuya presentación en un determinado experimento tiene una probabilidad p. La del suceso contrario, B, será, naturalmente 1 - p, que se designa por q, de manera que p + q = 1. Estas probabilidades

/p y q

p y q son constantemente las mismas en cada experimento o prueba que se efectúe. Realizadas n pruebas, el suceso A podrá haberse presentado cierto número de veces, variable de 0 a n, siendo este número de veces la variable aleatoria X, que puede tomar los valores 0, 1, 2, ..., x, ... n, cuyas probabilidades f (x) se quiere determinar.

Ejemplo: Sea una urna que contiene cinco bolas idénticas en todo, salvo el color, pues dos son blancas y tres son negras. Al sacar una bola de la urna, la probabilidad de que sea blanca será:

$$p = \frac{2}{5} = 0,4,$$

y la de que sea negra:

$$q = \frac{3}{5} = 0,6$$

Reintegrando la bola a la urna y realizando en igual forma sucesivas extracciones, estas probabilidades p y q serán las mismas en cada prueba.

Si se repite la prueba dos veces, n = 2, designando por A bola blanca y B bola negra, los casos posibles que se podrán presentar serán: AA, AB, BA y BB, esto es, dos bolas blancas, la primera blanca y la segunda negra, la primera negra y la segunda blanca y las dos negras. Las probabilidades de cada uno de estos casos, determinado cada uno de ellos por la aparición de dos sucesos independientes, son, en virtud del teorema de la probabilidad compuesta, los productos de las probabilidades respectivas de cada suceso, y serán, por tanto, pp, pq, qp y qq, esto es, los términos del producto $(p + q)(p + q)$.

Si se repite la prueba tres veces, n = 3, los casos posibles serán AAA, AAB, ABA, ABB, BAA, BAB, BBA y BBB, o sea, tres bolas blancas, dos blancas y una negra y dos negras y una blanca en distintos órdenes, y tres negras, siendo las respectivas probabilidades:

ppp, ppq, pqp, pqq, qpp, qpq, qqp, qqq,

que son los términos del producto $(p + q)(p + q)(p + q)$.

/Siendo la

Siendo la variable X el número de veces que se obtiene bola blanca, en el caso de dos pruebas los valores x pueden ser 0, 1 y 2, y sus probabilidades la suma de las de los casos posibles en que se presenten 0, 1 y 2 veces la letra p, o sea, q^2, $2pq$ y p^2, esto es, los términos del desarrollo de $(q + p)^2$. En el caso de $n = 3$ pruebas, los valores x podrán ser 0, 1, 2 y 3, y sus probabilidades respectivas q^3, $3q^2p$, $3qp^2$ y p^3, esto es, los términos del desarrollo de $(q + p)^3$; análogamente, en el caso de n pruebas, la probabilidad de obtener x bolas blancas y, en consecuencia $n - x$ bolas negras, será la suma del número de términos del producto $(p + q)(p + q) \ldots^n (p + q)$, en que figure p, x veces, esto es, el término de grado x en p, del desarrollo de $(q + p)^n$. Por tanto, se tiene:

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

en la que $\binom{n}{x}$ es el número combinatorio $C_n^x = \frac{n!}{x!(n-x)!}$

El hecho de que los distintos valores de f (x) sean los términos de la potencia enésima del binomio $(q + p)$ es al que se debe el nombre de distribución binomial.

Con el ejemplo de la urna considerada, en que $p = 0,4$ $q = 0,6$, los distintos valores de f (x) para diversos valores de n, vienen expresados en la siguiente tabla:

/Valores de

Valores de x	Valores de f (x)				
	n = 1	n = 2	n = 3	n = 4	n = 10
0	0,6	0,36	0,216	0,1296	0,0060
1	0,4	0,48	0,432	0,3456	0,0403
2	-	0,16	0,288	0,3456	0,1209
3	-	-	0,064	0,1536	0,2150
4	-	-	-	0,0256	0,2508
5	-	-	-	-	0,2007
6	-	-	-	-	0,1115
7	-	-	-	-	0,0425
8	-	-	-	-	0,0106
9	-	-	-	-	0,0016
10	-	-	-	-	0,0001

Dividiendo f (x + 1) por f (x), se tiene:

$$\frac{f(x+1)}{f(x)} = \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^{x} q^{n-x}} = \frac{\binom{n}{x+1} p}{\binom{n}{x} q} = \frac{n-x}{x+1} \frac{p}{q}$$

de donde

$$f(x+1) = \frac{n-x}{n+1} \cdot \frac{p}{q} \cdot f(x);$$

fórmula que nos permite calcular cada valor f (x + 1) deduciéndolo del anterior.

Moda en la distribución binomial

También se deduce la expresión anterior, que será

$$f(x+1) > f(x)$$

/si (n - x)

$$\text{si} \quad (n - x)p > (n + 1)q$$
$$\text{o} \quad np > xq + q + xp$$
$$\text{o bien} \quad np - q > x$$
$$\text{y si} \quad f(x + 1) < f(x)$$
$$\text{es} \quad np - q < x$$

de donde resulta que, al crecer f(x) cuando $x < np - q$, y decrecer si $x > np - q$, el valor modal x para el que f(x) es máximo, es el menor entero mayor que $np - q$. Pero si $np - q = x$ es $f(x + 1) = f(x)$, siendo entonces máximos e iguales estos valores, y x y $x + 1$ los valores modales. Así ocurre en el ejemplo anterior, cuando $n = 4$.

$$np - q = 4.0{,}4 - 0{,}6 = 1.$$

Para $x = 1$ y $x = 2$ la probabilidad es igual y máxima.

Media aritmética en la distribución binomial

Esta media es

$$\alpha = \frac{\sum x_i f(x_i)}{\sum f(x_i)}$$

siendo $f(x_i)$ las probabilidades o frecuencias relativas, cuya suma es necesariamente la unidad y, por tanto, $\alpha = \sum x f(x)$, esto es:

$$\alpha = 0.q^n + 1\binom{n}{1} pq^{n-1} + 2\binom{n}{2} p^2 q^{n-2} + \ldots + x\binom{n}{x} p^x q^{n-x} + \ldots + n\binom{n}{n} p^n$$

Quitando el primer término, igual a 0, y sacando np factor común de los demás, queda para cada uno:

$$x\binom{n}{x} p^x q^{n-x} = np. \frac{x.(n-1)!}{x!\,(n-x)!} p^{x-1} q^{n-x} = np \frac{(n-1)!}{(x-1)!\,(n-x)!} p^{x-1} q^{n-x} =$$

$$= np \binom{n-1}{x-1} p(x-1)q(n-x)$$

y haciendo la suma

$$\alpha = np\left[q^{n-1} + \binom{n-1}{1} pq^{n-2} + \ldots + \binom{n-1}{x-1} p^{x-1} q^{n-x} + \ldots + \binom{n-1}{n-1} p^{n-1} \right] =$$

$$= np, \left[(q + p)^{n-1} \right] = np$$

/puesto que

puesto que $q + p = 1$.

Es decir, el valor medio de la variable aleatoria X, en la distribución binomial, es $\alpha = np$. Si np es entero, el valor medio coincide con la moda.

Varianza en la distribución binomial

La varianza tenía por fórmula

$$s^2 = \frac{\sum x_i^2 n_i}{\sum n_i} - a^2$$

En este caso será $\sigma^2 = \sum x^2 f(x) - \alpha^2$, y como $a = np$,

$$\sigma = \sum x^2 f(x) - n^2 p^2.$$

El primer término del segundo miembro es:

$$0^2 q^n + 1^2 \binom{n}{1} pq^{n-1} + 2^2 \binom{n}{2} p^2 q^{n-2} + \ldots + x^2 \binom{n}{x} p^x q^{n-x} + \ldots + n^2 \binom{n}{n} p^n$$

El término general de esta suma, puesto que se vio antes que

$$x \binom{n}{x} p^x q^{n-x} = np \binom{n-1}{x-1} p^{x-1} q^{n-x},$$

y poniendo $x = 1 + (x - 1)$, se puede ver que es

$$x^2 \binom{n}{x} p^x q^{n-x} = xnp \binom{n-1}{x-1} p^{x-1} q^{n-x} = \left[1 + (x-1)\right] np \binom{n-1}{x-1} p^{x-1} q^{n-x} =$$

$$= np \left[\binom{n-1}{x-1} p^{x-1} q^{n-x} + (x-1) \binom{n-1}{x-1} p^{x-1} q^{n-x} \right]$$

Efectuando la suma de todos los términos, como

$$\sum \binom{n-1}{x-1} p^{x-1} q^{n-x} = (p+q)^{n-1} = 1$$

y

$$\sum (x-1) \binom{n-1}{x-1} p^{x-1} q^{n-x}$$

/será el

será el valor medio en el caso de n - 1 pruebas, o sea, (n - 1)p, se tendrá:

$$\sum x^2 f(x) = np\left[1 + (n-1)p\right] = np + n^2p^2 - np^2$$

y

$$\sigma^2 = \sum x^2 f(x) - n^2p^2 = np - np^2 = np\,(1-p) = npq.$$

Asi hemos llegado a la expresión de la varianza en la distribución binomial, que es

$$\sigma^2 = npq.$$

La desviación típica será, por consiguiente,

$$\sigma = \sqrt{npq}$$

Si arrojamos una moneda 100 veces al aire, siendo la probabilidad de que salga cara igual que la de que salga cruz, $p = q = 0{,}5$ y $n = 100$, el valor medio del número de caras que se obtenga será $np = 50$; la varianza, $npq = 25$, y la desviación típica, $\sigma = 5$.

