Curso de

Estadística para muestristas

Profesora: Dra. Clotilde Bula

agosto/diciembre 1955

REPUBLICA ARGENTINA

SERVICIO ESTADISTICO NACIONAL

CURSO DE

ESTADISTICA PARA MUESTRISTAS

Profesora: Dra. CLOTILDE BULA

agosto - diciembre 1955

CENTRO DE CAPACITACION TECNICA PARA FUNCIONARIOS ESTADISTICOS

Programa de Estadística para Muestristas

Prof.:Dra. Clotilde Bula

Bol. 1: Estadística. Definición-Concepto de variable y de atributo-Concepto de función estadística- Las series estadísticas-Distribuciones de frecuencia

Bol. 2: Representaciones gráficas.Finalidades de las representaciones gráficas.Representación de un fenómeno en función de una o dos variables.Diagramas.Histogramas.

Bol. 3: Estadísticas de variables-Series de frecuencia.Variable.Intervalo de clase-Distribución de frecuencias.Representación gráfica de las distribuciones de frecuencia.Curvas de frecuencia. Métodos de estudio de las distribuciones de frecuencias.

Bol. 4: Promedios.Clasificación. edia aritmética, geométrica,armónica, cuadrática. Definiciones, propiedades y cálculo,

Bol. 5: Medidas de posición. Modo.Mediana. Deciles.Cuartiles.Percentiles.Definiciones,propiedades,cálculo y determinacion gráfica.

Bol. 6: Medidas de variabilidad y de dispersión.Generalidades.Variación. Medidas de variación absoluta.Oscilación.Desviación media. Concepto. Cálculo.

Bol. 7: Desviación típica o desvío standard.Varianza.Conceptos.Cálculo Cálculos abreviados.Desviación cuartil.Error probable.Relaciones entre las diferentes medidas de variabilidad.

Bol. 8: Medidas de variabilidad relativa.Coeficiente de variabilidad. Medidas de asimetría.Curtosis.

Bol. 9: Momentos. Momentos de las distribuciones de frecuencia.Momentos respecto de un origen arbitrario. Momentos con origen en la media aritméticaCorrecciones de Sheppard.

Bol.10: Probabilidad.Definición.Teoremas fundamentales.Variable aleatoria.Esperanza matemática.

Bol.11: Distribuciones de probabilidad. Distribución binomial. Distribución de Poisson.Distribución normal. Uso de tablas.

Bol.12: Noción de muestra. Parametros y estadísticas. Distribución de las estadísticas. Estimación

Bol.13: Intervalo de confianza. Coeficiente de confianza.Regresión lineal. Mínimos cuadrados.

Bol.14: Elementos del analisis de la varianza.

ESTADISTICA CURSO " A " 1a. 2a. y 3a. Conferencia
Prof. Dra. Clotilte Bula

ESTADISTICA

El término Estadística tiene dos sentidos.

En el lenguaje común se usa generalmente como sinónimo del término " dato ". Así por ejemplo, se dice que hemos visto " estadísticas de cereales cosechados ". Se hablaría con más precisión, en este caso, si se dijera que hemos visto "datos de cereales cosechados".

El término " estadística " se aplica también a los principios y métodos científicos de descripción y análisis de datos numéricos.

El estudio de la Estadística es el estudio de los métodos usados para coleccionar, presentar, analizar e interpretar datos cuantitativos y cualitativos.

El estudio de la estadística agropecuaria, por ejemplo, consiste en un examen de aquellos métodos que pueden ser especialmente usados en el tratamiento de datos relacionados con las actividades agrícolas y ganaderas.

Los métodos de análisis estadísticos se emplean en muchos campos del conocimiento : biología, psicología, bacteriología, educación, etc., pero su uso se ha extendido más rápidamente, en los últimos años, al campo de la Economía y el Comercio.

El objeto del método estadístico son los hechos o fenómenos colectivos, porque o es la misma colectividad lo que nos interesa o bien una muestra de ella que puede conducir al conocimiento de la característica deseada.

Reunidas las unidades estadísticas debe realizarse el despojo para obtener aquellos datos que constituyen el fin del relevamiento. " Despojar " una unidad significa " repartirla " en grupos según la calidad o la intensidad de los caracteres o de las circunstancias que para cada una de estas unidades fueron determinadas.

Se trata entonces de formar, sobre la base de un esquema de despojo, tantos grupos de unidades como modalidades " cualitativas " o " cuantitativas " se desean estudiar.

Un carácter se llama " cualitativo " cuando no puede ser indicado más que con expresiones verbales. Por ejemplo, la nacionalidad es un carácter cualitativo porque se expresa por la voz derivada del país de origen de los individuos considerados.

Un carácter se llama " Cuantitativo " cuando puede ser expresado por cifras. Por ejemplo, las estaturas, las edades, etc.

El carácter cuantitativo puede variar de modo : 1) continuo y 2) discontinuo.

Varía de modo " continuo " cuando al pasar de un valor x a un valor x + 1 puede tomar todos los infinitos valores intermedios.

Varía en forma "discontínua" cuando procede por saltos.

Son de variación continua : la estatura, el peso, la edad, ya que fijados dos valores para cada uno de ellos, se puede concebir la existencia de individuos que posean uno cualquiera de los infinitos valores intermediarios.

Son de variación discontinua: la población, la fecundidad, medida por el número de hijos, porque crecen o decrecen de uno en uno, saltando todos los valores fraccionarios intermedios.

En el agrupamiento que se realiza para la presentación de una observación estadística, referida a caracteres cuantitativos, debe tenerse en cuenta las necesidades de la investigación que posteriormente debe realizarse.

Conviene realizar el despojo de las unidades en clases más o menos amplias. La amplitud de estas clases se llama " módulo ".

Cuando el módulo tiene siempre la misma amplitud, se dice que la observación está presentada en una serie " monomodular " y se llama " plurimodular " cuando en la clasificación se emplean módulos de distinta amplitud.

A veces resulta necesario dejar abierta la primera o la última clase de una serie, o ambas a la vez en caso que ellas contengan pocas observaciones, se dice entonces que estas clases son abiertas. Cuando el límite inferior de la primera clase y el límite superior de la última clase están perfectamente determinados se dice que ambas clases son cerradas.

Hemos dicho que uno o los caracteres observados en una investigación estadística pueden tomar una modalidad " cuantitativa " o " cualitativa ". En el primer caso se los designa con el nombre de

" variables " y en el segundo con el de "atributos ".

Los caracteres observados pueden presentarse con distinta frecuencia. Cuando cada carácter se presenta una sóla vez se dice que no hay ponderación, sobreentendimiento que ella es equivalente a la unidad. Cuando cada carácter se presenta más de una vez, se dice que la observación es ponderada.

EJEMPLOS:

Nº 1.- Exportaciones Argentinas en 1950

Productos x	Valores en millones de m$n y
de la Ganadería	2.703,4
de la Agricultura	2.321,8
Forestales	229,5
de la Minería	12,7
de la Caza y Pesca	6,1
Diversos artículos	154,1

caso de carácter " cualitativo " o " atributo "

Nº 2.- Estaturas de 300 hombres argentinos

Estatura en cms. x	Nº de hombres y
150 - 152	2
152 - 154	6
154 - 156	12
156 - 158	25
158 - 160	50
160 - 162	61
162 - 164	70
164 - 166	40
166 - 168	25
168 - 170	8
170 - 172	1
	300

El ejemplo anterior trata de observaciones de un carácter " cuantitativo " o " variable " cuyo agrupamiento se ha efectuado por clases iguales, por tanto es una presentación " monomodular " y además " cerrada " porque el primer y último intervalo tienen sus límites bien establecidos.

Nº 3 .- Clasificación por edades de 522 obreros de una fábrica

Edad en años x	Nº de obreros y
menos 18	150
18 - 21	280
21 - 25	75
25 - 35	10
35 - 45	4
45 - 55	2
más de 55	1
	522

en este ejemplo tratamos con valores de una " variable " que se ha agrupado en clases desiguales, es por tanto una presentación " plurimodular " y además la primera y última clases son abiertas porque faltan, respectivamente, los límites inferior y superior.

En los tres ejemplos se trata de observaciones " ponderadas " estando la ponderación indicada en la segunda columna de cada cuadro. El número que indica cada ponderación se designa con el nombre de " frecuencia " . Así en el ejemplo Nº 2 las estaturas comprendidas entre 160 y 162 cms. se han presentado con una " frecuencia " de 61.

CONCEPTO DE FUNCION ESTADISTICA

En matemática el concepto de " función " se expresa diciendo: Si a cada valor de una variable x se hace corresponder un valor de otra variable y, la segunda variable es " función de la primera.

La variable x se denomina " variable independiente ".

El concepto estadístico de función no es distinto.

Si a los valores que expresan la edad de los obreros hacemos corresponder el número de obreros, en el ejemplo Nº 2, el fenómeno " edad " es la variable independiente y el " número de obreros " es la función; se dice que el número de obreros varía en función de la edad o que es función de la edad.

Por analogía se extiende el significado de variable independiente a los " atributos " aún cuando no tome valores, sino modalidades diversas.

Así se puede decir que se estudia la frecuencia de los individuos en función del color de sus cabellos; los valores de las exportaciones en función del origen de los productos; la renta en función del sexo, etc.

LAS SERIES ESTADISTICAS - DISTRIBUCIONES DE FRECUENCIAS

Las sucesiones de cifras a que conduce el despojo y la sistematización tabular se designan por " series estadísticas ".

En una tabla definitiva, la primera columna, expresa los valores (o clases de valores) de una variable x o la modalidad de un atributo cualitativo.

La otra o las otras columnas expresan las frecuencias o las intensidades de un fenómeno y considerado como función de la variable x o atributo respectivo.

Ej. Producción de Tártaro en toneladas

Años	Mendoza	San Juan	Otras regiones	Total
(1)	(2)	(3)	(4)	(5)
1933	80,0	45,0	25,0	150,0
1934	78,0	54,6	25,0	148,0
1935	94,4	45,0	41,0	200,0
1936	123,0	82,0	45,0	250,0
1937	139,1	92,8	58,1	290,0

tenemos así, cuatro series estadísticas, relacionando la primera columna con cada una de las otras cuatro que indican frecuencias.

Los resultados de la observación de un carácter cuantitativo (variable) dan origen a dos tipos de series estadísticas;
1) las series cronológicas, históricas o temporales cuando el fenómeno estudiado en función del tiempo (ejemplo anterior);
2) las distribuciones de frecuencias que son el resultado de la observación, en un momento dado, de una masa de unidades distribuída según la modalidad cuantitativa de otro fenómeno (ejemplos Nº 2 y 3).-

En el curso que nos ocupa sólo trataremos de las distribuciones de frecuencias.

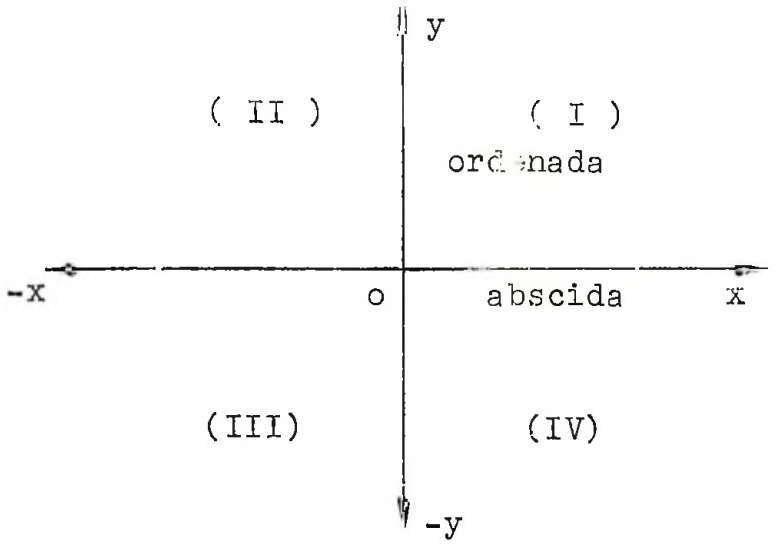
REPRESENTACION GRAFICA DE LAS DISTRIBUCIONES DE FRECUENCIAS

Los resultados del despojo, además de presentarse en tablas y cuadros numéricos, pueden ser expresados por representaciones gráficas. Estas tienen la ventaja de dar una visión rápida y suficientemente exacta del movimiento del fenómeno aún cuando esté expresado por series muy largas. Facilitan también la comparación de fenómenos diversos, en otras palabras, las representaciones gráficas no sólo tienen una finalidad propia, sino que constituyen un óptimo instrumento de investigación.

La representación de una distribución de frecuencias pue

de hacerse gráficamente por medio de segmentos o de superficies. Habitualmente se emplea como referencia para el gráfico un sistema de coordenadas cartesianas.

Sobre el eje horizontal (eje de las x) se mide la intensidad de la variable independiente y sobre el eje vertical (eje de las y) la intensidad de la función (frecuencia)



El plano queda dividido en cuatro cuadrantes de los cuales el primero es usado para valores positivos de la variable y de la función; el segundo para valores negativos de la variable y positivos de la función; el tercero para valores negativos de la variable y negativos de la función y finalmente, el cuarto para valores positivos de la variable y negativos de la función.

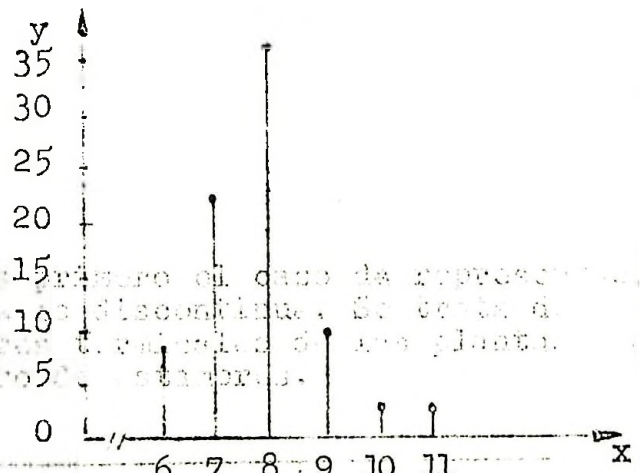
En el estudio de fenómenos colectivos, naturales o sociales, que son objeto de la investigación estadística, es difícil que se estudien fenómenos en función de variables con valores negativos y es más difícil aún que se estudien funciones con valores negativos. Por esto resulta casi siempre usado el primer cuadrante.