Representación gráfica

Formando el polígono de frecuencias de la distribución binomial obtenemos la representación gráfica de ésta, que, en el caso de ser $p = q$, tiene que ser evidentemente simétrica. Si es $p \neq q$, ya no lo es, pero se puede ver que tiende rápidamente a serlo si se aumenta el número n de pruebas.

Distribución normal. Ajuste de una distribucion normal a una distribucion muestral de frecuencia 1/

La distribución normal fue considerada por primera vez por De Moivre en 1753, pero fue pronto relegada al olvido, hasta que, a principios del siglo XIX, Gauss y Laplace la pusieron de actualidad, y por ello en la literatura estadística se la conoce también con el nombre de distribución de Laplace-Gauss.

1/ En la Enseñanza Media no se estudian integrales. El conocimiento por parte del alumno de ligeras nociones de Cálculo integral nos hubiera permitido desarrollar el tema con más rigor matemático.

/El nombre

El nombre de normal tiene solamente carácter histórico, ya que se creyó que, en la práctica, la mayoría de las distribuciones eran de este tipo normal, y las restantes anormales; lo de "normal" es sólo un nombre, y hoy día esta distribución es tan corriente como otra cualquiera de las que se utilizan en Estadística.

En el tema 3º nos referimos a un tipo de curvas de frecuencia que adoptaban forma de campana o forma encorvada unimodal. Casi todos los rasgos biológicos, como la talla y el peso, cuando se miden y anotan en gran número, presentan este tipo de curva.

Se ha discutido mucho por qué tantos caracteres biológicos, cuando se miden y ordenan, su representación gráfica adopta la forma de una curva de campana. Una teoría asegura que la probabilidad de muerte en los extremos de la escala de valores de la variable es mayor y menor en el centro. Algunas pruebas parecen apoyar esta hipótesis. El coeficiente de mortalidad entre los hombres muy pequeños o muy altos es mucho mayor que en el hombre de estatura media. Sin embargo, los pesos de los niños recién nacidos también adoptan forma de campana, dando así lugar a la hipótesis de que la herencia es la fuerza que determina esta clase de forma.

También hay que hacer notar la acción de las influencias del medio. El psicólogo La Piere encontró que de cien conductores de automóvil que llegaron a un importante cruce de calles en un distrito residencial sin señales de tráfico ni guardia urbano, 1 paró por completo, 21 disminuyeron la velocidad considerablemente, 65 aminoraron algo, 12 no hicieron nada y 1 aumentó la velocidad. El comportamiento de estos conductores, que se adapta a una curva en forma de campana, puede difícilmente atribuirse a factores genéticos. Luego el medio se comporta del mismo modo que la herencia.

En la historia de la Estadística, la distribución normal tiene una gran importancia, ya que en principio se creyó que la mayoría de las distribuciones eran de este tipo; según Yule "la distribución normal fue para los primeros estadísticos lo que el círculo para los astrónomos de la época de Ptolomeo".

/Aunque las

Aunque las primitivas exageraciones sobre la importancia de la distribución normal, hoy día no se pueden mantener, es evidente que desempeña un importante papel en la teoría y en la práctica estadística.

A continuación veremos algunos conceptos que utilizaremos a lo largo del tema.

Curva de probabilidad

En la Tabla 19 aparecen 200 individuos clasificados por su estatura. En la columna 3 se han escrito las frecuencias relativas. Los 200 individuos constituyen una muestra obtenida de un conjunto más amplio denominado universo o población, que puede abarcar a todos los habitantes de un país. En la figura 1 representamos el histograma de frecuencias relativas correspondiente a la variable estadística x.

Tabla 19

(1) Intervalos	(2) Frecuencias absolutas n_i	(3) Frecuencias relativas	(4) Frecuencias relativas acumuladas
1,57 - 1,59	1	1/200	1/200
1,59 - 1,61	2	2/200	3/200
1,61 - 1,63	7	7/200	10/200
1,63 - 1,65	19	19/200	29/200
1,65 - 1,67	44	44/200	73/200
1,67 - 1,69	50	50/200	123/200
1,69 - 1,71	42	42/200	165/200
1,71 - 1,73	25	25/200	190/200
1,73 - 1,75	8	8/200	198/200
1,75 - 1,77	1	1/200	199/200
1,77 - 1,79	1	1/200	200/200
Totales	200	1	

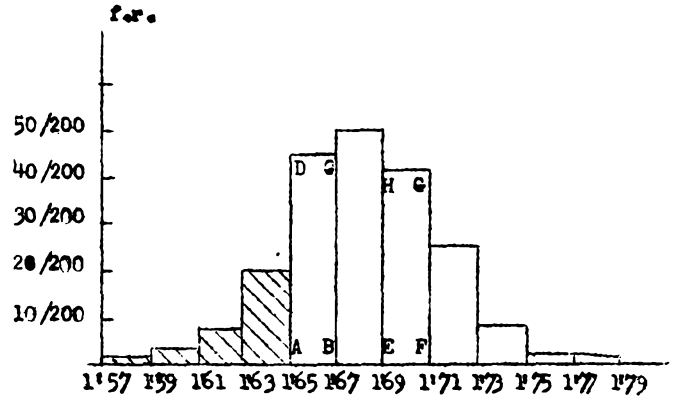
/El área

El área total encerrada por el histograma es igual a la unidad, ya que el área debe ser igual a la suma de las frecuencias relativas.

Por otra parte, observamos las siguientes relaciones:

Area ABCD = frecuencia relativa o proporción de individuos que tienen una estatura comprendida entre 1,65 y 1,67.

Figura 1



Si designamos abreviadamente la frecuencia relativa por f.r, la anterior expresión podemos escribirla así:

$$\text{Area ABCD} = f.r(1{,}65 < x \leq 1{,}67)$$

De forma análoga:

$$\text{Area rayada} = f.r(x \leq 1{,}65)$$

$$\text{Area ABCD} + \text{Area EFGH} =$$

$$= f.r(1{,}65 < x \leq 1{,}67) + f.r\ (1{,}69 < x \leq 1{,}71)$$

Si aumentamos indefinidamente el número de elementos de la muestra y hacemos tender a cero la amplitud de los intervalos, el contorno del histograma adoptará la forma de una curva que se denomina curva de probabilidad. (Fig. 2)

/Idealizando, en

Idealizando, en virtud de la Ley del azar, al aumentar el número de observaciones las frecuencias relativas tienden a estabilizarse alrededor de la probabilidad, el histograma de frecuencias relativas se convierte en la curva de probabilidad y la variable estadística continua en una variable aleatoria continua, que designaremos por X.

Las anteriores relaciones para las frecuencias relativas quedan ahora establecidas así:

El área total encerrada por la curva de probabilidad y el eje X es igual a la unidad.

Area ABCD = Probabilidad de que un individuo tenga una estatura comprendida entre 1,65 y 1,67.