No faltan, sin embargo, en Estadística, ejemplos de variables que pueden tomar valores negativos. Se puede, por ejemplo, clasificar un grupo de mediciones o estimaciones, según la amplitud del " error " que lo perturba. En este caso, la variable (intensidad del error) puede tomar valores positivos o negativos.

Clasificando un grupo de personas, según el patrimonio poseído, no puede excluirse que se hallen también individuos con patrimonios negativos, gravados con un pasivo que supera al activo. En este caso se usarán los cuadrantes 1º y 2º.

Veremos primero el caso de representación cuando la variable independiente es discontinua. Se trata de una observación efectuada en las flores terminales de las plantas dedaleras mediante el recuento de número de estambres.

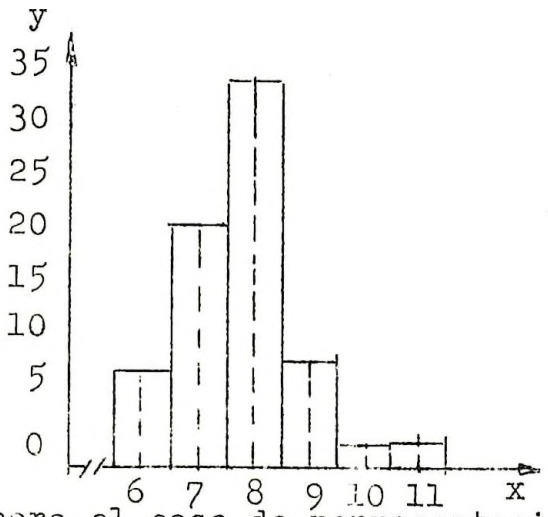
Nº de estambres x	Nº de plantas y
6	7
7	22
8	35
9	8
10	2
11	2
	76



En este caso se trazan simplemente, las ordenadas en correspondencia con los valores de las variables.

No corresponde en este caso unir los extremos superiores de las ordenadas entre sí, por tratarse de variable discontinua; sin embargo a veces para tener una visión más clara de las variaciones suelen ser unidos.

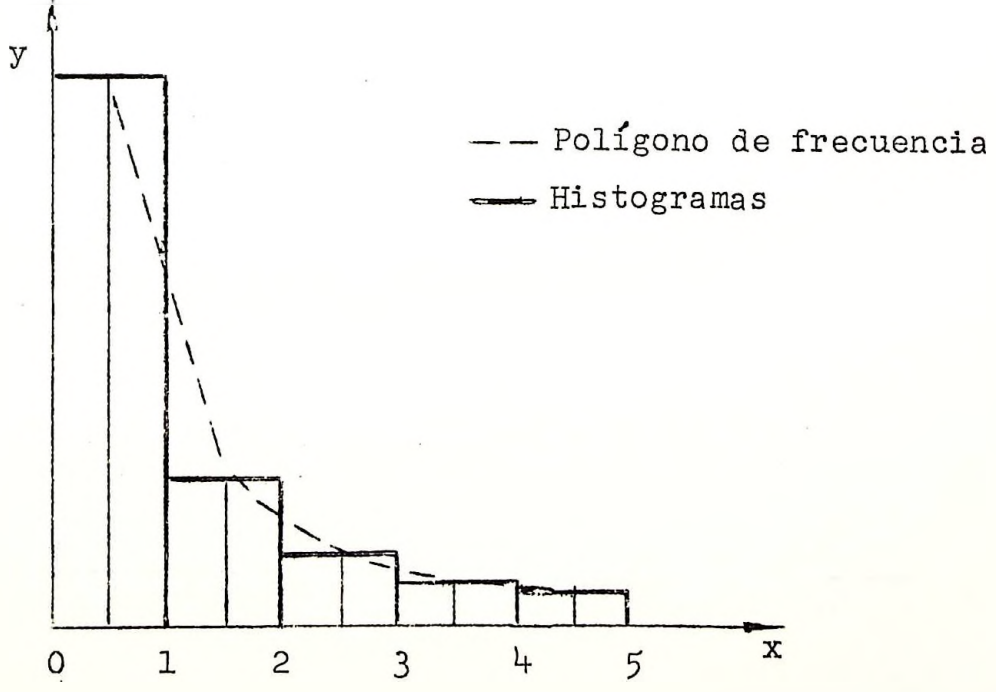
También puede representarse las frecuencias, en este caso, por rectángulos de base igual a la unidad y con su altura indicando las respectivas frecuencias, como puede verse a continuación:



Veremos ahora el caso de representación de observaciones en las cuales la variable independiente es contínua

Ej. Mortalidad infantil en Rosario

Edades x	Nº de muertos y
1 año	736
2 años	182
3 años	92
4 años	58
5 años	43



Como el intervalo de variación es de año en año (edades) levantamos una ordenada que represente las frecuencias respectivas en el punto medio de cada intervalo. El diagrama resultante puede completarse de dos maneras:

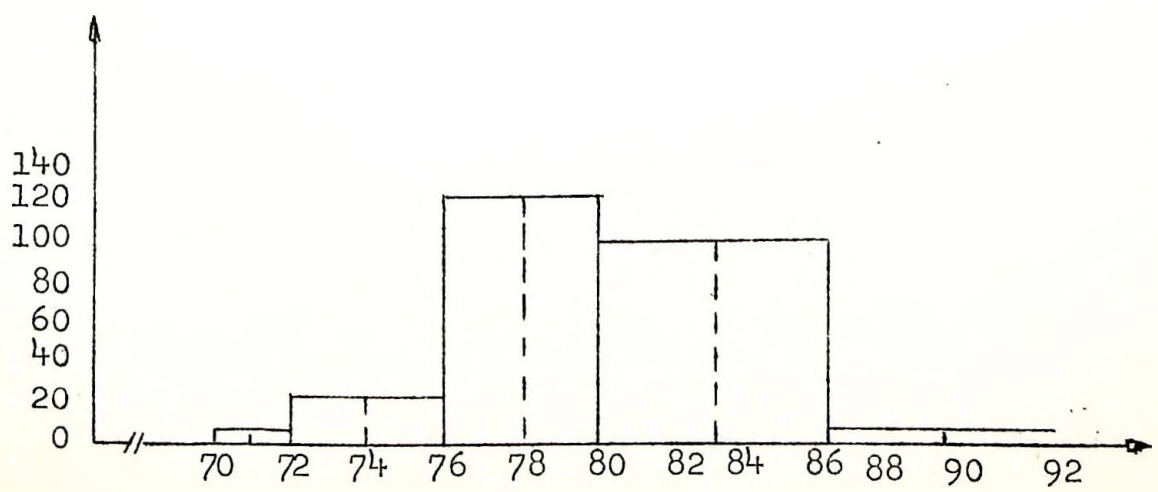
I) Como " polígono de frecuencias " trazando rectas que unen entre sí los puntos marcados en las verticales.

II) Como "histograma" o diagrama de columnas, trazando segmentos horizontales de recta por los puntos marcados en las verticales, líneas estas que forman ahora los ejes centrales de una serie de rectángulos representativos de las frecuencias.

Por lo general es más exacta la representación por histogramas que por polígonos de frecuencias. Veamos ahora un ejemplo en el que los intervalos de clase no son uniformes, es decir que tenemos una distribución plurimodular:

Indices cefálicos medidos entre niñas de 7 a 15 años

Indice cefálico en cms. x		Nº de niñas y	
70 - 72	(2)	7	(7)
72 - 76	(4)	52	(26)
76 - 80	(4)	254	(127)
80 - 86	(6)	303	(101)
86 - 92	(6)	26	(7)
		642	



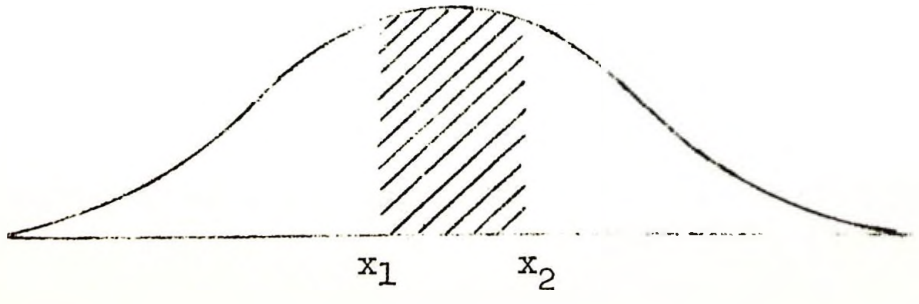
Observando el ejemplo vemos que existen tres clases de módulos distintos. El primero de amplitud 2, los dos siguientes de amplitud 4 y los dos últimos de amplitud 6. Si consideramos el primer intervalo como unitario, los dos siguientes representan el doble y los dos últimos el triple de amplitud. Para que los rectángulos representen las frecuencias exactamente, deben ser tales que su superficie coincida con ellas. Para obtener tal resultado es necesario que las frecuencias de los dos segundos grupos sean divididas por dos (base del rectángulo que las contendrá) y sus valores así obtenidos representarlos como ordenadas levantadas en los puntos medios de los intervalos dados. En cuanto a los dos últimos módulos, por ser triples que el primero, sus frecuencias para ser representadas deberán dividirse por tres y proceder luego como en el caso anterior. La primer columna de números entre paréntesis del cuadro anterior indican las amplitudes de los módulos, y la segunda columna el resultado de haber dividido las frecuencias por 2 y por 3 según el caso.

CURVAS DE FRECUENCIAS

Si los intervalos de clase se hiciesen cada vez más pequeños y se aumentase al mismo tiempo el número de observaciones de modo que las frecuencias de clase permanecieran finitas, tanto el polígono de clase, como el histograma, se aproximarían cada vez más a una línea curva. Este límite ideal del polígono y del histograma, recibe el nombre de " curva de frecuencia " y es un concepto de máxima importancia en Estadística.

En la curva de frecuencia el área comprendida entre dos ordenadas cualesquiera es proporcional al número de observaciones incluídas entre los valores correspondientes de la variable.

Así el número de observaciones comprendidos entre los valores x_1 y x_2 de la variable, del gráfico adjunto, será proporcional al área de la zona rayada en la figura.



El número de valores observados, mayores que x_2 vendrá dado por el área de la curva situada a la derecha de la ordenada trazada por x_2 y así sucesivamente.

TIPOS CORRIENTES DE DISTRIBUCIONES DE FRECUENCIAS

Las formas que presentan las series de datos son casi infinitas en su variedad, pero ofrecen un número relativamente pequeño de tipos simples.

Mediante estos tipos simples pueden hacerse, muchas veces el análisis de distribuciones más complicadas.

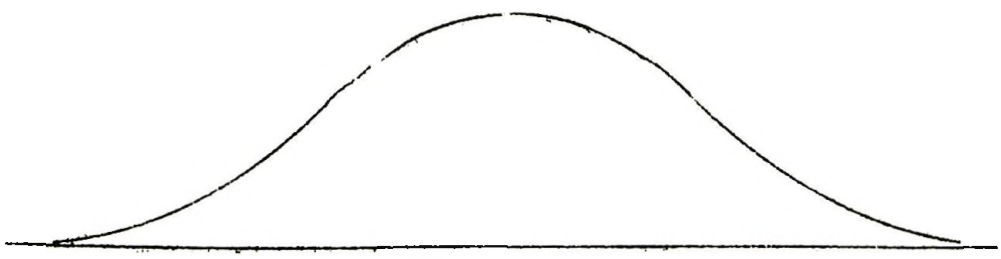
Los tipos básicos son cuatro:

1.- distribución simétrica
2.- distribución moderadamente asimétrica
3.- distribución muy asimétrica o en forma de J
4.- distribución en forma de U

1.- Distribución simétrica

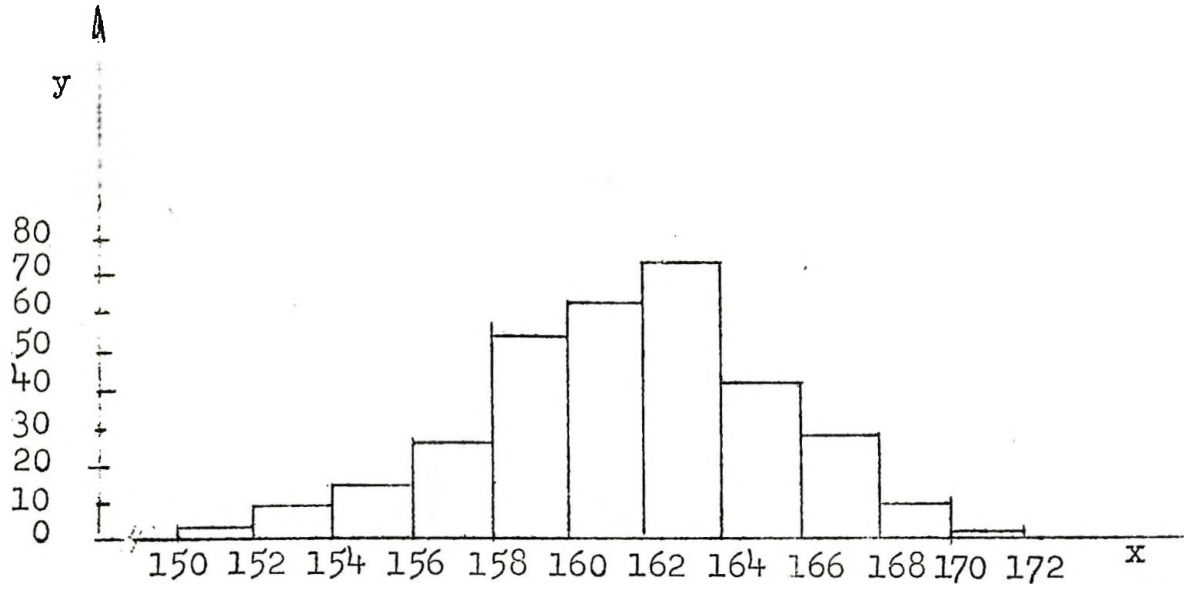
En este tipo de frecuencias de clase van disminuyendo hasta cero, de manera simétrica, a uno y otro lado de un valor máximo central.