Abreviadamente, la anterior expresión podemos escribirla así:

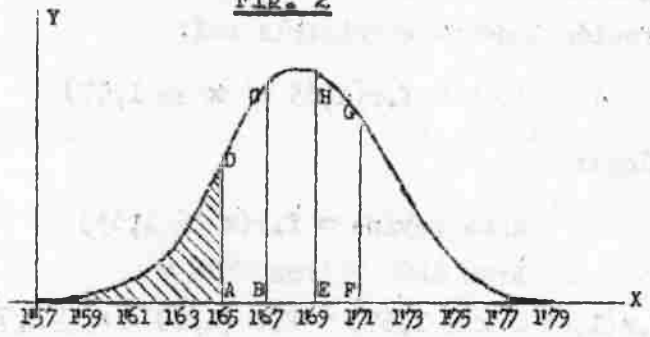
$$\text{Area ABCD} = \text{Prob}\ (1{,}65 < X \leq 1{,}67)$$

De forma análoga:

$$\text{Area rayada} = \text{Prob}\ (X \leq 1{,}65)$$

$$\text{Area ABCD} + \text{Area EFGH} = \text{Prob}\ (1{,}65 < X \leq 1{,}67) + \text{Prob}\ (1{,}69 < X \leq 1{,}71)$$

Fig. 2



Conocidas estas probabilidades, podríamos determinar con bastante aproximación la composición de la población de la que se obtuvo la muestra.

Si suponemos que el efectivo total es N = 10.000.000, tendríamos:

/Nº de

Nº de individuos con una estatura comprendida entre 1,65 y 1,67 =

$$= 10.000.000 \; X \; \text{Prob} \; (1{,}65 < X \leq 1{,}67)$$

Obsérvese que la probabilidad de que la variable aleatoria tome un valor determinado es nula, por ejemplo:

$$\text{Prob} \; (X = 1{,}65) = 0$$

ya que el área del trapecio mixtilíneo quedaría reducida al área de un trapecio de altura cero.

Función de densidad de la probabilidad

Si la curva de probabilidad tiene una expresión analítica $y = f(x)$, a $f(x)$ se la denomina función de densidad de la probabilidad. Debe cumplir la condición $f(x) \geq 0$ para cualquier valor de x.

Una variable aleatoria continua X quedará definida, cuando se de el recorrido expresado por el intervalo (a, b) y la función de densidad $f(x)$. Simbólicamente:

$$X \begin{cases} a < X \leq b \\ y = f(x) \geq 0 \end{cases}$$

Función de distribución

Si formamos la distribución de frecuencias relativas acumuladas (columna 4 de la Tabla 19) y por los extremos superiores de los intervalos de clase levantamos ordenadas equivalentes a dichas frecuencias relativas, obtenemos un polígono acumulativo (Fig. 3) donde

$$AB = f.r \; (x \leq 1{,}67) \qquad [1]$$

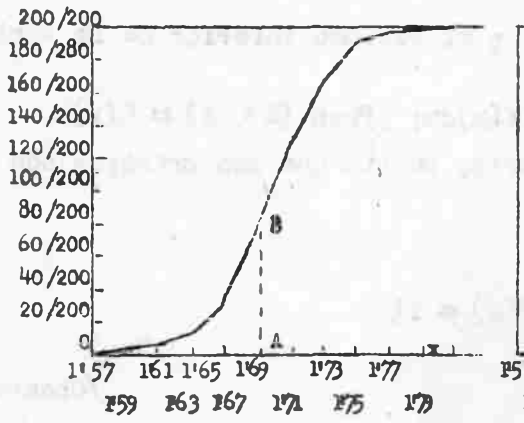


Fig. 3

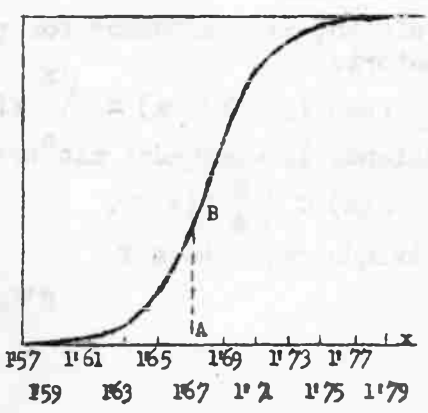


Fig. 4

/Si aumentamos

Si aumentamos indefinidamente el número de elementos de la muestra y hacemos tender a cero la amplitud de los intervalos, el polígono acumulativo adoptará la forma de una curva, cuya expresión analítica $y = F(x)$ se llama <u>función de distribución</u>.

La expresión [1] quedará convertida en la siguiente (Fig. 4):

$$AB = \text{Prob}\ (X \leq 1{,}67).$$

La función de distribución da la probabilidad de que la variable aleatoria X tome un valor igual o menor que un x dado, es decir:

$$\text{Prob}\ (X \leq x) = F(x)$$

La relación que liga la función de distribución con la f. de densidad es que

$$F'(x) = f(x)$$

o sea, que la <u>derivada de la f. de distribución es la f. de densidad</u>.[1]

A veces la función de distribución adopta una forma como la de la figura 5.

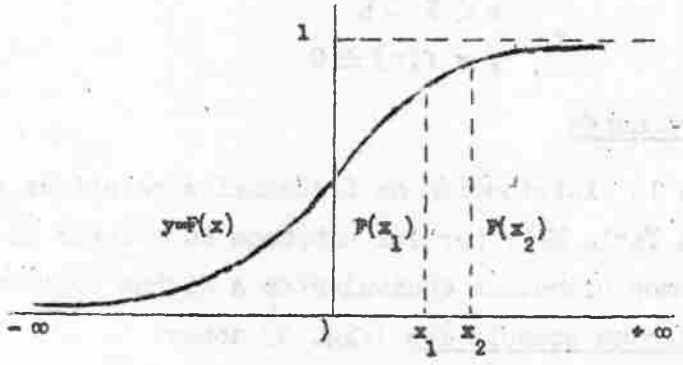


Fig. 5

1/ En efecto, si designamos por a el extremo inferior de la variable aleatoria

$$\text{Prob}\ (a < X \leq x) = \int_a^x f(x)dx; \quad \text{Prob}\ (X \leq x) = F(x)$$

Igualando los segundos miembros, puesto que los primeros son iguales,

$$F(x) = \int_a^x f(x)dx.$$

Derivando respecto a x

$$F'(x) = f(x)$$

/Observando la

Observando la figura se pone de manifiesto:

1) $F(x)$ es una función monótona no decreciente.

2) $F(-\infty) = 0$; $F(+\infty) = 1$.

3) Prob $(x_1 < X \leq x_2) = F(x_1) - F(x_2)$.

En la función de distribución las probabilidades vienen determinadas por las ordenadas y en la de densidad por las áreas.

La distribución normal

Una variable aleatoria continua X, cuyo recorrido es de $-\infty$ a $+\infty$ y que tiene por función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}}$$

se dice que sigue la distribución normal. Se puede demostrar, con ayuda del Cálculo integral, que los parámetros α y σ que determinan por completo la f. de densidad son precisamente la media y la desviación típica de la variable aleatoria continua X.[1/]

Gauss llegó a la función [2] al investigar la ley de distribución a que deben obedecer los errores de observación para que la media aritmética de una serie de medidas sea el valor más probable de la "verdadera" magnitud.

En matemáticas superiores se demuestra, mediante un cambio de variable y haciendo uso de la integral de Gauss, que el área encerrada por la curva y el eje de las x es igual a la unidad. Por otra parte, la [2], cualquiera que sea el valor de x, nunca toma valores negativos, es decir, $f(x) \geq 0$. Por tanto, [2] cumple las condiciones exigidas a una función de densidad.

La distribución normal de parámetros α y σ se la designa abreviadamente por Normal (α, σ).

1/ α y σ son dos parámetros poblacionales. Si fueran desconocidos podríamos calcularlos aproximadamente mediante la media y la desviación típica de una distribución muestral de frecuencias obtenida de la población, siempre que la distribución de frecuencias contuviese un gran número de observaciones.

/Estudio analítico

Estudio analítico de la curva

A continuación hacemos el estudio analítico de la f. de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

con objeto de dibujar la curva.

Asíntotas

Si $x \longrightarrow -\infty$; $f(x) \longrightarrow 0$; asíntota el eje de las x con contacto en $-\infty$

Si $x \longrightarrow -\infty$; $f(x) \longrightarrow 0$; asíntota el eje de las x con contacto en $+\infty$

Primera derivada

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)}{2\sigma^2}} \left[-\frac{1}{2\sigma^2} \cdot 2(x-\alpha) \right] = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} (x-\alpha)$$

Derivada segunda

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}} \left[e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \cdot \frac{1}{\sigma^2}(x-\alpha)^2 + e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \right] = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \left[1 - \frac{(x-\alpha)^2}{\sigma^2} \right]$$

Máximo

Si $x = \alpha$,, $f'(x) = 0$,, $f''(x) < 0$. Para $x = \alpha$ existe un máximo con ordenada

$$f(\alpha) = \frac{1}{\sigma\sqrt{2\pi}}$$

Puntos de inflexión

Se obtendrán anulando la segunda derivada, para lo cual

$$1 - \frac{(x-\alpha)^2}{\sigma^2} = 0 \quad ,, \quad \frac{(x-\alpha)^2}{\sigma^2} = 1 \quad ,, \quad \frac{x-\alpha}{\sigma} = \pm 1$$

/De $\frac{x-\alpha}{}$

De $\frac{x - \alpha}{\sigma} = +1$,, se obtiene $x = \alpha + \sigma$

De $\frac{x - \alpha}{\sigma} = -1$,, se obtiene $x = \alpha - \sigma$

Existen dos puntos de inflexión, con abscisas $\alpha + \sigma$ y $\alpha - \sigma$ siendo sus ordenadas respectivas.

$$f(\alpha + \sigma) = f(\alpha - \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}}$$

Concavidad

Para valores $|x - \alpha| > \sigma$, $f''(x) > 0$, o sea, en los intervalos $(-\infty , \alpha - \sigma)$, $(\alpha + \sigma, \infty)$, la curva vuelve la concavidad hacia la región positiva de Oy.

Para $|x - \alpha| < \sigma$, $f''(x) < 0$, es decir, en el intervalo $(\alpha - \sigma, \alpha + \sigma)$ la concavidad se vuelve hacia la región negativa de Oy.

Simetría

Si en la expresión [2] sustituimos x por $\alpha + k$ y por $\alpha - k$, se verifica:

$$f(\alpha + k) = f(\alpha - k)$$

La curva de la distribución normal es simétrica respecto a la recta $x = \alpha$, luego en esta distribución

$$\text{Media} = \text{Moda} = \text{Mediana} = \alpha$$

Representación gráfica

Suele tomarse una unidad mayor en el eje de ordenadas que en el de abscisas para que la forma de la curva normal no resulte muy aplanada (Fig. 6).

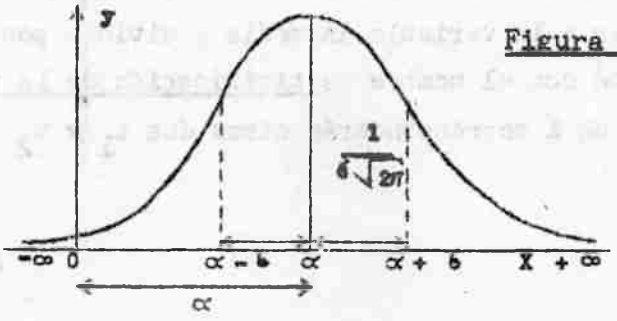
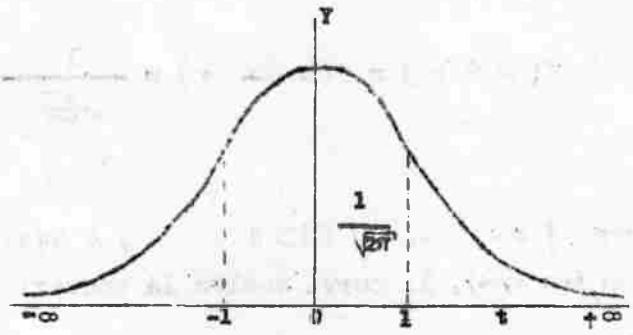


Figura 6

/Figura 7

Figura 7



Tablas

Para el uso práctico de la distribución normal existen diversos tipos de tablas, que se clasifican en dos grupos: tablas de áreas y tablas de ordenadas.

Tablas de áreas

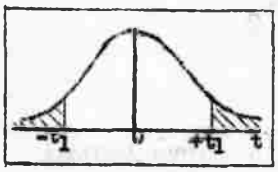
A veces interesa calcular Prob $(x_1 < X \leq x_2)$, es decir, la probabilidad de que la variable aleatoria X esté comprendida entre dos valores dados x_1 y x_2.

Como los parámetros poblacionales α y σ varían de una distribución a otra, interesa con una sola tabla resolver todos los problemas que se nos presenten en la práctica.

Para ello se efectúa el cambio de variable $\frac{X - \alpha}{\sigma} = t$. A esta operación de restar a la variable la media y dividir por la desviación típica se la conoce con el nombre de tipificación de la variable. A dos valores, x_1 y x_2, de X corresponderán otros dos t_1 y t_2 dados por las

/Fig. 8

Fig. 8

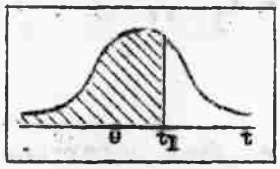


PROB $(/t/ \geq t_1)$

TABLA 20

t	Probabilidad
0'0	1'00000
0'1	0'92034
0'2	0'84148
0'3	0'76418
0'4	0'68916
0'5	0'61708
0'6	0'54850
0'7	0'48392
0'8	0'42372
0'9	0'36812
1'0	0'31732
1'1	0'27134
1'2	0'23014
1'3	0'19360
1'4	0'16152
1'5	0'13362
1'6	0'10960
1'7	0'08914
1'8	0'07186
1'9	0'05744
2'0	0'04550
2'1	0'03572
2'2	0'02780
2'3	0'02144
2'4	0'01640
2'5	0'01242
2'6	0'00932
2'7	0'00694
2'8	0'00512
2'9	0'00374
3'0	0'00270
3'1	0'00194
3'2	0'00138
3'3	0'00096
3'4	0'00068
3'5	0'00046
3'6	0'00032
3'7	0'00022
3'8	0'00014
3'9	0'00010
4'0	0'00006

Fig. 9

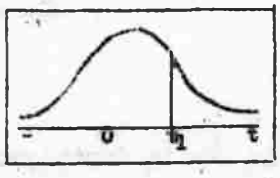


PROB $(t \leq t_1)$

TABLA 21

t	Probabilidad
0'0	0'50000
0'1	0'53989
0'2	0'57926
0'3	0'61791
0'4	0'65542
0'5	0'69146
0'6	0'72575
0'7	0'75804
0'8	0'78814
0'9	0'81594
1'0	0'84134
1'1	0'86433
1'2	0'88493
1'3	0'90320
1'4	0'91924
1'5	0'93319
1'6	0'94520
1'7	0'95543
1'8	0'96407
1'9	0'97128
2'0	0'97725
2'1	0'98214
2'2	0'98610
2'3	0'98928
2'4	0'98180
2'5	0'99379
2'6	0'99534
2'7	0'99653
2'8	0'99744
2'9	0'99819
3'0	0'99865
3'1	0'99903
3'2	0'99931
3'3	0'99952
3'4	0'99966
3'5	0'99977
3'6	0'99984
3'7	0'99989
3'8	0'99993
3'9	0'99995
4'0	0'99997

Fig. 10



ORDENADA PARA $t = t_1$

TABLA 22

t	Probabilidad
0'0	0'3989
0'1	0'3970
0'2	0'3910
0'3	0'3814
0'4	0'3683
0'5	0'3521
0'6	0'3332
0'7	0'3123
0'8	0'2899
0'9	0'2661
1'0	0'2420
1'1	0'2179
1'2	0'1942
1'3	0'1714
1'4	0'1497
1'5	0'1295
1'6	0'1109
1'7	0'0941
1'8	0'0790
1'9	0'0656
2'0	0'0540
2'1	0'0440
2'2	0'0355
2'3	0'0283
2'4	0'0224
2'5	0'0175
2'6	0'0136
2'7	0'0104
2'8	0'0079
2'9	0'0060
3'0	0'0044
3'1	0'0033
3'2	0'0017
3'3	0'0012
3'4	0'0024
3'5	0'0009
3'6	0'0006
3'7	0'0004
3'8	0'0003
3'9	0'0002
4'0	0'0001

fórmulas

$$\frac{x_1 - \alpha}{\sigma} = t_1 \quad ,, \quad \frac{x_2 - \alpha}{\sigma} = t_2$$

Se puede demostrar que el área encerrada por la curva Normal (α, σ) y las ordenadas correspondientes a las abscisas x_1 y x_2, es la misma que la determinada por la curva

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \qquad [3]$$

y las ordenadas correspondientes a las abscisas t_1 y t_2.[1/]

La curva [3] puede ponerse en la forma

$$\frac{1}{1 \cdot \sqrt{2\pi}} e^{-\frac{(t-0)^2}{2 \cdot 1^2}}$$

que es la f. de densidad de la distribución normal con <u>media cero y desviación típica igual a la unidad,</u> por cuyo motivo se la simboliza por Normal (0,1). Su estudio analítico puede hacerse de forma análoga al de la Normal (α, σ). Su representación gráfica en la figura 7.

Las áreas de la Normal (0, 1) están tabuladas. En la Tabla 20 (véase figura 8) aparecen rayadas las áreas de las dos colas de la curva.