El gráfico siguiente da una idea de la forma ideal de esta distribución:



Esta forma simétrica pura es relativamente poco frecuente. Las series biológicas y antropométricas, especialmente las que contienen medidas lineales, tales como alturas, más bien que medidas de dos o tres dimensiones como la circunferencia toráxica o el peso, conducen a distribuciones de frecuencias que son casi simétricas.

Representando los datos en el ejemplo Nº 2 tenemos:



2.- Distirbución moderadamente asimétrica

En este caso las frecuencias de clase disminuyen, con rapidez, notablemente mayor hacia uno de los lados del valor máximo que al otro como ocurre en los casos representados a continuación:

Estas son las formas más corrientes de distribuciones de frecuencias, presentándose ejemplos de ellas en estadísticas de las más diversas fuentes.

Las distribuciones de frecuencias que se estudian en las ciencias sociales son asimétricas y a menudo hacia la derecha. Sólo raramente se encuentra una distribución de frecuencias asimétricas hacia la izquierda. Las curvas asimétricas reciben también el nombre de " oblicuas ".

Hemos de considerar, oportunamente, la asimetría con mayor detenimiento y examinaremos diversos métodos para medirla. en particular veremos que la asimetría tiene un signo y podemos ya indicar

que la asimetría se llama " positiva " si la rama más larga de la curva está a la derecha y " negativa " si está a la izquierda.

Ejemplos : IV Censo Escolar

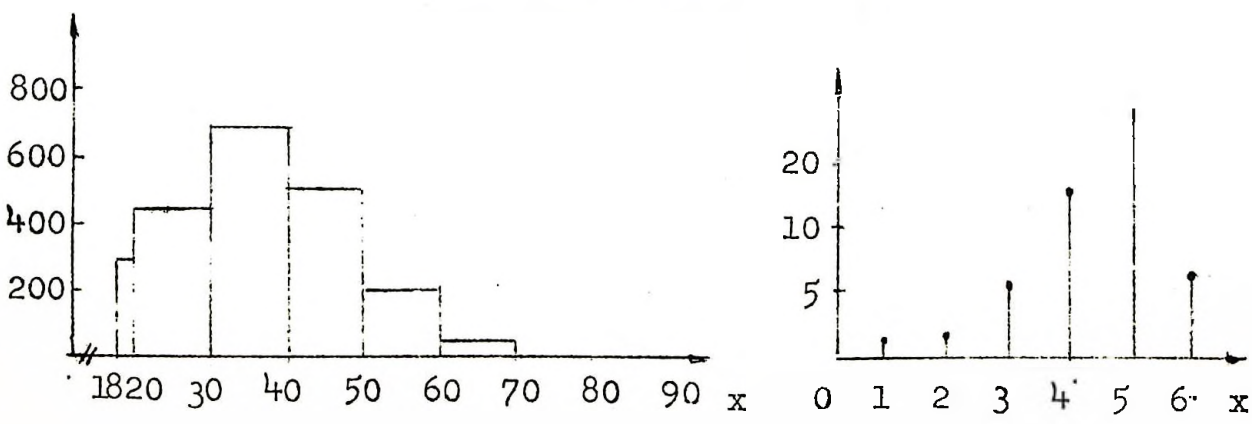
Edad de las madres con hijos de menos de 22 años x	Nº de madres en miles y		
18 - 20	67	x(5)	335
20 - 30	441		
30 - 40	717		
40 - 50	504		
50 - 60	215		
60 - 70	42		
70 - 90	7	:(2)	3,5

es caso de asimetría positiva.

Ejemplo de asimetría negativa:

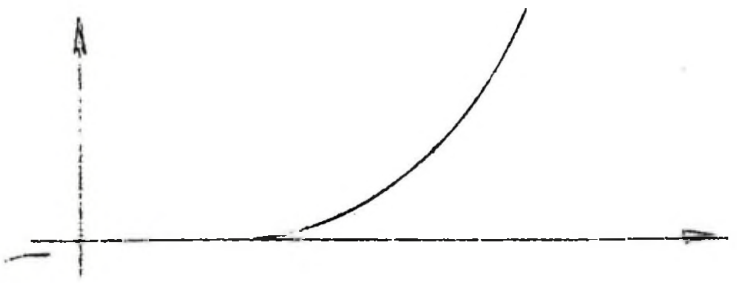
Poda de viñedos

Posición de la yema	Racimos por yema
x	y
1	1
2	1
3	5
4	13
5	19
6	7



3.- Distribución en forma de "J"

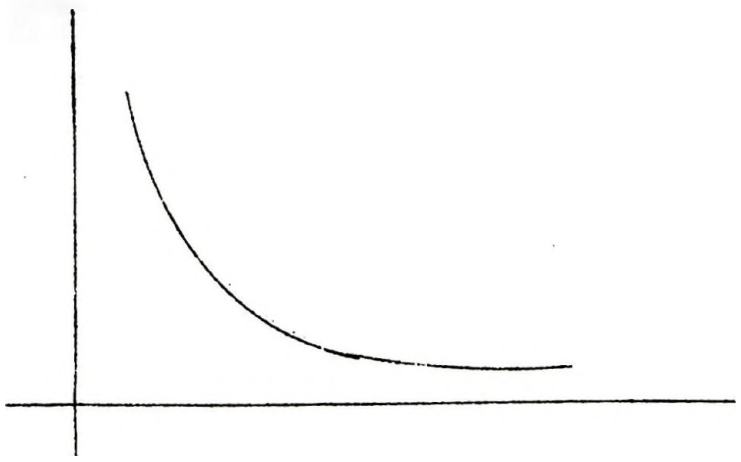
En este tipo de distribución las frecuencias de clase van creciendo hasta un valor máximo situado en uno de los extremos del campo de variación, como en la figura que sigue:



Puede considerarse como una forma límite del tipo anterior y en realidad no siempre pueden distinguirse entre sí las distribuciones de uno y otro tipo por métodos elementales si no se dispone de los datos primitivos.

En estadística económica, esta forma es especialmente característica de la distribución de la riqueza en la población global, como lo demuestran los datos del impuesto sobre la renta y del valor de las propiedades.

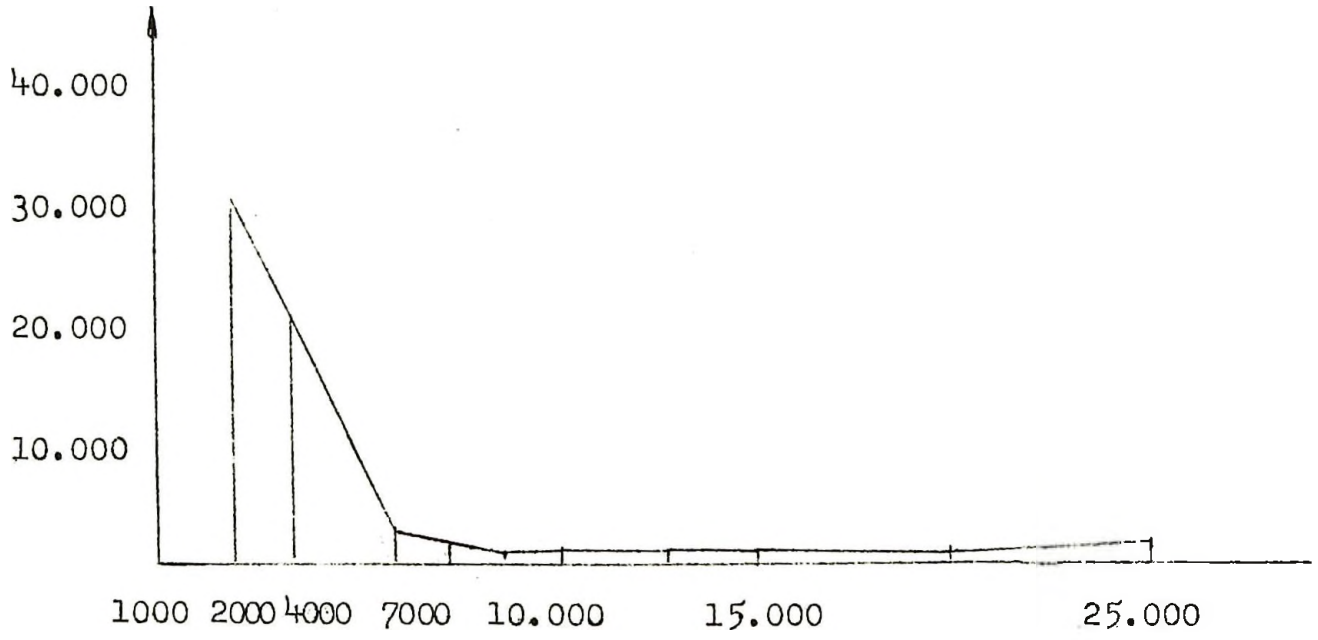
La curva a que el fenómeno de las rentas da lugar se llama "curva de Pareto" en honor a Vilfredo Pareto que llamó la atención de los economistas sobre ella y semeja una "J" a la inversa.



Ejemplos:

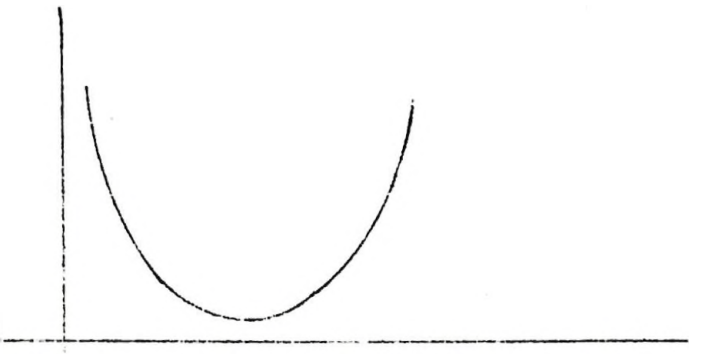
Rentas globales

Rentas en m$n x			Nº de contribuyentes y		
1.000	a	2.000	32.518		32.518
2.000	a	4.000	47.202	: (2)	23.601
4.000	a	7.000	5.502	: (3)	1.834
7.000	a	10.000	1.866	: (3)	622
10.000	a	15.000	1.085	: (5)	217
15.000	a	25.000	670	: (10)	67
más	de	25.000	645		



4.- Distribución en forma de "U"

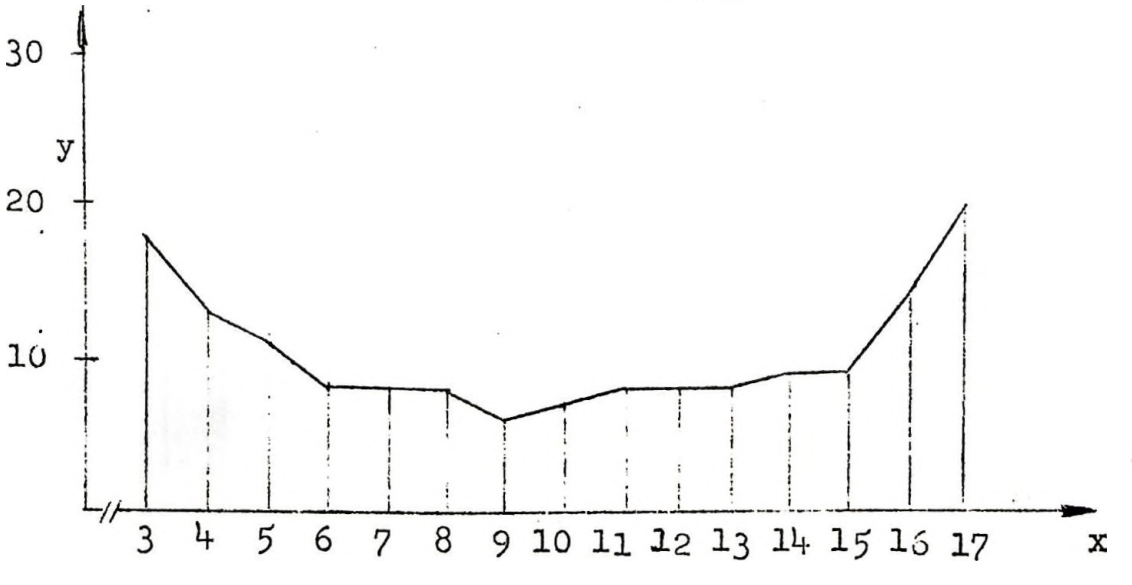
Este tipo de distribución presenta los valores de frecuencias máximas en los extremos del campo de variación y un mínimo hacia el centro, como se ve en la figura siguiente:



Se trata de una forma de distribución rara pero interesante por el contraste que ofrece con las anteriores.

Ejemplo: Mortalidad entre 3 y 17 años en Rosario (1926)

Edad x	Nº de muertos y	Edad x	Nº de muertos y
3	18	11	8
4	13	12	8
5	11	13	8
6	8	14	9
7	8	15	9
8	8	16	14
9	6	17	20
10	7		

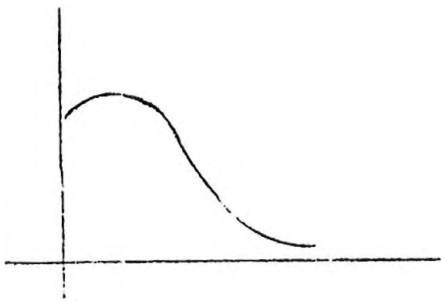


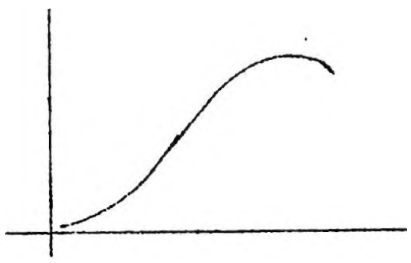
5.- Otras formas de las distribuciones de frecuencias.

a) Formas truncadas:

Los cuatro tipos que hemos considerado se presentan, a veces, en forma incompleta.

Ciertas limitaciones en el campo de variación pueden originar una especie de truncamiento en uno u otro de los extremos:



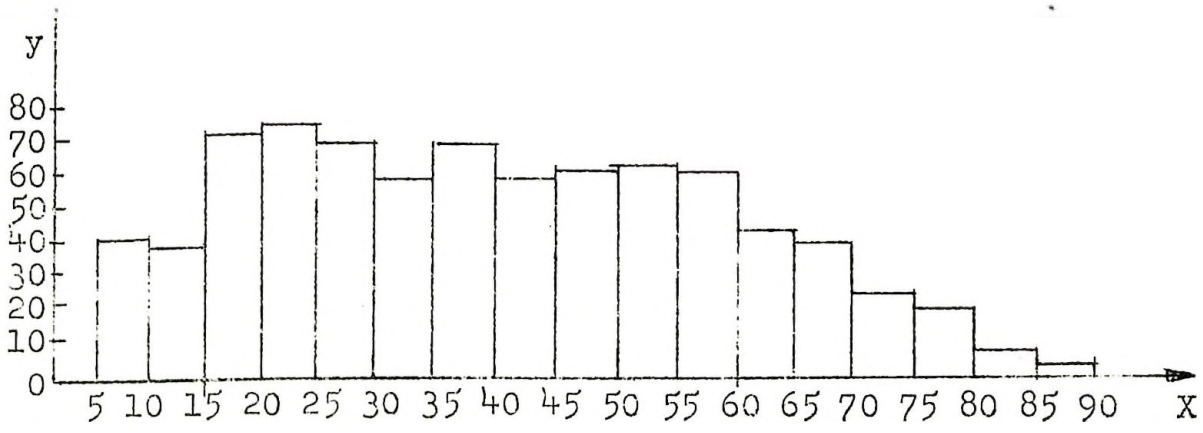


No es raro que, en especial, las variables discontinuas den lugar a irregularidades de esta clase.

b) Distribuciones complejas:

La tabla que sigue nos da el número de. defunciones de varones argentinos en Rosario (1926) por grupos de edades.

Edades x	Defunciones y	Edades x	Defunciones y
0 - 5	687	45 - 50	58
5 -10	40	50 - 55	60
10 -15	38	55 - 60	58
15 -20	72	60 - 65	41
20 -25	74	65 - 70	39
25 -30	66	70 - 75	21
30 -35	57	75 - 80	17
35 -40	65	80 - 85	3
40 -45	55	85 - (0	3



Como se ve en el histograma, la distribución tiene cuatro máximos, en los grupos de edades: 0 - 5; 20 - 25; 35 - 40 ; y 50 - 55.

Sin detenernos en el detalle de esta curva de mortalidad, podemos notar que la elevada frecuencia del grupo inicial se debe, indudablemente al fuerte coeficiente de mortalidad infantil.

El máximo para las edades 20 - 25 podría atribuirse a la tuberculosis; el que corresponde a los 35 - 40 años al cáncer y para 50 - 55 a la arterioesclerosis.

Podemos, por otra parte, considerar a la distribución como formada por la combinación de varias distribuciones: una en "J" inversa para los primeros años, otra truncada desde los 10 a los 30; la que sigue asimétrica desde los 30 a los 40 y finalmente otra truncada desde los 40 hasta las edades extremas.

Este es un ejemplo del hecho que una distribución compleja puede a veces, ser descompuesta en otras más simples. En particular este tipo de análisis tiene importancia en los estudios acutuariales y en investigaciones sobre causas de muerte.

METODOS DE ESTUDIO DE LAS DISTRIBUCIONES DE FRECUENCIAS

La clasificación de los datos cuantitativos y la elaboración de una distribución de frecuencias constituyen una de las etapas más importantes de los trabajos de elaboración y análisis.

Por medio de la clasificación puede ser revelada la estructura íntima de los datos y puesto de manifiesto la armonía que es esencial exista en toda masa de materiales.

Pero todo esto es tan sólo el primer paso en el análisis estadístico, restando aún encontrar métodos propios para medir y expresar más concisamente las características más importantes de un conjunto de datos.

Para ciertos propósitos, la propia distribución de frecuencias, debe resumirse y condensarse, destilando su esencia, por decirlo así, hasta dejarla reducida a tres o cuatro números expresivos.

Si cada distribución de frecuencias constituyera un problema nuevo y único, que obedeciera a leyes particulares en cada caso, presentaría su estudio enorme dificultad, pero, afortunadamente no ocurre así.

Los datos cuantitativos, aunque procedan de campos diferentes, al agruparse en las distribuciones de frecuencias, muestran algunas características comunes que obedecen a ciertas leyes generales. La experiencia adquirida en uno cualquiera de los campos de procedencia de los datos, constituirá, por lo tanto, una guía segura para los trabajos que se realicen en los demás.

La uniformidad en la manera de conducir las masas de datos hace posible el desarrollo de un método general para ordenar, analizar y comparar medidas realizadas en los campos más diversos de la investigación científica.

El primer paso para hacer inteligible una serie de observaciones consiste en resumir los datos de la respectiva distribución de frecuencias. Pero esto no basta en la práctica, especialmente cuando se deben comparar dos o más series.

Tenemos que avanzar más para poder definir cuantitativamente las características de una distribución de frecuencias mediante unos cuantos números, tan pocos como sea posible.

Debemos recordar que, generalmente, las distribuciones procedentes de materiales semejantes presentan formas parecidas y el empleo práctico de las diversas medidas estadísticas que examinaremos ahora se basan, precisamente, en este hecho.

Ya hemos visto como se distribuyen, a lo largo de una escala de valores correspondientes a la totalidad de las observaciones, los valores de las frecuencias.

La distribución de frecuencias puede estudiarse por la elección de un solo valor de la escala que represente perfectamente el total de la distribución y puesto que las frecuencias varían, es evidente que debe elegirse aquel valor que se presente mayor número de veces o en otros términos el valor de la variable para el cual tiene lugar la concentración máxima. Este valor constituye una " medida de tendencia central " de la distribución. Así el grupo de renta al que corresponde el mayor número de rentistas sirve para representar la distribución.

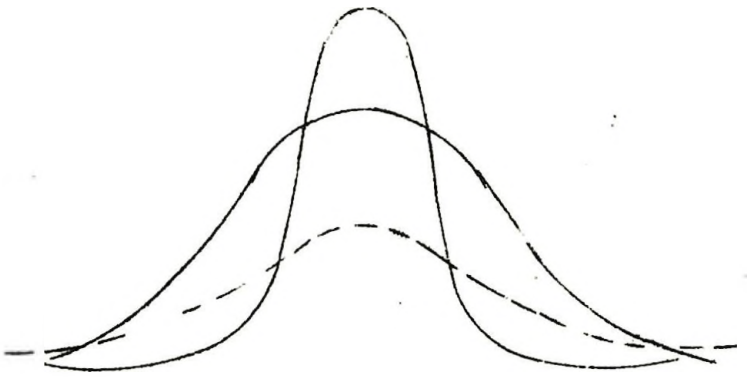
Se observa que este valor más corriente, es tan solo una de las diversas medidas posibles de la tendencia central de una distribución dada.

Estas medidas según sus características particulares se clasifican en " promedios"y " medidas de posición".

Un valor representativo singular, tal como el citado, tiene muchas aplicaciones, pero no muestra, por sí mismo, diversos aspectos que se relacionan con la distribución.

Reviste gran importancia el carácter de la distribución alrededor del promedio pues puede ocurrir que los casos tabulados se concentren apretadamente o se dispersen ampliamente.

Por ejemplo



La propiedad representativa de un promedio depende de como se ajustan a él estrechamente los demás valores, o sea del grado de concentración alrededor de la tendencia central. Por tanto, debe complementarse el promedio con una " medida de variabilidad " o de " dispersión " alrededor del valor central.

Deberá estudiarse detenidamente, el grado de simetría que la distribución presenta, porque es sumamente importante conocer si existe igual distribución de casos a cada lado del punto de máxima concentración o si la curva se extiende hacia un lado más que hacia el otro. Cuando la curva no sea simétrica, habrá que determinar su grado de asimetría y para ello se emplean las " medidas de asimetría ".

Es posible también medir el grado de agudeza de las curvas de frecuencias. Esta característica se denomina " kurtosis " y su medida constituye la etapa final en el estudio de una distribución de frecuencias.

Una vez realizadas todas estas medidas, está encauzada la labor del análisis estadístico, pues la masa de datos originales se han destilado produciendo tres o cuatro medidas de gran significación.

Estas operaciones no sólo nos revelan las características de la distribución dado sino que facilitan las comparaciones con otras análogas.

Sería, por ejemplo, imposible efectuar la comparación de las cifras no ordenadas de las rentas de 10.000 rentistas argentinos con datos análogos de Estados Unidos.

Pero si hallamos un valor de renta promedio o valor más representativo para cada nación y realizamos un estudio de las distribuciones de las rentas alrededor de cada valor central, habremos encontrado una base sólida en que apoyar nuetro estudio comparativo.

Llamaremos valores medios, en general, a aquellos deducidos por cálculos sencillos, de primera aproximación, en el menor número posible, pero con la mayor significación, que fuesen suficientes para representar lo que hubiere de más importante en la distribución considerada.

Est. "A"

4º y 5º. Conferencia

PROMEDIOS Y MEDIDAS DE PRECISION

De cualquier modo que definamos un promedio, debe tenerse muy presente que es siempre un valor de la variable y que por tanto, tendrá necesariamente las mismas dimensiones que ella, es decir, si la variable es una longitud, su promedio será una longitud; si la variable es un porcentaje, su promedio será un porcentaje.

Pero hay diferentes maneras de definir la posición de una distribución de frecuencias, esto es, hay diferentes formas de promedios y surgen por tanto estas preguntas: Con qué criterio hemos de juzgar los méritos relativos de las diferentes formas de promedios? Cuáles son, en realidad, las propiedades que deberá reunir un buen promedio?

a) En primer lugar un promedio debe estar rigurosamente definido y no ser susceptible de distintas interpretaciones. Un promedio que fuese simplemente una estimación dependería tanto del observador como de los datos.

b) Un promedio debe basarse en todas las observaciones efectuadas. En caso contrario, no sería en realidad una característica de toda la distribución.

c) El promedio debe ser sencillo y claro, es decir que no debe tener un caracter matemático demasiado abstracto.

d) Es importante que el promedio elegido se preste fácilmente al cálculo algebraico. Si por ejemplo se dan dos o más series de observaciones procedentes de materiales análogos, el promedio de las series combinadas debe poder expresarse fácilmente en función de los promedios de las series componentes. Si una variable puede expresarse como suma de dos o más, el promedio del conjunto debe poder expresarse en función de los promedios de los respectivos sumandos. Una medida que no permita establecer simples relaciones de esta clase es casi seguro que tiene aplicación muy limitada.

CLASIFICACION DE LOS PROMEDIOS

Los promedios se clasifican en dos grandes grupos:

I.- Aquellos que dependen del orden de los elementos. Permiten comparar tantos conjuntos de elementos cualitativos como cuantitativos. Se los designa con el nombre de "medidas de posición".

II.- Los que dependen de la magnitud de sus elementos, Se los designa con el nombre general de promedios. Opina Lucien March que estos valores permiten comparaciones más precisas que las medidas de posición, cuando los elementos son cuantitativos.

Son medidas de posición: el modo, la mediana, los cuartilos decilos y centilos

Son promedios: la media aritmética, la media geométrica, la media armónica y las medias de precisión (cuagrática, cúbica, cuártica, etc.)

MEDIA ARITMETICA

Concepto: La media aritmética de una serie de valores de una variable $X_1, X_2, X_3, \ldots\ldots X_N$, en número de N es el cociente de dividir la suma de sus valores por el número de las variables consideradas.

Llamando $\bar{X}$ a la media aritmética, resulta:

$$\bar{X} = \frac{X_1 + X_2 + \ldots\ldots + X_N}{N} = 1/N \sum_{i=1}^{N} X_i \qquad (1)$$

Se utiliza también el término "promedio" ó "media" para designar este valor.

En el caso de distribuciones de frecuencias la media aritmética se define de igual forma pero en su expresión matemática debemos poner en evidencia la frecuencia (ponderación) que corresponde a cada valor de la variable, en la observación.

Así si tuviésemos los valores de las frecuencias y_i en correspondencia con los valores X_i de la variable.

X_i	y_i	$X_i\ y_i$
X_1	y_1	$X_1\ y_1$
X_2	y_2	$X_2\ y_2$
X_3	y_3	$X_3\ y_3$
....	....	.....
X_N	y_N	$X_N\ y_N$
$\sum_{i\ 1}^{N} y_i$		$\sum_{i=1}^{N} X_i\ y_i$

Por definición resulta:

$$\bar{X} = \frac{X_1\ y_1 + X_2\ y_2 + \ldots\ldots + X_N\ y_N}{y_1 + y_2 + \ldots\ldots + y_N} = \frac{\sum_{i=1}^{N} X_i\ y_i}{\sum_{i=1}^{N} y_i} \qquad (2)$$

Ejemplos: Nº 1 (Caso sin ponderación)

X
6
8
9
11
14
48

$$X = \frac{48}{5} = 9,6$$

Ejemplo : Nº 2 (Caso ponderado)

X	y	X_y
3	2	6
10	15	150
14	11	154
40	2	80
	30	390

$$\bar{X} = \frac{390}{30} = 13$$

Propiedades de la media aritmética:

1.- La suma de los desvíos a la media aritmética es igual a cero.