1/ En efecto:

$$\text{Prob}\,(x_1 < X \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} dx$$

Con el cambio de variable $\frac{X - \alpha}{\sigma} = t$, $dx = \sigma\, dt$, y sustituyendo en la expresión anterior el integrando, los límites de la integral y el elemento dx, tendremos

$$\text{Prob}\,(x_1 < X \leq x_2 = \int_{t_1}^{t_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

/Las áreas

Las áreas dan

$$\text{Prob}\ (-\infty < t \leq t_1) = \text{Prob}\ (t_1 < t \leq \infty) = \text{Prob}\ (|t| \geq t_1) = 2\ \text{Prob}\ (t > t_1)$$

o también

$$\text{Prob}\left(-\infty < \frac{X-\alpha}{\sigma} \leq t_1\right) = \text{Prob}\left(t_1 < \frac{X-\alpha}{\sigma} \leq \infty\right) =$$

$$= \text{Prob}\left(\left|\frac{X-\alpha}{\sigma}\right| \geq t_1\right) = 2\ \text{Prob}\left(\frac{X-\alpha}{\sigma} \geq t_1\right)$$

Ahora bien

$$\text{Prob}\left(\left|\frac{X-\alpha}{\sigma}\right| \geq t_1\right) = \text{Prob}\ (|X-\alpha| \geq t_1\,\sigma)$$

o sea, la Tabla 20 nos da la probabilidad de obtener una desviación respecto a la media que en valor absoluto sea igual o mayor que un cierto múltiplo de la desviación típica.

En la tabla 20 sólo figuran valores positivos de t.

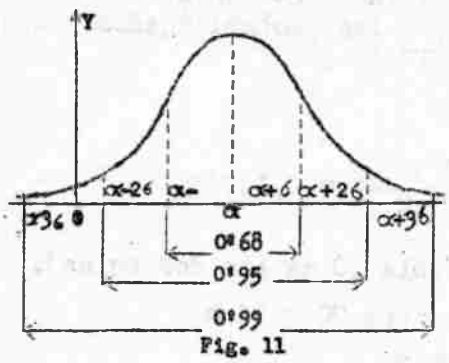


Fig. 11

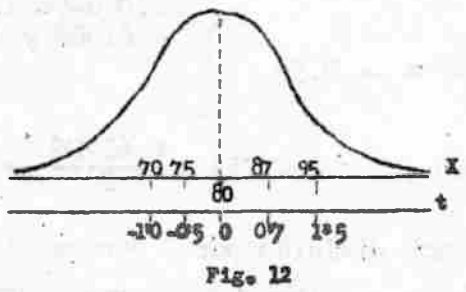


Fig. 12

Areas bajo la curva normal

De las tablas de áreas de la distribución normal se obtienen las siguientes relaciones (véase figura 11):

/Prob (—

$$\text{Prob}\ (\alpha - \sigma < X \leq \alpha + \sigma) = 0{,}68268 \simeq 0{,}68$$
$$\text{Prob}\ (\alpha - 2\sigma < X \leq \alpha + 2\sigma) = 0{,}95450 \simeq 0{,}95$$
$$\text{Prob}\ (\alpha - 3\sigma < X \leq \alpha + 3\sigma) = 0{,}99730 \simeq 0{,}99$$

es decir,

$\text{Prob}\ (-1 < t \leq +1) \simeq 0{,}68$ $\qquad$ $\text{Prob}\ (|X - \alpha| > \sigma) \simeq 0{,}32$

$\text{Prob}\ (-2 < t \leq +2) \simeq 0{,}95$ o también $\text{Prob}\ (|X - \alpha| > 2\sigma) \simeq 0{,}05$

$\text{Prob}\ (-3 < t \leq +3) \simeq 0{,}99$ $\qquad$ $\text{Prob}\ (|X - \alpha| > 3\sigma) \simeq 0{,}01$

En la práctica estas relaciones son de uso frecuente. Obsérvese que para valores de la variable fuera del intervalo $(\alpha - 3\sigma, \alpha + 3\sigma)$, la curva se confunde prácticamente con el eje de las X.

Ejemplo: En una distribución Normal (80, 10), N = 1.000. Encontrar el número de elementos con las siguientes características:

a) Comprendidos entre 70 y 75.

b) Comprendidos entre 87 y 95.

c) Comprendidos entre 75 y 87.

N = 1.000 es el efectivo total del colectivo. Calculadas las probabilidades y multiplicadas por N, tendríamos el número de elementos que cumplen determinadas características. (Véase figura 12.)

a) Valores tipificados:

$$\frac{70 - 80}{10} = -1{,}0$$

$$\frac{70 - 80}{10} = -0{,}5$$

Entrando en la Tabla 20 con los valores 0,5 y 1,0 obtendremos las probabilidades respectivas 0,61708 y 0,31732.

$$\text{Prob}\ (70 < X \leq 75) = \frac{0{,}61708}{2} - \frac{0{,}31732}{2} = 0{,}14988$$

(Hemos dividido por 2 porque la Tabla 20 da las dos colas).

Número de elementos comprendidos entre 70 y 75 =

$$= 1.000 \times 0{,}14988 \simeq 150$$

b) Valores tipificados:

$$\frac{95 - 80}{10}$$

$\frac{95 - 80}{10} = 1{,}5$

$\frac{87 - 80}{10} = 0{,}7$

Entrando en la Tabla 20 con los valores 0,7 y 1,5 obtendremos las probabilidades 0,48392 y 0,13362.

$$\text{Prob } (87 < X \leq 95) = \frac{0{,}48392}{2} - \frac{0{,}13362}{2} = 0{,}17515$$

Número de elementos comprendidos entre 87 y 95 =

$$= 1.000 \times 0{,}17515 \simeq 175$$

c) Los valores tipificados para X = 75 y X = 87 ya fueron hallados.

$$\text{Prob } (75 < X \leq 87) = \text{Prob } (75 < X \leq 80) + \text{Prob } (80 < X \leq 87) =$$

$$= 0{,}5 - \frac{0{,}61708}{2} + 0{,}5 - \frac{0{,}48392}{2} = 0{,}44950$$

(Téngase en cuenta que el área encerrada por la mitad de la curva vale 0,5).

Número de elementos comprendidos entre 87 y 95 =

$$= 1.000 \times 0{,}44950 \simeq 450$$

Otro tipo de Tabla de áreas

La Tabla 21 (Fig. 9) da las áreas situadas a la izquierda de la ordenada correspondiente a un valor t_1. La Tabla da

$$\text{Prob } (t \leq t_1)$$

sólo aparecen valores positivos de t; si t_1 fuera negativo,

$$\text{Prob } (t \leq - t_1) = 1 - \text{Prob } (t \leq t_1)$$

Teniendo en cuenta los valores tipificados obtenidos, el problema anterior se resolvería de la siguiente forma:

a) $\text{Prob } (70 < X \leq 75) = \text{Prob } (-1{,}0 < t \leq -0{,}5) =$

$$= \text{Prob } (t \leq -0{,}5) - \text{Prob } (t \leq -1{,}0) = 1 - \text{Prob } (t \leq 0{,}5) - [1 - \text{Prob } (t \leq 1{,}0)]$$

Entrando en la Tabla 21 con los valores tipificados 0,5 y 1,0 obtenemos las probabilidades 0,69146 y 0,84134, luego

$\text{Prob } (70 < X \leq 75) = 1 - 0{,}69146 - (1 - 0{,}84134) = 0{,}14988.$

b) $\text{Prob } (87 < X \leq 95) = \text{Prob } (0{,}7 < t \leq 1{,}5) =$

$$= \text{Prob } (t \leq 1{,}5) - \text{Prob } (t \leq 0{,}7)$$

/Entrando en

Entrando en la Tabla con los valores tipificados 1,5 y 0,7 obtenemos las probabilidades 0,93319 y 0,75804, luego

$$\text{Prob}\ (87 < X \leq 95) = 0{,}93319 - 0{,}75804 = 0{,}17515.$$

c) $\text{Prob}\ (75 < X \leq 87) = \text{Prob}\ (-0{,}5 < t \leq 0{,}7) =$

$$= \text{Prob}\ (t \leq 0{,}7) - \text{Prob}\ (t \leq -0{,}5) = \text{Prob}\ (t \leq 0{,}7) - [1 - \text{Prob}\ (t \leq 0{,}5)]$$

Entrando en la Tabla con los valores tipificados 0,7 y 0,5 obtenemos las probabilidades 0,75804 y 0,69146, luego

$$\text{Prob}\ (75 < X \leq 87) = 0{,}75804 - (1 - 0{,}69146) = 0{,}44950.$$

Obsérvese que conocida la Tabla 20 se calcula fácilmente la Tabla 21. Basta dividir por 2 las probabilidades de la primera y restar de 1. (Véase la relación entre las áreas de las figuras 8 y 9).