Vamos a demostrar que, en el caso sin ponderación, es

$$\sum_{i=1}^{N} (\bar{X} - X_i) = \sum_{i=1}^{N} x_i = 0$$

Desarrollando el paréntesis:

$$N\bar{X} - \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} x_i$$

y en virtud de (1)

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} X_i - \sum_{i=1}^{N} X_i = 0$$

En el caso ponderado, deberá verificarse:

$$\sum_{i=1}^{N} (X - X_i)y_i = \sum_{i=1}^{N} x_i y_i = 0$$

Desarrollando el primer miembro

$$\bar{X} \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} X_i y_i = 0 \qquad (3)$$

En virtud de (2) resulta

$$\sum_{i=1}^{N} X_i y_i - \bar{X} \sum_{i=1}^{N} y_i$$

Por tanto (3) puede escribirse

$$\sum_{i=1}^{N} x_i y_i = \bar{X} \sum_{i=1}^{N} y_i - \bar{X} \sum_{i=1}^{N} y_i = 0 \qquad (4)$$

que es lo que queríamos demostrar.

Ejemplo: Nº 1

X_i	$x_i \quad \bar{X} - X_i$
6	3.6
8	1.6
9	0.6
11	- 1.4
14	- 4.4
48	- 5.8
	- 5.8
	0

$$\bar{X} = \frac{48}{5} = 9.6$$

Nº 2

X_i	y_i	$X_i y_i$	$x_i = \bar{X} - X_i$	$x_i y_i$
3	2	6	10	20
10	15	150	3	45
14	11	154	-1	-11
40	2	80	- 27	-54
	30	390		65
				-65
				0

$$\bar{X} = \frac{390}{30} = 13$$

2.- La suma de los cuadrados de los desvíos, respecto de la media aritmética, es un mínimo.

Para el caso sin ponderación, demostraremos que:

$$\sum_{i=1}^{N} (\bar{X} - X_i)^2 = \text{mínimo}$$

hacemos $\bar{X} - X_i = x_i$, simbolizando como antes por x_i a los desvíos y desarrollando el cuadro anterior, se tiene:

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} X^{-2} - 2\sum_{i=1}^{N} \bar{X} X_i + \sum_{i=1}^{N} X_i^2$$

$$= N X^2 - 2 X \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} X_i^2$$

por ser en virtud de (1)

$$N \bar{X} = \sum_{i=1}^{N} X_i$$

reemplazando este valor en el segundo término del segundo miembro de la expresión anterior, podemos escribir:

$$\sum_{i=1}^{N} x_i^2 = N\bar{X}^2 - 2 \bar{X} N.\bar{X} + \sum_{i=1}^{N} X_i^2$$

reuniendo los dos primeros términos del segundo miembro, resulta

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} X_i^2 - N \bar{X}^2 \qquad (5)$$

Si en vez de trabajar con $\bar{X}$, lo hacemos con $\bar{X} \pm a$ resulta:

$$\sum_{i=1}^{N} (\bar{X} \pm a - X_i)^2 = \sum_{i=1}^{N} (\bar{X}^2 + a^2 + X_i^2 \pm 2\bar{X} a - 2 \bar{X} X_i - 2 a X_i)$$

$$= N X^{-2} + N a^2 + \sum_{i=1}^{N} X_i^2 \pm 2N \bar{X} a - 2 \bar{X} \sum_{1=1}^{N} X_i \pm 2a \sum_{i=1}^{N} X_i$$

reemplazando $\sum_{i=1}^{N} X_i = N \bar{X}$ como hicimos antes en virtud de (1) se tiene:

$$\sum_{i=1}^{N} (\bar{X} \pm a - X_i)^2 = N \bar{X}^2 + N a^2 \sum_{i=1}^{N} X_i^2 \pm 2 N \bar{X} a - 2 \bar{X} N \bar{X} \pm 2 a N \bar{X}$$

reuniendo los términos semejantes nos queda:

$$\sum_{i=1}^{N} (\bar{X} \pm a - X_i)^2 = \sum_{i=1}^{N} X_i^2 - N \bar{X}^2 a^2 N \qquad (6)$$

Comparando resulta que

$$(5) < (6)$$

puesto que en (6) aparece sumado un término que es positivo por contener a^2 multiplicado por el número de observaciones, luego la expresión (5) cumple la condición de mínimo.

Para el caso ponderado, tenemos,

$$\sum_{i=1}^{N} (\bar{X} - X_i)^2 y_i = \sum_{i=1}^{N} x_i^2 y_i = \text{mínimo}$$

desarrollando el cuadrado:

$$\sum_{i=1}^{N} x_i^2 y_i = \sum_{i=1}^{N} \bar{X}^2 y_i - 2 \sum_{i=1}^{N} \bar{X} X_i y_i + \sum_{i=1}^{N} X_i^2 y_i$$

$$\sum_{i=1}^{N} x_i^2 y_i = \bar{X}^2 \sum_{i=1}^{N} y_i - 2 \bar{X} \sum_{i=1}^{N} X_i y_i + \sum_{i=1}^{N} X_i^2 y_i$$

En virtud de (2) resulta $\bar{X} \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} X_i y_i$ (7)

reemplazando esta expresión en la fórmula anterior, se tiene:

$$\sum_{i=1}^{N} x_i^2 y_i = \bar{X}^2 \sum_{i=1}^{N} y_i - 2 \bar{X} \bar{X} \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} X_i^2 y_i$$

reuniendo los dos primeros términos del segundo miembro por ser semejantes resulta:

$$\sum_{i=1}^{N} x_i^2 y_i = \sum_{i=1}^{N} X_i^2 y_i - \bar{X}^2 \sum_{i=1}^{N} y_i \qquad (8)$$

Si como hicimos antes, trabajamos con $\bar{X} \pm a$ en vez de $\bar{X}$, tenemos:

$$\sum_{i=1}^{N} x_i^2 y_i = \sum_{i=1}^{N} (\bar{X} \pm a - X_i)^2 y_i$$

desarrollando:

$$\sum_{i=1}^{N} x_i^2 y_i = \sum_{i=1}^{N} (\bar{X}^2 + a^2 + X_i^2 \pm 2 \bar{X} a - 2 \bar{X} X_i \pm 2 X_i a) y_i$$

$$= \bar{X}^2 \sum_{i=1}^{N} y_i + a^2 \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} X_i^2 y_i \pm 2 \bar{X} a \sum_{i=1}^{N} y_i - 2 \bar{X} \sum_{i=1}^{N} X_i y_i \pm 2a \sum_{i=1}^{N} X_i y_i$$

haciendo la sustitución de valores expresada en (7)

$$\sum_{i=1}^{N} x_i^2 y_i = \bar{X}^2 \sum_{i=1}^{N} y_i + a^2 \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} X_i^2 y_i \pm 2 \bar{X} a \sum_{i=1}^{N} y_i - 2 \bar{X} \bar{X} \sum_{i=1}^{N} y_i \pm 2a \bar{X} \sum_{i=1}^{N} y_i$$

reuniendo términos semejantes, resulta:

$$\sum_{i=1}^{N} x_i^2 y_i = \sum_{i=1}^{N} X_i^2 y_i - \bar{X}^2 \sum_{i=1}^{N} y_i + a^2 \sum_{i=1}^{N} y_i \qquad (9)$$

Comparando resulta que:

$$(8) < (9)$$

luego la (8) cumple la condición de mínimo.

Ejemplo Nº 1

X_i	$x_i = \bar{X} - X_i$	x_i^2	$x_i' = X' - X_i$	$(x_i')^2$	$x_i'' = \bar{X}'' - X_i$	$(x_i'')^2$
6	3,6	12,96	3	9	4	16
8	1,6	2,56	1	1	2	4
9	0,6	0,36	0	0	1	1
11	-1,4	1,96	-2	4	-1	1
14	-4,4	19,36	-5	25	-4	16
48		37,20		39		38
$\bar{X} = \frac{48}{5} = 9,6$			$\bar{X}' = 9$		$\bar{X}'' = 10$	

Se observa cómo el valor de la suma de los desvíos al cuadrado es mínima cuando se calculan respecto de la media aritmética.

Ejemplo Nº 2

X_i	y_i	$X_i y_i$	x_i	x_i^2	$x_i^2 y_i$	x_i'	$(x_i')^2$	$(x_i')^2 y_i$	x_i''	$(x_i'')^2$	$(x_i'')^2 y_i$
3	2	6	10	100	200	9	81	162	13	169	338
10	15	150	3	9	135	2	4	60	6	36	540
14	11	154	-1	1	11	-2	4	44	2	4	44
40	2	80	-27	729	1458	-28	784	1.568	-24	576	1.152
	30	390			1804			1.834			2.074
$\bar{X} = \frac{390}{30} = 13$						$\bar{X}' = 12$			$\bar{X}'' = 16$		

En el caso ponderado ocurre lo mismo, es decir que es mínima la suma de los desvíos al cuadrado calculados desde la media aritmética.

3.- Si se suma o resta un valor constante a los valores de la variable observada, la media aritmética de los nuevos valores es igual a la media aritmética de los valores primitivos aumentada o disminuída en la cantidad constante.

Indicando por $\bar{X}'$ a la media aritmética de los nuevos valores y por h la constante que se suma o resta, se tendrá:

$$\bar{X}' = \bar{X} \pm h$$

En efecto; para el caso sin ponderación:

$$\bar{X}' = \frac{\sum_{i=1}^{N} (X_i \pm h)}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \pm \frac{Nh}{N}$$

simplificando, resulta:

$$\bar{X}' = \bar{X} \pm h$$

Para el caso ponderado:

$$\bar{X}' = \bar{X} \pm h$$

$$\bar{X}' = \frac{\sum_{i=1}^{N} (X_i \pm h) y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} X_i y_i \pm h \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i}$$

$$\bar{X}' = \frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} y_i} \pm \frac{h \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i}$$

$$\bar{X}' = \bar{X} \pm h$$

Ejemplo Nº 1

X	X' = X + 3	X'' = X - 7
6	9	-1
8	11	1
9	12	2
11	14	4
14	17	7
48	63	13
$\bar{X} = \frac{48}{5} = 9,6$	$\bar{X}' = \bar{X} + 3$	$\bar{X}'' = \bar{X} - 7$
	$\bar{X}' = \frac{63}{5} = 12,6$	$\bar{X}'' = \frac{13}{5} = 2,6$

Ejemplo Nº 2

X	y	Xy	X'=X+3	X'y	X'' = X - 14	X'' y
3	2	6	6	12	- 11	-22
10	15	150	13	195	- 4	-60
14	11	154	17	187	0	0
40	2	80	43	86	26	52
	30	390		480		-82
						52
						-30
$\bar{X} = \frac{390}{30} = 13$			$\bar{X}' = \bar{X} + 3$		$\bar{X}'' = \bar{X} - 14$	
			$\bar{X}' = \frac{480}{30} = 16$		$\bar{X}'' = \frac{-30}{30} = -1$	

CALCULO DE LA MEDIA ARITMETICA

Vamos a considerar el caso general:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i \, y_i}{\sum_{i=1}^{N} y_i}$$

que comprende como caso particular al promedio de una observación en que todos los valores de y_i se igualan a 1 ó también cuando todas las frecuencias son uniformes es decir que la ponderación es constante.

METODO BASICO

Es el que ya hemos aplicado en los ejemplos anteriores, donde para facilitar las operaciones se disponen los valores en un cuadro como el que sigue:

X_i	y_i	$X_i \, y_i$
X_1	y_1	$X_1 \, y_1$
X_2	y_2	$X_2 \, y_2$
X_3	y_3	$X_3 \, y_3$
.	.	.
.	.	.
.	.	.
X_N	y_N	$X_N \, y_N$
	$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} X_i \, y_i$

$$X = \frac{\sum_{i=1}^{N} X_i \, y_i}{\sum_{i=1}^{N} y_i}$$

METODO ABREVIADO

El cálculo de la media aritmética suele ser bastante laborioso sobre todo cuando las frecuencias son elevadas.

Puede sin embargo simplificarse notablemente mediante un cambio de origen de la variable, en virtud de la última propiedad estudiada.

Para ello se elige un valor conveniente de la variable como origen aribitrario y se calcula la media aritmética (que simbolizaremos por m) de los desvíos entre los valores de la variable y el valor elegido como origen arbitrario.

Esta media aritmética m, sumada al valor elegido como origen, da la media aritmética buscada.

La simplicidad del cálculo depende del acierto en la elección del origen arbitrario y por lo general es más conveniente el valor central del intervalo de clase al que corresponde la mayor frecuencia.

X_i	y_i	$x_i = X_i - X_j$	$x_i\, y_i$
X_1	y_1	$x_1 = X_1 - X_j$	$x_1\, y_1$
X_2	y_2	$x_2 = X_2 - X_j$	$x_2\, y_2$
X_3	y_3	$x_3 = X_3 - X_j$	$x_3\, y_3$
.	.	. . .	. .
.	.	. . .	. .
.	.	. . .	. .
x_j	y_j	$x_j = X_j - X_j$	$x_j\, y_j$
.	.	. . .	. .
.	.	. . .	. .
.	.	. . .	. .
X_N	y_N	$X_N = X_N - X_j$	$x_N\, y_N$
	$\sum_{i=1}^{N} y_i$		$\sum_{i=1}^{N} x_i\, y_i$

$$m = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} (X_i - X_j)\, y_i}{\sum_{i=1}^{N} y_i}$$

$$= \frac{\sum_{i=1}^{N} X_i\, y_i}{\sum_{i=1}^{N} y_i} - \frac{\sum_{i=1}^{N} X_j\, y_i}{\sum_{i=1}^{N} y_i}$$

$$m = \bar{X} - X_j$$

luego:

$$\bar{X} = m + X_j$$

Hasta aquí hemos considerado siempre que los intervalos de clases fuesen iguales, es decir que se trate de distribuciones monomodulares y que la amplitud del intervalo se reemplaza por su valor central a los efectos del cálculo.

En este mismo caso es aún posible simplificar más los cálculos, puesto que siendo todos los desvíos múltiplos de la amplitud

del intervalo, al dividirlos por ella, se tiene la serie natural de los números en los dos sentidos a partir de cero que corresponde al valor que se elija como origen arbitrario.

Teniendo en cuenta esta observación, no es necesario realizar cálculo alguno sino simplemente situar la serie natural de los números en ambas direcciones con signo negativo en la parte superior al valor cero y positivo en la inferior.

El resto del cálculo es idéntico al caso anterior pero se opera con valores más pequeños.

Al valor medio m así calculado hay que multiplicarlo por la amplitud del intervalo, para expresarlo en unidades iguales a las de los valores de la distribución y naturalmente sumarle el valor de la variable que se tomó como origen. El cuadro de cálculo es el siguiente:

X_i	y_i	$X_i = \frac{X_i - X_j}{h}$	$X_i\ y_i$
X_1	y_1	$X_1=(X_1 - X_j):h$	$X_1\ y_1$
X_2	y_2	$X_2=(X_2 - X_j):h$	$X_2\ y_2$
.	.	.	. .
.	.	.	. .
.	.	.	. .
X_j	y_j	$X_j=(X_j - X_j):h$	$X_j\ y_j$
.	.	.	. .
.	.	.	. .
.	.	.	. .
X_n	y_n	$X_n=(X_n - X_j):h$	$X_n\ y_n$

$$m = \frac{\sum_{i=1}^{n} x_i\ y_i}{\sum_{i=1}^{n} y_i} : h$$

$$m = (\bar{X} - X_j) : h$$

En el caso de distribuciones plurimodales el procedimiento de cálculo es similar, pero debe tenerse especial cuidado al realizar el cambio de origen.

Ejemplo: Calcular la media aritmética en el siguiente ejemplo, por los tres procedimientos explicados:

INDICES CEFALICOS		NUMEROS DE ALUMNOS	PRIMER METODO		SEGUNDO METODO		TERCER METODO	
X_i	X_i	y_i	X_i	$X_i\ y_i$	$X'_i = X_i - 81$	$X'_i\ y_i$	$X''_i = \frac{X_i - 81}{h}$	$X''_i\ y_i$
70-72	71	1	71	71	-10	-10	-5	-5
72-74	73	1	73	73	- 8	- 8	-4	-4
74-76	75	6	75	450	- 6	-36	-3	-18
76-78	77	14	77	1078	- 4	-56	-2	-28
78-80	79	21	79	1659	- 2	-42	-1	-21
80-82	81	23	81	1863	0	-152	0	-76
82-84	83	14	83	1162	2	28	1	14
84-86	85	8	85	680	4	32	2	16
86-88	87	2	87	174	6	12	3	6
88-90	89	1	89	89	8	8	4	4
90-92	91	1	91	91	10	10	5	5
		92		7390		90		45
						-62		-31

$$\bar{X} = \frac{7390}{92} = 80,326$$

$$\bar{X} = \bar{X}' + 81 = -\frac{62}{92} - 81 = -0,674 \quad 81 = 80,326$$

$$\bar{X} = h\bar{X}'' + 81 = -\frac{31}{92} \cdot 2 + 81 = -0,337 \times 2 + 81 = -0,674 + 81 = 80,326$$

Observaciones

La media aritmética se emplea en general cuando se desea encontrar un valor que esté en relación con la totalidad de una observación.

Debemos sin embargo tener muy presente que la media aritmética está influenciada por los valores extremos y los cercanos al valor central. A veces esta característica perjudica la solución del problema.

Vamos por ahora sólo a constatar prácticamente esta influencia y dejaremos para más tarde considerar la conveniencia de su elección en comparación a otros valores medios.

X_i	y_i	$X_i\ y_i$
320	3	960
500	16	8000
700	9	6300
2000	2	4000
	30	19260

$$\bar{X} = \frac{19260}{30} = 642$$

X_i	y_i	$X_i\ y_i$
320	1	320
500	16	8000
700	9	6300
2000	4	8000
	30	22620

$$\bar{X} = \frac{22620}{30} = 754$$

Existen otros valores medios que no sufren las influencias de los valores extremos. En cada caso debe decidirse qué valor medio se calculará según convenga o no al problema por resolver la influencia de los valores extremos.

Est. A

6a. Conferencia

MEDIA GEOMETRICA

Concepto: La media geométrica, que simbolizaremos por M_g, de una serie de valores de la variable:

$$X_1, X_2, X_3, \ldots\ldots X_N$$

se define por la relación:

$$M_g = \sqrt[N]{X_1 . X_2 . X_3 . \ldots\ldots X_N} = \sqrt[N]{\prod_{i=1}^{N} X_i} \qquad (1)$$

definición que puede expresarse en forma logarítmica:

$$\log. M_g = \frac{1}{N} \sum_{i=1}^{N} \log X_i \qquad (2)$$

es decir que el logaritmo de la media geométrica de una serie de valores es la media aritmética de sus logaritmos.

En el caso de una distribución de frecuencias, como cada valor de la variable se observa un cierto número de veces y_i, resulta:

$$M_g = \sqrt[\sum y_i]{X_1^{y_1} \; X_2^{y_2} \ldots\ldots X_N^{y_N}} = \sqrt[\sum_{i=1}^{N} y_i]{\prod_{i=1}^{N} X_i^{y_i}} \qquad (3)$$

o en forma logarítmica:

$$\log M_g = \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i \log X_i \qquad (4)$$

Propiedades de la media geométrica

1.- La media geométrica es menor que la media aritmética, aplicadas ambas para los mismos valores de la variable.

Demostraremos que $M_g \leq \bar{X}$

Supongamos que X_N y X_1 son respectivamente los valores mayor y menor de la variable.

Sea a/2 la diferencia entre la media aritmética y la media geométrica de estos valores. Se verificará entonces que:

$$\frac{X_N + X_1}{2} = \sqrt{X_N X_1} + a/2$$

$$X_N + X_1 = 2\sqrt{X_N X_1} + a$$

Despejando el valor de a:

$$a = X_N - 2\sqrt{X_N X_1} + X_1 = \left(\sqrt{X_N} - \sqrt{X_1}\right)^2$$

Por consiguiente si a es igual a un cuadrado resulta positiva y también lo será a/2

Si fuese $X_N = X_1$ entonces resultaría a=0 y por tanto a/2=0
En virtud de estas relaciones, resulta:

$$\frac{X_N + X_1}{2} \geq \sqrt{X_N \; X_1}$$

Si se sustituyen los valores X_N y X_1 por $\frac{X_N + X_1}{2}$, no se afecta el valor de la media aritmética de toda la serie.

Sin embargo, el valor de la media geométrica aumenta porque, como vimos, cuando $X_N \neq X_1$ resulta:

$$\frac{X_N + X_1}{2} > \sqrt{X_N X_1}$$

y se tiene que la contribución de $(\frac{X_N + X_1}{2})^2$ a la media geométrica es mayor que la contribución original de $X_N X_1$

Repitiendo este proceso reiteradamente para el mayor y el menor de los valores restantes se obtiene un valor cada vez mayor para M_g que se aproxima a $\bar{X}$

Cálculo de la media geométrica

En pocos casos puede prescindirse para el cálculo de la media geométrica de los logaritmos, veremos por tanto, esquemáticamente el procedimiento a seguir, utilizando la fórmula (4)

Se tiene el siguiente cuadro:

X_i	y_i	Log X_i	Y_i Log X_i
X_1	y_1	Log X_1	y_1 log X_1
X_2	y_2	Log X_2	y_2 log X_2
.	.	.	.
.	.	.	.
.	.	.	.
X_N	y_N	Log X_N	y_N log X_N
	$\sum_{i=1}^{n} y_i$		$\sum_{i=1}^{n} y_i \log X_i$

$$\text{Log } M_g = \frac{\sum_{i=1}^{n} y_i \log X_i}{\sum_{i=1}^{n} y_i}$$

En caso de tener intervalos de clase, se procede como para el cálculo de la $\bar{X}$, es decir, se sustituyen los valores del intervalo de clase por el valor central.

Ejemplo: Multas impuestas en la Oficina Química

Multas en m$n X_i	Números de multas y_i	Log X_i	y_i log X_i
20	207	1,301.030	269,313.210
50	74	1,698.970	125,723.780
100	24	2,000.000	48,000.000
200	4	2,301.030	9,204.120
	309		452,242.110

$$\text{Log } M_g = \frac{452,242.110}{309} = 1,463.566$$

$$M_g = \text{antilog } 1,463.566 = 29,078$$

Observaciones

Se utiliza la media geométrica especialmente en la estadística económica, aplicándola a los números índices de precios, en los que adquieren la mayor importancia las variaciones relativas.

Un alza de precios representada por el cambio de 50 a 100 es tan importante como un alza de 100 a 200. Pero esta equivalencia no se manifiesta en la media aritmética que concede doble peso al segundo de estos cambios porque opera por diferencias en vez de hacerlo por cocientes.

Un ejemplo frecuentemente citado es el de dos casos de variación de precio: uno que representa un aumento de 10 veces el precio de 100 a 1000 y otro una disminución de una décima, de 100 a 10

la $\bar{X}$ de 1000 y 10 es 505; la M_g es $\sqrt{1000 \times 10} = 100$

Cuando se usa como promedio la M_g, esta nos hace ver que las razones de cambio se compensan mutuamente, resultando completamente incorrecta la $\bar{X}$ como promedio de relaciones entre los cambios de precios.

Existe para la aplicación de la M_g un grave inconveniente y es que no puede utilizarse en el caso que algún valor de la variable sea menor o igual a cero, pues entonces no tiene significación real.

En efecto, si $X_i \leq 0$ (es decir si X_i fuese negativa o nula) resulta:

$$M_g = \sqrt[\sum_{i=1}^{N} y_i]{\prod_{i=1}^{N} X_i^{y_i}} \qquad \sqrt[\sum_{i=1}^{N} y_i]{a}$$

siendo:

$$a \leq c$$

Cuando la variable toma valores según una progresión aritmética y la distribución de frecuencias es aceptablemente simétrica la $\bar{X}$ es preferible a la M_g.

Est. A
7a. Conferencia

MEDIA ARMONICA

Concepto

La media armónica de una serie de valores de la variable, es la recíproca de la media aritmética de los recíprocos de cada una de aquellas.

Simbolizando por M_a a la media armónica, se tendrá para los valores de la tabla siguiente que:

X_i	$1/X_i$
X_1	$1/X_1$
X_2	$1/X_2$
.	.
.	.
.	.
X_N	$1/X_N$

luego:

$$\overline{\frac{X}{X}}_1 = \frac{\sum_{i=1}^{N} \frac{1}{X_i}}{N} \qquad (1)$$

$$M_a = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}$$

En el caso de una distribución de frecuencias resultará:

X_i	$1/X_i$	y_i	$y_i \; 1/X_i$
X_1	$1/X_1$	y_1	$y_1 \; 1/X_1$
X_2	$1/X_2$	y_2	$y_2 \; 1/X_2$
X_3	$1/X_3$	y_3	$y_3 \; 1/X_3$
.	.	.	.
.	.	.	.
.	.	.	.
X_N	$1/X_N$	y_N	$y_N \; 1/X_N$
		$\sum_{i=1}^{N} y_i$	$\sum_{i=1}^{N} y_i \frac{1}{X_i}$

luego

$$\overline{\frac{X}{X}}_1 = \frac{\sum_{i=1}^{N} \frac{y_i}{X_i}}{\sum_{i=1}^{N} y_i} \qquad (2)$$

$$M_a = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i \frac{1}{X_i}}$$

Cálculo de la media armónica

La aplicación de las fórmulas (1) y (2) es teóricamente muy sencilla, sin embargo el operar con las recíprocas de los valores hace el cálculo muy laborioso.

Puede darse una expresión para el cálculo de la M_a en la que no es necesario utilizar las inversas de las frecuencias, pero también resulta una forma compleja.

Si son dos valores:

$$M_a = \frac{\sum_{i=1}^{2} y_i}{\frac{y_1}{X_1} + \frac{y_2}{X_2}} = \frac{X_1 X_2 \sum_{i=1}^{2} y_i}{y_1 X_2 + y_2 X_1} \qquad (3)$$

para tres valores:

$$M_a = \frac{\sum_{i=1}^{3} y_1}{\frac{y_1}{X_1} + \frac{y_2}{X_2} + \frac{y_3}{X_3}} = \frac{X_1 X_2 X_3 \sum_{i=1}^{3} y_i}{y_1 X_2 X_3 + y_2 X_1 X_3 + y_3 X_1 X_2} \qquad (4)$$

Generalizando la fórmula para n valores de la variable:

$$M_a = \frac{X_1 X_2 X_3 \ldots X_N \sum_{i=1}^{N} y_1}{y_1 X_2 X_3 \ldots X_N + y_2 X_1 X_3 \ldots X_N + \ldots + y_N X_1 X_2 \ldots X_{n-1}} \qquad (5)$$

Las fórmulas (3), (4) y (5) se simplifican para el caso de ponderación unitaria:

$$M_a = \frac{2 X_1 X_2}{X_1 + X_2} \qquad (3')$$

$$M_a = \frac{2 X_1 X_2 X_3}{X_2 X_3 + X_1 X_3 + X_1 X_2} \qquad (4')$$

$$M_a = \frac{n X_1 X_2 \ldots X_N}{X_2 X_3 \ldots X_N + X_1 X_3 \ldots X_N + \ldots + X_1 X_2 \ldots X_{N-1}} \qquad (5')$$

En las distribuciones de frecuencias a base de intervalos de clase puede calcularse M_a tomando como valores de X_i, los puntos medios de los intervalos cuyos límites son las inversas de los de cada clase.

Obteniendo la $\overline{X}$ de la distribución así corregida, se llega a la M_a calculando la inversa del valor conseguido.

Este proceso conduce a cálculos complicados.