Tablas de ordenadas

Dan las ordenadas para determinados valores de t en la curva Normal (0,1):

$$f(t) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}$$

Si se trata de una distribución Normal (α , σ) de f. de densidad

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$$

tipificando la variable mediante el cambio de variable $\frac{X-\alpha}{\sigma} = t$, la Normal (α , σ) queda en la forma

$$f(x) = \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}$$

o sea

$$f(x) = \frac{1}{\sigma} \cdot f(t) \qquad [4]$$

Como las ordenadas de la f(t) están tabuladas, para hallar la ordenada correspondiente a un valor x, se divide por σ el valor de f(t).

Con las tablas de ordenadas se pueden calcular las probabilidades de la distribución binomial mediante las ordenadas de la normal, ya que en Estadística matemática se demuestra que el límite de la distribución binomial, cuando el parámetro $n \to \infty$, es la distribución normal y la aproximación es aceptable siempre que

$$p \leq \frac{1}{2} \quad ,, \quad np > 5$$

/En la

En la Tabla 22 (Fig. 10) aparecen las ordenadas de la Normal (0,1) para valores positivos de t. Para valores negativos de t serían iguales a los positivos correspondientes, a causa de la simetría de la curva.

Ejemplo 1:

Se lanzan cien monedas 1.000 veces. ¿Qué número aproximado de veces es de esperar que se obtengan 55 caras?

Se trata de una distribución binomial, donde

$$n = 100 \quad x = 55 \quad ,, \quad p = q = \frac{1}{2} \quad N = 1.000$$

La probabilidad de obtener 55 caras sería

$$\binom{100}{55} \left(\frac{1}{2}\right)^{55} \cdot \left(\frac{1}{2}\right)^{45}$$

El cálculo de la anterior expresión sería en extremo laborioso, pero siendo $p = \frac{1}{2}$, $np = 50 > 5$, podemos calcular la probabilidad pedida mediante la Tabla de ordenadas de la distribución normal, considerando que la media y la desviación típica de esta última distribución son aproximadas al valor medio teórico y a la desviación típica de la binomial.

En el tema "Distribuciones de probabilidad", vimos que

$$\alpha = np = 100 \cdot \frac{1}{2} = 50 \quad ,, \quad \sigma = \sqrt{npq} = \sqrt{100 \times 0,5 \times 0,5} = 5$$

En la Normal (50, 5) buscaremos la ordenada correspondiente a $x = 55$. Tipificando la variable tenemos:

$$t = \frac{X - \alpha}{\sigma} = \frac{55 - 50}{5} = 1,0$$

Entrando en la Tabla 22 con el valor 1,0 obtenemos la ordenada 0,2420.

Mediante la relación [4], tenemos

$$\text{Prob de 55 caras} = f(x) = \frac{1}{\sigma} f(t) = \frac{1}{5} \cdot 0,2420$$

/Número de

Número de 55 caras en 1.000 tiradas = $1.000 \cdot \frac{1}{5} \cdot 0{,}2420 \simeq 48$.

Un resultado de 55 caras es de esperar ocurra unas 48 veces en 1.000 tiradas.

Mediante la Tabla de áreas de la distribución normal puede resolverse el siguiente problema.

Ejemplo 2.

Se lanzan cien monedas 1.000 veces. ¿Qué número aproximado de veces es de esperar que se obtengan 55 o más caras?

Deberíamos calcular una suma de probabilidades del tipo siguiente:

$$\text{Prob (55 caras)} + \text{Prob (56 caras)} + \ldots + \text{Prob (100 caras)} =$$

$$= \binom{100}{55}\left(\frac{1}{2}\right)^{55}\left(\frac{1}{2}\right)^{45} + \ldots\ldots + \binom{100}{100}\left(\frac{1}{2}\right)^{100} \qquad [5]$$

Por tratarse de una distribución binomial la variable aleatoria es discreta. Para hacer uso de las Tablas de áreas de la distribución normal debemos hacer la consideración de que un resultado de 55 caras en la binomial equivale a un intervalo (54,5, 55,5) en la escala continua de la curva normal.

En el problema anterior se calcularon

$$\alpha = 50 \qquad ,, \qquad \sigma = 5$$

y siguen subsistiendo las mismas condiciones para aproximar la distribución binomial por la Normal (50,5).

Tipificando la variable para X = 54,5,

$$t = \frac{54{,}5 - 50}{5} = 0{,}9$$

La suma [5] será equivalente a calcular Prob ($t \geq 0{,}9$).

Entrando en la Tabla 21 con t = 0,9

Prob ($t \geq 0{,}9$) = 1 — Prob ($t \leq 0{,}9$) = 1 — 0,81594 = 0,18406

Número de veces que aparecerán 55 o más caras =

= 1.000 X 0,18406 $\simeq$ 184 veces.

/Ajuste de

Ajuste de una distribución normal a una distribución muestral de frecuencias

Para ajustar una distribución normal a la distribución muestral de frecuencias de la Tabla 19, dispondremos los cálculos según la Tabla 23.

Tabla 23

Intervalos	Marcas de clase x_i	Frecuencias	x''_i	$x''_i n_i$	$x''^2_i n_i$	Valores tipificados t	Prob. acumuladas	Prob. de los intervalos (*)	$N \cdot p_i$	n_t
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.57 - 1.59	1.58	1	-5	-5	25	$\frac{1.59 - 1.68}{0.03} = -3.0$	0.00135	0.00135	0.270	0
1.59 - 1.61	1.60	2	-4	-8	32	$\frac{1.61 - 1.68}{0.03} = -2.3$	0.01072	0.00937	1.874	2
1.61 - 1.63	1.62	7	-3	-21	63	$\frac{1.63 - 1.68}{0.03} = -1.7$	0.04457	0.03385	6.770	7
1.63 - 1.65	1.64	19	-2	-38	76	$\frac{1.65 - 1.68}{0.03} = -1.0$	0.15866	0.11409	22.818	23
1.65 - 1.67	1.66	44	-1	-44	44	$\frac{1.67 - 1.68}{0.03} = -0.3$	0.38209	0.22343	44.686	45
1.67 - 1.69	1.68	50	0	0	0	$\frac{1.69 - 1.68}{0.03} = 0.3$	0.61791	0.23582	47.164	47
1.69 - 1.71	1.70	42	1	42	42	$\frac{1.71 - 1.68}{0.03} = 1.0$	0.84134	0.22343	44.686	44
1.71 - 1.73	1.72	25	2	50	100	$\frac{1.73 - 1.68}{0.03} = 1.7$	0.95543	0.11409	22.818	23
1.73 - 1.75	1.74	8	3	24	72	$\frac{1.75 - 1.68}{0.03} = 2.3$	0.98928	0.03385	6.770	7
1.75 - 1.77	1.76	1	4	4	16	$\frac{1.77 - 1.68}{0.03} = 3.0$	0.99865	0.00937	1.874	2
1.77 - 1.79	1.78	1	5	5	25	$\frac{1.79 - 1.68}{0.03} = 3.7$	0.99989	0.00124	0.248	0
Totales		200		9	495					200

La significación de las columnas de la Tabla 23 es la siguiente:

1. Los intervalos de clase
2. Las marcas de clase de los intervalos
3. Frecuencias absolutas

(*) 0.00135 es la probabilidad de que una observación se encuentre comprendida en el intervalo $(-\infty, 1.59)$.

(4. Desviaciones

4. Desviaciones respecto al origen de trabajo $0_x = 1{,}68$; las desviaciones han sido medidas tomando como unidad la amplitud del intervalo.

5. Producto de la columna (4) por la (3).

6. Producto de (5) por (4)

La media aritmética de la distribución de frecuencias es:

$$a = 0_x + \frac{\Sigma x_i'' n_i}{n} \cdot c = 1{,}68 + \frac{9}{200} \cdot 0{,}02 \simeq 1{,}68$$

y la desviación típica

$$s = c\sqrt{\frac{\Sigma x_i''^2 n_i}{n} - \left(\frac{\Sigma x_i'' n_i}{n}\right)^2} = 0{,}02\sqrt{\frac{495}{200} - \left(\frac{9}{200}\right)^2} \simeq 0{,}03$$

Estimamos los parámetros α y σ de la distribución normal que son desconocidos por a y s media y desviación típica muestrales, por tanto,

$\alpha = 1{,}68$,, $\sigma = 0{,}03$.

7. Valores tipificados de los extremos superiores de los intervalos, obtenidos mediante la fórmula

$$\frac{X - \alpha}{\sigma} = t$$

8. Probabilidades acumuladas.

Así, para calcular la Prob $(-\infty < t \leq -3{,}0)$, entramos en la Tabla 20 con el valor 3,0, que da la probabilidad 0,00270; dividiendo por 2 se obtiene 0,00135. Para calcular Prob $(-\infty < t \leq 0{,}3)$, entramos en la Tabla con 0,3, que da la probabilidad 0,76418; dividiendo por 2 se obtiene 0,38209, que restada de 1 da 0,61791.

9. Probabilidad de cada intervalo. Cada cifra de la columna (8) se resta de la siguiente. Así, 0,01072 — 0,00135 = 0,00937.

10. Frecuencias teóricas n_t obtenidas al ajustar la distribución normal a la distribución muestral de frecuencias.

$$n_t = N \times p_i$$

siendo $N = 200$ y p_i la probabilidad correspondiente a cada intervalo.

/11. Las

11. Las cifras de la columna (10) se han redondeado con objeto de que el total de dicha columna sea 200.

Otro ejemplo

En una investigación se halló que la relación entre los diámetros longitudinal y transversal de 815 hojas de tabaco se distribuia como figura en la Tabla 24. Ajustar una distribución normal.

Tabla 24

Intervalos (1)	Frecuencias n_i (2)	Valores tipificados t (3)	Probab. acumuladas (4)	Probab. de los intervalos (*) (5)	$N_x P_i$ (6)	Frecuencias teóricas u_T (7)
1.55 - 1.65	4	-2.8	0.00256	0.00256	2.09	2
1.65 - 1.75	9	-2.2	0.01390	0.01134	9.24	9
1.75 - 1.85	31	-1.6	0.05480	0.04090	33.33	33
1.85 - 1.95	75	-1.0	0.15866	0.10386	84.65	85
1.95 - 2.05	183	-0.3	0.38209	0.22343	182.10	182
2.05 - 2.15	204	0.3	0.61791	0.23582	192.19	192
2.15 - 2.25	157	0.9	0.81594	0.19803	161.39	162
2.25 - 2.35	97	1.5	0.93319	0.11725	95.56	96
2.35 - 2.45	40	2.1	0.98214	0.04895	39.89	40
2.45 - 2.55	12	2.8	0.99744	0.01530	12.47	12
2.55 - 2.65	3	3.4	0.99966	0.00222	1.81	2
Totales ...						815

La marcha general es la misma que la seguida en el ejemplo precedente.

Suprimimos en la Tabla anterior las columnas necesarias para calcular la media y la desviación típica.

$$a = 2{,}106 \qquad \text{,,} \qquad s = 0{,}161$$

/El primer

El primer valor tipificado de la columna (3) es

$$\frac{1,65 - 2,106}{0,161} = -2,8$$

Se han subrayado las cifras forzadas en las décimas, cuando la cifra de las centésimas era igual o mayor que cinco.

La probabilidad de que t esté comprendida entre $-\infty$ y $+\infty$ es la unidad, por tanto, la probabilidad de que una observación estuviera entre 2,65 e ∞ sería 1 — 0,99966 = 0,00034, que al multiplicar por 815 da 0,27710, menos que media unidad.

Si en vez de limitar en las décimas los valores tipificados, tuvieran éstos más cifras decimales, se efectuaría una interpolación lineal para calcular las probabilidades.

Tanto en este ejemplo como en el anterior el ajuste es bastante bueno, pues las diferencias entre las frecuencias observadas y las teóricas son pequeñas.

Importancia de la distribución normal en la teoría y en la práctica estadística

En la práctica, a un gran número de fenómenos corresponden variables que siguen la ley normal más o menos aproximadamente. Recordemos especialmente, en Biometría, las distribuciones de frecuencias para tallas y pesos.

En Psicología se ha comprobado que las aptitudes de los individuos varían en intensidad según la curva normal.

En la industria se distribuyen normalmente las dimensiones de las piezas fabricadas por una máquina.

Los errores accidentales o de observación en Astronomía, Geodesia, etc., siguen también, aproximadamente, la ley normal, a la cual puede llegarse deduciéndola de ciertos postulados compatibles con la experiencia.

La gran importancia teórica de la distribución normal radica en que, bajo determinadas condidiciones son muchas las distribuciones que tienden a la ley normal, esto es, es un "dominio de atracción".

/También tiene

También tiene gran importancia teórica en las hipótesis de normalidad, que son fundamentales en la teoría de muestras; así, por ejemplo, al examinar la concordancia entre teoría y hechos, suele suponerse que la distribución de las medias en las muestras grandes es más o menos aproximadamente normal, lo que es cierto en muestras procedentes de poblaciones más o menos normales, y tiende a serlo en general cuando crece el número de datos.

D. IDEAS FUNDAMENTALES SOBRE LA TEORIA DE MUESTRAS

Estadística descriptiva y Estadística inductiva

Dos aspectos cabe considerar en Estadística conocidos con los nombres de Estadística descriptiva y Estadística inductiva.

En la Estadística descriptiva se trata de describir el fenómeno estudiado mediante unas cuantas características que resuman los datos estadísticos.

La Estadística inductiva o inferencia estadística tiene por objeto inducir las características de un colectivo observando una parte del mismo.

Muestra y población.

Si nos interesara conocer la edad media de los estudiantes de una Universidad, nuestros cálculos deberían efectuarse con los datos registrados para cada alumno. Tal procedimiento sería muy laborioso y emplearíamos gran cantidad de tiempo. Sería mucho más fácil para nuestro propósito seleccionar una muestra de estudiantes y efectuar el análisis con los datos obtenidos, con lo que induciríamos ciertas conclusiones. La mayor parte de los estudios estadísticos se realizan con muestras y no con enumeraciones completas de los datos. Una muestra estadística representa solamente una parte de un conjunto más amplio. El conjunto de donde se obtuvo una muestra se denomina población o universo.

Muestreo Aleatorio

El muestreo es un proceso mediante el cual se selecciona de un conjunto determinado llamado población universo o colectivo, un conjunto parcial que recibe el nombre de muestra.

Si examinamos una población con un número finito de elementos, extraemos un elemento, anotamos la característica que nos interese y lo devolvemos a la población, hemos realizado un muestreo con reemplazamiento. Si el elemento extraído no se devuelve a la población, el muestreo realizado se denomina sin reemplazamiento. En el primer caso, el segundo

/elemento extraído

elemento extraído tiene la misma probabilidad que el primero. No ocurre así en el muestreo sin reemplazamiento, aunque si la población está compuesta por un gran número de elementos, podemos considerar que la extracción de un elemento no ha hecho variar la estructura probabilística de la población.

Debe evitarse cualquier subjetividad en la elección de la muestra. Supongamos una muestra de estudiantes que fueron seleccionados con objeto de estimar la inteligencia mediana del conjunto de alumnos de un centro docente. Una persona encargada de la tarea de elegir la muestra que no estuviera enterada de los procedimientos estadísticos adecuados, podrían tener tendencias en diferentes sentidos: puede no incluir bastantes casos próximos a la mediana, eligiendo individuos de gran y pequeña inteligencia, creyendo que unos equilibrarían a otros; pueden separar los casos extremos eligiendo sólo personas de inteligencia media; o puede pensar que las personas de escasa inteligencia son excepcionales y no incluirlas en la muestra.

Para evitar toda consideración subjetiva se obtienen muestras aleatorias. Para la realización práctica del proceso del muestreo un procedimiento es el siguiente: Si tratamos de elegir domicilios en una ciudad con objeto de llevar a cabo una muestra de viviendas a cada domicilio le asignamos un número. Los números pueden ser elegidos obteniéndolos de un bombo giratorio o, mejor aún, de unas tablas de números obtenidos al azar. Las viviendas cuyos domicilios han sido elegidos de esa forma se incluyen en la muestra y de esas viviendas obtendríamos los datos que nos interesara estudiar. De esta manera hemos obtenido una muestra objetiva, habiendo eliminado las subjetividades del ejemplo anterior. En la Teoría de muestras los errores debidos al libre albedrío de la persona que extrae la muestra, se conocen con el nombre de sesgos debidos al proceso de selección.