Ejemplo:

Calcular la M_a en:

longitudes	X_i	y_i	$1/X_i$	$y: 1/X_i$
76-78	77	6	0,01298	0,07788
78-80	79	10	0,01278	0,12780
80-82	81	20	0,01234	0,24680
82-84	83	8	0,01204	0,09632
		44		0,54880

$$M_a = \frac{44}{0,54880} = 80,17$$

Utilizando la fórmula (4)

$$M_a = \frac{44 \times 77 \times 79 \times 81 \times 83}{6 \times 79 \times 81 \times 83 + 10 \times 77 \times 81 \times 83 + 20 \times 77 \times 79 \times 83 + 8 \times 77 \times 79 \times 81}$$

$$M_a = 80,32$$

Propiedad de la M_a

La M_a es menor o igual a la media geométrica es decir que:

$$M_a \leq M_g$$

Sean X_1 y X_N respectivamente el más pequeño y el mayor de los valores de una serie.

Demostraremos antes que:

$$\frac{X_1 + X_N}{2} \geq \sqrt{X_1 X_N}$$

luego:

$$X_1 + X_N \geq 2\sqrt{X_1 X_N}$$

multiplicando ambos miembros por $\sqrt{X_1 X_N}$ se tiene:

$$\sqrt{X_1 X_N}\,(X_1 X_N) \geq 2 X_1 X_N$$

$$\sqrt{X_1 X_N} \geq \frac{2 X_1 X_N}{X_1 + X_N} \qquad (5)$$

Pero observamos que el segundo miembro no es otra cosa que la M_a, en efecto:

$$\frac{2 X_1 X_N}{X_1 + X_N} = \frac{2}{\frac{X_1 + X_N}{X_1 X_N}} = \frac{2}{\frac{1}{X1} + \frac{1}{X_N}} = M_a$$

por tanto podemos por (5) escribir que:

$$M_g \geq M_a$$

Si se sustituyen X_1 y X_N por su media armónica $\frac{2 X_1 X_N}{X_1 + X_N}$ no cambia el valor de M_a para toda la serie; pero M_g disminuye pues vimos que $\sqrt{X_1 X_N} > \frac{2X_1 X_2}{X_1 + X_2}$ cuando $X_1 \neq X_N$ y así la contribución de $\left(\frac{2 X_1 X_N}{X_1 + X_N}\right)^2$ a la M_g sería menor que la contribución $X_1 X_2$

Repitiendo este proceso para el más pequeño y el mayor restantes, se obtiene un valor cada vez más pequeño para M_g que se aproxima a la M_a.

Observaciones

La media armónica es un tipo de promedio de aplicación restringida que se emplea especialmente para evitar errores en la elaboración de ciertas clases de datos.

Deberá usarse cuando se promedian velocidades.

Por ejemplo: Un automóvil ha recorrido 40 Kms. a la velocidad de 80 Kms. por hora y otros 40 Kms. a 120 Kms. por hora. Se desea saber cual fué el promedio de la velocidad.

La $\bar{X}$ de las dos velocidades es de 100 Kms. pero este promedio no es correcto.

El automóvil marchó durante 30' a 80 Kms. por hora

80 Kms. ---------- 60'
40 Kms. ---------- 30'

y marchó 20' a 120 Kms. por hora

120 Kms. ---------- 60'
40 Kms. ---------- 20'

Recorrió por tanto 80 Kms. en 50' y el promedio de la velocidad fué:

50' ---------- 80 Kms.
60' ---------- $\frac{80 \times 60}{50}$ = 96 Kms. por hora

Equivale este resultado a calcular la $\bar{X}$ de 80 y 120 ponderado el primero por 30 y el segundo por 20.

Puede obtenerse directamente este resultado calculando la M_a de las dos velocidades, es decir:

$$M_a = \frac{2 \times 8 \times 120}{80 + 120} = 96 \text{ Kms.}$$

Est.. A
8a.. Conferencia

MODA

Concepto

Hemos dicho que el valor de la variable X para el que se presenta la concentración máxima, en una distribución de frecuencias, puede considerarse un valor representación de toda la distribución.

Este valor se denomina modo (promedio típico, moda o norma) y el intervalo de clase al que corresponde es el intervalo modal.

En la representación gráfica de una distribución de frecuencias, el modo es el valor de la variable X que corresponde a la ordenada máxima.

Simbolizaremos al modo por M_o

El concepto de valor modal es fácilmente comprendido: es la renta más general, la estatura más corriente, en pocas palabras, es el valor más denso de la variable.

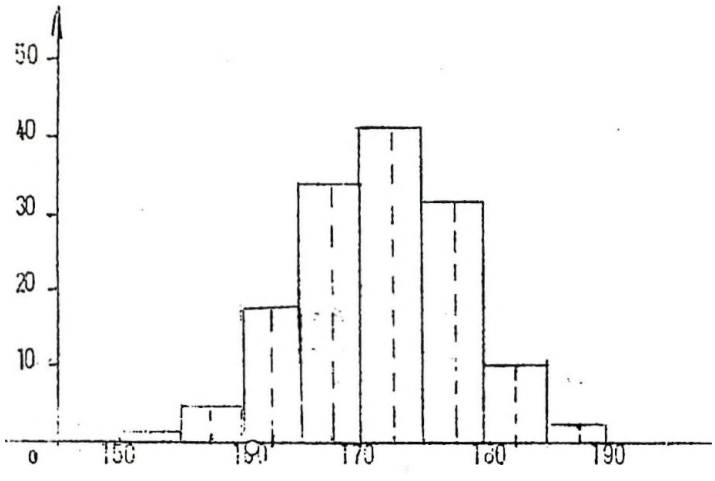
No es fácil, sin embargo, fijar el valor modal que corresponde a una distribución de frecuencias dada.

Existen dos métodos, uno aproximado y otro riguroso para determinar el modo.

Veremos ahora el método aproximado que da resultados prácticos suficientes.

Para explicar el método aproximado que se sigue para la determinación del valor modal en los trabajos estadísticos, vamos a utilizar el siguiente ejemplo:

ESTATURA EN CMS.	x_i	Nº DE ENFERMOS y_i
150-155	152,5	1
155-160	157,5	5
160-165	162,5	16
165-170	167,5	34
170-175	172,5	41
175-180	177,5	32
180-185	182,5	10
185-190	187,5	2



El intervalo de clase cuyos límites son 170-175 contiene la mayor frecuencia, es esta la clase modal y su valor medio que es 172,5 puede aceptarse como valor aproximado del modo.

Si la distribución fuese simétrica, este valor central da, sin error, el valor real del modo, pero en nuestro ejemplo se observa que las frecuencias de las clases inferiores son numéricamente mayores que las frecuencias de las clases superiores a la clase modal.

En la hipótesis de que dentro de la clase modal el modo se aproxima tanto más a uno de sus límites cuanto mayor es la frecuencia de la clase adyacente, tendríamos que el M_o se aproximará más a 170 que a 175.

Conviniendo en que la distancia del M_o a cada extremo del intervalo sea inversamente proporcional a la frecuencia de clase adyacente, puede calcularse su valor así:

Sea X_{l_i} M_o el límite inferior de la clase modal

X_{l_s} M_o el límite superior de la clase modal

y_o la frecuencia de la clase modal

y_1 la frecuencia de la clase que sigue a la clase modal

y_{-1} la frecuencia de la clase que antecede a la clase modal

h amplitud del intervalo de clase

Dividiendo a h en dos partes inversamente proporcionales a y_1 e y_{-1} tendremos medidas en unidades de la distribución, las distancias d_{-1} y d_1 a $X_{l_i}M_o$ y a $X_{l_s}M_o$ respectivamente

del M_o

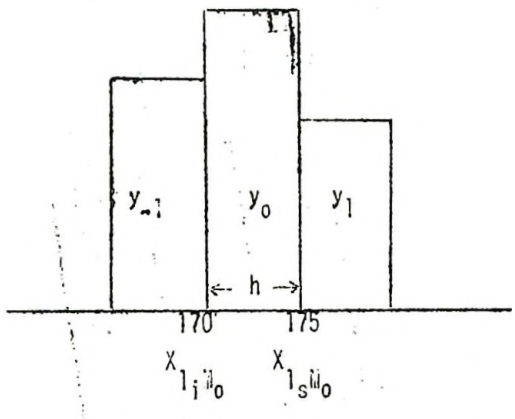
Será :

$$d_{-1} = \frac{h}{\frac{1}{y_1} + \frac{1}{y_{-1}}} \cdot \frac{1}{y_{-1}} = \frac{h \; y_1 \; y_{-1}}{(y_1 + y_{-1}) \; y_{-1}} = \frac{y_1}{y_1 + y_{-1}} \cdot h$$

luego:

$$M_o = X_{1_i M_o} + d_{-1} = X_{1_i M_o} + \frac{y_1}{y_1 + y_{-1}} .h \quad (1)$$

en el caso que se utilice el límite inferior



Si se considerase el límite superior, sería:

$$d_1 = \frac{h}{\frac{1}{y_1} + \frac{1}{y_{-1}}} \cdot \frac{1}{y_1} = \frac{y_{-1}}{y_1 + y_{-1}} .h$$

luego:

$$M_o = X_{1_s M_o} - d_1 = X_{1_s M_o} - \frac{y_{-1}}{y_1 + y_{-1}} h \quad (2)$$

Aplicando la fórmula (1) a los datos de nuestro ejemplo resulta:

$$M_o = 170 + \frac{32}{34 + 32} .5 = 170 + \frac{160}{66} = 170 + 2,42 = 172,42$$

Si se usara la fórmula (2) se tendría:

$$M_o = 175 - \frac{34}{34+32} \cdot 5 = 175 - \frac{170}{66} = 175 - 2{,}58 = 172{,}42$$

Estas fórmulas permiten mediante un sencillo cálculo determinar el valor del modo con un error que no tiene importancia en la mayor parte de los casos.

Si existiese irregularidad en las frecuencias, para conseguir mayor aproximación, pueden considerarse las de dos o más intervalos a cada lado del de mayor frecuencia.

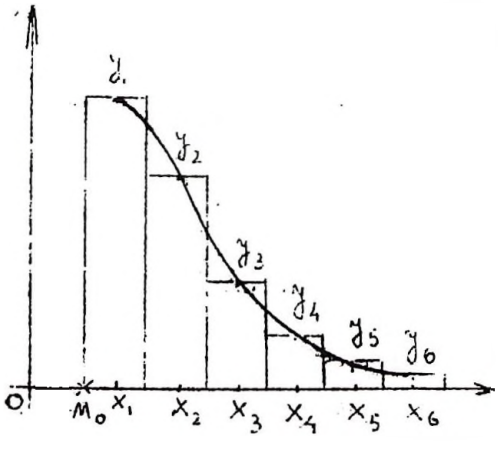
Basta para ello sustituir en (1) y (2)

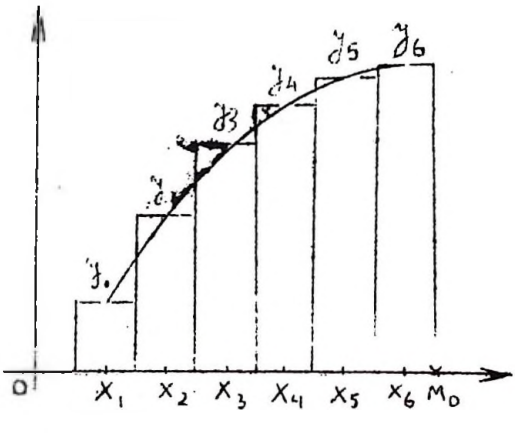
y_1 por $y_1 + y_2 + \cdots\cdots$

y_{-1} por $y_{-1} + y_{-2} + \cdots$

Es preciso, cuando la amplitud de los intervalos próximos no sea la misma, hacer las correcciones necesarias, para que tanto y_1 como y_{-1} se refieran a intervalos de igual amplitud, para lo cual hay necesidad de dividir ambas frecuencias por la amplitud de los intervalos respectivos o bien sumar las de varios intervalos a uno de los lados para que la amplitud del intervalo de la suma sea igual a la del otro lado.

En las distribuciones que dan origen a curvas de concentración la mayor frecuencia se encuentra siempre en uno de los intervalos extremos





En estos casos puede considerarse que la distribución sigue,

dentro del intervalo, la misma forma que en el conjunto total. Por ello, si el intervalo de máxima frecuencia es el primero, se atribuye al modo un valor igual a su límite inferior y si es el último igual a su límite superior.

En algunos casos se complica el problema a causa de la existencia de diversos puntos de concentración, en vez del único punto supuesto en nuestras explicaciones.

Si las distribuciones presentan dos puntos de concentración se las llama bimodales y su representación gráfica da una curva de frecuencias con dos valores máximos; pero si presenta más de dos puntos de concentración se las designa con el nombre genérico de plurimodales.