Otros sesgos pueden ser debidos a sustitución de unidades en la muestra. Si se trata de estimar el número de personas por vivienda y el entrevistador sustituye una vivienda cerrada por la más próxima, se comprende fácilmente que las familias numerosas aparecerán por exceso,

/ya que

ya que una vivienda de un matrimonio sin hijos es muy posible esté cerrada en determinadas horas del día por encontrarse ambos cónyuges en el trabajo.

Tablas de números aleatorios.

Una Tabla de números aleatorios es una tabla de números obtenidos al azar. Están constituídas por series de columnas de números dígitos obtenidos por métodos experimentales que garantizan en lo posible las ausencias de tendencias o sesgos.

Copiamos un fragmento de las tablas espuestas por R. Clay en su libro Elementary Statistics:

06	43	38	...	...
39	29	84	...	...
89	88	45	...	...
61	51	23	...	...
99	65	34	...	...
.....				
.....				

Supongamos que de una población de 700 elementos queremos sacar una muestra de 70 elementos. Como necesitamos sacar números de 001 al 700, cogeremos tres columnas de dígitos de la tabla y sacaremos los 70 primeros números que aparezcan menores que 701, es decir, los números 064, 392, 615, etc. Hemos excluído el 898 y el 996 por ser mayores que 701.

Estimación y parámetro.

Una función de las observaciones de una muestra recibe el nombre de estadígrafo o estimador. Algunos tratadistas le dan el nombre de estadístico.

Si de una población obtenemos una muestra de n elementos x_1, x_2, ..., x_n (muestra de tamaño n), un estadístico será una función de estas observaciones que emplearemos para estimar algún parámetro de la población. Para estimar la media de una población empleamos la función

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

/Una medida

Una medida calculada a partir de una muestra se llama una estimación. Si x_1. x_2, ..., x_n representan las tallas de un grupo de personas, $\bar{X}$ será una estimación (un número) que nos servirá para estimar el correspondiente parámetro de la población, es decir, la talla media del universo del que se obtuvo la muestra.

Casi nunca la estimación y el parámetro serán idénticos. Por otra parte, la estimación variará de muestra a muestra porque estará afectada por las tallas del grupo particular de personas que entraron a formar parte de la muestra, y el azar ha ejercido su influencia en la selección de las personas.

Distribución muestral de un Estadístico

Puesto que el azar determina qué casos particulares están incluídos en cada muestra de una población, inevitablemente habrá alguna variabilidad entre las muestras estadísticas. Si de una población de 9.000 estudiantes con una estatura media de 1,68 se obtienen una serie de muestras de 50 estudiantes cada una, la media obtenida para cada muestra probablemente no será 1,68. El valor del estadístico para la primer muestra puede ser 1,66; para la segunda, 1,70, y para la tercera, 1,69.

Si este proceso se repitiera muchas veces, algunos valores del estadístico se presentarán con más frecuencia que otros. Los valores del estadístico alrededor del parámetro 1,68 se presentarán más a menudo que los valores más alejados de él. Habrá una concentración de medias muestrales alrededor de 1,68 y serán tanto menos frecuentes las medias muestrales cuanto más difieran de 1,68: las medias muestrales con 1,55 se presentarán con menos frecuencia que las de 1,67, mientras que las de 1,70 serán más frecuentes que las de 1,78.

Con un gran número de medias muestrales podría formarse una distribución de frecuencias relativas. La media, la desviación típica y otras medidas podrían calcularse en esta distribución de frecuencias relativas. Si el número de muestras fuese muy grande, el contorno del histograma tendería hacia la curva de probabilidad y la distribución de frecuencias relativas en una distribución de probabilidad que se denomina la distribución muestral del estadístico.

/La media

La media de las medias muestrales sería la media de la población como se verá más adelante, y en este caso sería 1,68, y la desviación típica de la distribución muestral del estadístico indicaría la dispersión de las medias muestrales. Esta desviación típica recibe el nombre de error típico o error de muestreo.

Posteriormente, veremos las distribuciones muestrales de algunos estadísticos, tales como medias y proporciones. El conocimiento de tales distribuciones muestrales sirve para contestar a muchas de las críticas del muestreo. Aunque los resultados muestrales varíen, el conocimiento del error de muestreo sirve para efectuar generalizaciones acerca de los parámetros de la población.

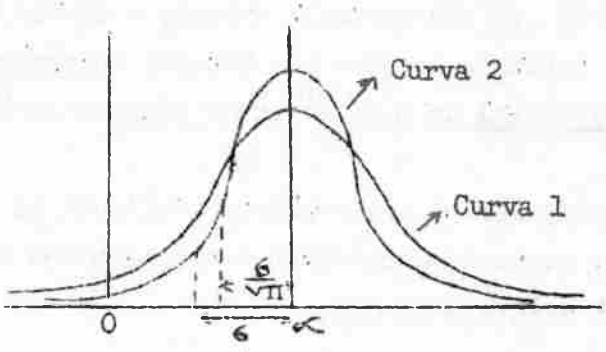
La distribución muestral de la media y la curva normal.

Si el tamaño de la muestra es igual o mayor que 30, la distribución muestral de la media aritmética es aproximadamente normal, aunque la población de donde procedan las muestras no sea normal. Este hecho, que se puede observar experimentalmente, tiene una considerable importancia práctica. En Estadística matemática se demuestra que si una población tiene varianza finita σ^2 y media μ, la distribución de la media muestral tiende a la distribución normal con desviación típica $\frac{\sigma}{\sqrt{n}}$ y media α al aumentar el tamaño de la muestra. La condición de que la varianza sea finita no es una restricción muy fuerte en lo que se refiere a la estadística aplicada, ya que prácticamente siempre será finito el recorrido de la variable aleatoria, en cuyo caso la varianza será necesariamente finita.

Podemos decir que las medias muestrales se distribuyen normalmente con media la de la población y desviación típica (error de muestreo) $\frac{\sigma}{\sqrt{n}}$ En la figura 1 la curva 1 es la de una distribución Normal (α, σ), y la curva 2 representa la distribución de las medias muestrales que es Normal $(\alpha \quad \frac{\sigma}{\sqrt{n}})$. Como la desviación típica de la curva 2 es menor que la de la 1, ésta resulta más achatada.

/Figura 1

Figura 1



Intervalos de confianza.

Hay una probabilidad de 0,95 de que la media muestral $\bar{X}$ caiga dentro del intervalo

$$\left(\alpha - 2 \frac{\sigma}{\sqrt{n}}, \alpha + 2 \frac{\sigma}{\sqrt{n}} \right)$$

es decir

$$\text{Prob} \left(\alpha - 2 \frac{\sigma}{\sqrt{n}} < X < \alpha + 2 \frac{\sigma}{\sqrt{n}} \right) = 0,95$$

expresión que, mediante sencillas transformaciones, es equivalente a la siguiente:

$$\text{Prob} \left(\bar{X} - 2 \frac{\sigma}{\sqrt{n}} < \alpha < X + 2 \frac{\sigma}{\sqrt{n}} \right) = 0,95$$

significando que hay una probabilidad de 0,95 de que el intervalo aleatorio

$$\left(\bar{X} - 2 - \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + 2 \frac{\sigma}{\sqrt{n}} \right)$$

contenga a la media α de la población.

Si obtenemos una media muestral particular $\bar{X}_1$, ya no será una variable aleatoria, sino un número; si α cae dentro del intervalo.

$$\left(\bar{X}_1 - 2 \right.$$

$$\left(\bar{X}_1 - 2 \frac{\sigma}{\sqrt{n}}, \; \bar{X}_1 + 2 \frac{\sigma}{\sqrt{n}} \right)$$

la probabilidad de que ocurra es la unidad, y si cae fuera la probabilidad será cero. Entonces decimos que tenemos un <u>coeficiente de confianza</u> o <u>probabilidad fiducial</u> de 0,95 de que α caiga dentro del intervalo anterior.

Ejemplo: Calcular el intervalo de confianza al 95 por 100 para la media α de una población, habiéndose obtenido una media muestral $\bar{X}$ = 1,675, siendo σ = 0,08 y n = 81.

$$2 \frac{\sigma}{\sqrt{n}} = 0,017 \; ; \quad \bar{X}_1 - 2 \frac{\sigma}{\sqrt{n}} = 1,658 \; ; \quad \bar{X}_1 + 2 \frac{\sigma}{\sqrt{n}} = 1,692$$

El intervalo de confianza es (1,658, 1,692).

El intervalo de confianza al 99 por 100 sería:

$$\left(\bar{X}_1 - 3 \frac{\sigma}{\sqrt{n}}, \; \bar{X}_1 + 3 \frac{\sigma}{\sqrt{n}} \right)$$