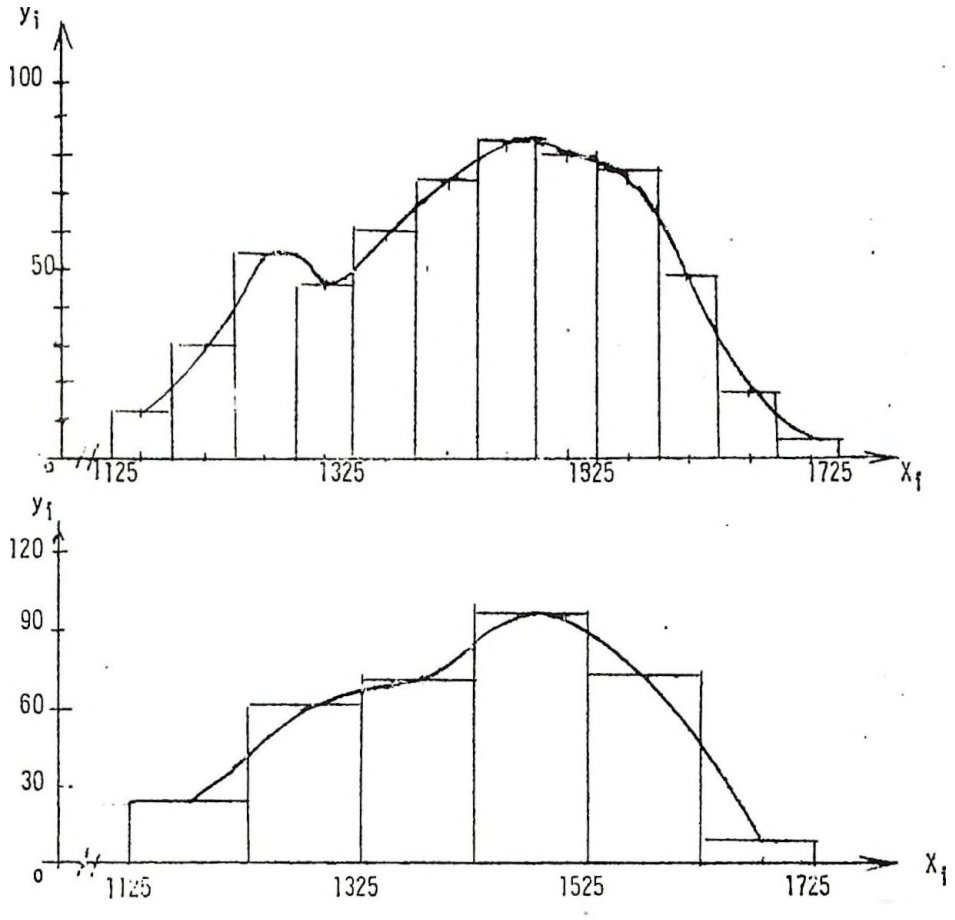
Refiriéndonos a las curvas bimodales la existencia del doble modo puede tener su explicación en la coexistencia de elementos heterogéneos (ejemplo: dos razas distintas en el caso de las poblaciones), bien en la insuficiencia de la experiencia, si se trata de datos homogéneos.

Puede también tener origen en el uso de un intervalo demasiado pequeño con relación al número de casos comprendidos en la clasificación.

En estos casos suele convertirse la distribución en monomodular (unimodal) variando los límites de los intervalos de clase, ampliándolos.

Ejemplo:

ESTATURAS DE NIÑAS Xi	NUMERO DE NIÑAS Yi	NUEVO REAGRUPAMIENTO DE ESTATURAS
1.125 - 1.175	15	51 ÷ 2 = 25,5
1.175 - 1.225	36	
1.225 - 1.275	63	118 ÷ 2 = 59
1.275 - 1.325	55	
1.325 - 1.375	69	144 ÷ 2 = 72
1.375 - 1.425	75	
1.425 - 1.475	91	181 ÷ 2 = 90,5
1.475 - 1.525	90	
1.525 - 1.575	85	137 ÷ 2 = 68,5
1.575 - 1.625	52	
1.625 - 1.675	10	11 ÷ 2 = 5,5
1.675 - 1.725	1	



Si a pesar de los cambios realizados en la amplitud de los intervalos subsistiese la distribución bimodal, será debido a la falta de homogeneidad de los datos, esto es, a que casualmente se hayan combinado dos distribuciones diferentes, casos que ocurren, con frecuencia, en las observaciones de carácter biométrico.

La existencia de dos especies de animales donde solo se sospechaba la existencia de una, ha quedado revelada, a veces, por este procedimiento.

El significado completo de una distribución de frecuencias desaparecerá si los datos no son homogéneos, hecho que es cierto, lo mismo en el campo de la estadística económica, como en cualquier otro.

Propiedad del modo

1.- En una distribución simétrica, el modo y la media aritmética coinciden.

Considerando una distribución en que las frecuencias disminuyen de la máxima con la misma intensidad en los dos sentidos, es decir, una distribución simétrica, al ser:

$$y_1 = y_N \qquad X_r - X_1 = X_N - X_r$$

$$y_2 = y_{N-1} \qquad X_r - X_2 = X_{N-1} - X_r$$

$$y_3 = y_{N-2} \qquad X_r - X_3 = X_{N-2} - X_r$$

$$\cdots\cdots\cdots \qquad \cdots\cdots\cdots\cdots\cdots$$

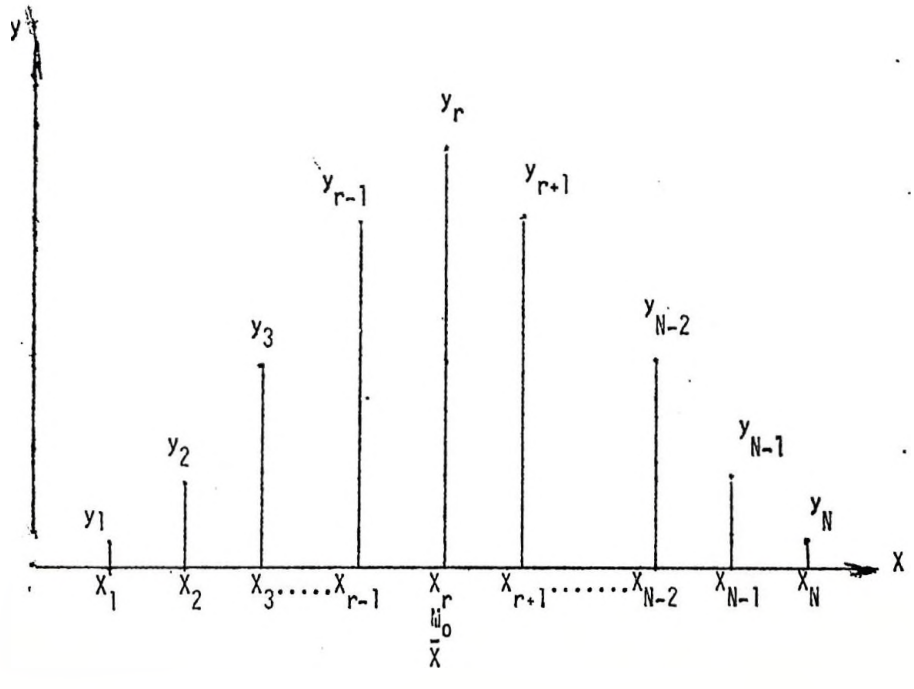
$$y_{r-1} = y_{r+1} \qquad X_r - X_{r-1} = X_{r+1} - X_r$$

resulta, multiplicando ordenadamente los primeros y segundos miembros de las igualdades anteriores entre sí y sumando:

$$\sum_{i=1}^{r-1} (X_r - X_i) y_i = \sum_{i=r+1}^{N} (X_i - X_r) y_i$$

Como esta fórmula expresa la primera propiedad de la media aritmética, por la cual la suma de los desvíos negativos es igual a la suma de los desvíos positivos, en valor absoluto, resulta:

$$X_r = M_o = \bar{X}$$



Obs nes

Las fluctuaciones externas al intervalo modal no pesan sobre valor como sucede con la media aritmética, lo que lo hace muy ú

til en los casos en que la forma de la distribución sea muy irregular.

Hay algunas distribuciones que presentan varios modos, en este caso deben calcularse todos e interpretarse como diferentes valores normales de las mismas.

Un examen minucioso de los diferentes modos y de los datos de la distribución puede ser el origen del conocimiento de circunstancias de gran importancia en el análisis del fenómeno que se estudia.

A veces se presentan varios modos porque el número de observaciones revificadas es escaso, no siendo por ello representativa la distribución.

Otras veces lo que sucede es que no existe homogeneidad en las observaciones, habiéndose mezclado las de varios fenómenos distintos.

Esto suele ocurrir en la clasificación por algún caracter físico de los habitantes de paises de población heterogénea desde el punto de vista racial, en la que el tipo predominante de cada grupo origina un modo.

En la clasificación de los fallecidos por edades se da tambien este fenómeno por existir, en la población, varios grupos determinados por la edad, afectados por un tipo distinto de mortalidad y como los comprendidos en algunas edades participan simultáneamente de dos tipos, se superponen las distintas distribuciones, dando como resultado una curva ondulada que presenta varios máximos.

En todo caso la existencia de varios modos, indica la necesidad de realizar un estudio por separado de los grupos que los originan.

Est. A
9a. Conferencia

MEDIANA

Concepto

Ordenados los valores de la variable por su magnitud en forma creciente o decreciente, se designa con el nombre de mediana al valor central de la variable, o sea el valor respecto del cual los mayores y los menores que él se presentan con igual frecuencia.

Por ejemplo, consideremos los valores de las rentas de siete personas:

$x_1 = 550$
$x_2 = 915$
$x_3 = 1.000$
$x_4 = 1.[illegible]$
$x_5 = 14[illegible]$
$x_6 = 17[illegible]$
$x_7 = 1890$

El valor de $ 1.320 es la mediana, pues quedan tres valores de la variable a cada lado.

El problema es algo diferente cuando se trata de un número par de datos.

Por ejemplo: promedio de salario por hora obrero trabajadas en el 1er. trimestre de 1949 en fábricas de la Provincia de Buenos Aires:

Industria	Promedio Salario
Aceites comestibles	[illegible],80
Hilados y tejidos	[illegible],00
Tejidos y artículos de punto	[illegible],76
Tejidos de seda	[illegible],92
Papel	3,38
Fibras textiles	[illegible],75
Vidrios	[illegible],14
Astilleros	[illegible],07

Ordenando los promedios de menor a mayor tenemos:

2,50-2,75-2,78-2,80-2,90-2,92-3,00-3,07-3,14-3,38-3,50

De acuerdo a la definición cualquier valor que deje cuatro promedios a cada lado será válido, por tanto, todo valor comprendido entre 2,90 y 2,92, que son los que ocupan la posición central, servirá.

En estas condiciones en que la mediana está realmente indeterminada, puede admitirse convencionalmente cualquier valor comprendido entre los expresados valores límites centrales.

Por lo general, a fin de restar arbitrariedad a la elección se promedian los dos valores centrales.

En el caso que estamos considerando, tendríamos:

$$\frac{2,90 + 2,92}{2} = 2,91$$

El valor obtenido para la mediana no coincide con el salario de ninguna de las industrias enumeradas, lo que ocurre siempre que se tenga un número par de valores de la variable.

Si las variables que se estudian son discontinuas o discretas, se obtendrá, en caso de darsenos un número par de valores, un valor irreal para la mediana, inconsistente con la naturaleza de los datos.

Por ejemplo: Si el número de hijos que tienen 6 familias fueran:

1, 3, 4, 5, 7, 9

la mediana sería 4,5 valor sin sentido.

Como puede observarse, la mediana depende de la posición de los términos en la ordenación y no de su amplitud.

En el caso que estamos considerando, es decir, en aquel en que los valores de la variable sólo se constatan una vez o tienen todos iguales frecuencias, el valor de la mediana, es decir del orden del término que debe tenerse en cuenta es:

$$M_a = \frac{N+1}{2}$$

Mediana de las distribuciones de frecuencias

En una distribución de frecuencias la mediana es aquel valor de la variable cuya ordenada divide a la población o universo consi

dorado en dos partes iguales...

Para determinar su valor se procede por etapas, pero conviene recordar que a los fines del cálculo suponemos que la distribución de las frecuencias es uniforme dentro de cada intervalo de clase.

Las operaciones sucesivas son:

1°) Se divide por dos el total de los datos, lo que nos da el número de estos que ha de quedar a cada lado del valor de la variable que se trata de determinar:

$$\frac{\sum_{i=1}^{N} y_i}{2}$$

2°) Sumamos, partiendo del extremo inferior de la escala, las frecuencias de cada grupo sucesivo hasta alcanzar el límite del grupo que contiene el valor de la mediana buscada.

3°) Determinamos el número de datos de este intervalo que se deben sumar al total de frecuencias obtenidas, para hallar un número igual a:

$$\frac{\sum_{i=1}^{N} y_i}{2}$$

4°) Dividimos este número adicional por el número total de datos que existen en el grupo que comprende a la mediana lo que nos dará la fracción del grupo dentro del cual deben hallarse.

5°) Multiplicamos el valor del intervalo de clase por la fracción antes hallada.

6°) Sumamos el resultado determinado en la 5a. etapa al límite inferior del intervalo que contiene a la mediana y obtenemos así su valor.

El proceso puede invertirse comenzando por el extremo superior de la escala y contando hacia abajo.

En este caso la operación final consistirá en restar del límite superior del intervalo de clase que contiene a la mediana el valor hallado en la operación 5a.

Este procedimiento es válido para N par o impar.

En este último caso será $\frac{N}{2}$ fraccionario, pero no por ello se invalida el mecanismo de cálculo.

Cálculo de la Mediana

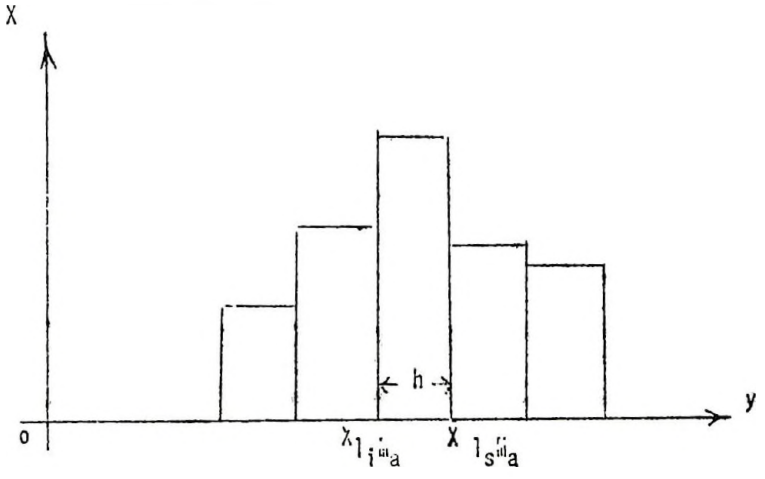
En el caso de una distribución de frecuencias simbolizaremos por:

M_a = mediana

$X_{l_i M_a}$ = límite inferior del intervalo de clase en que se halla la mediana.

$X_{l_s M_a}$ = límite superior del intervalo de clase en que se halla la mediana.

h = amplitud del intervalo de clase



Resulta de acuerdo a las etapas de cálculo descriptas que:

$$M_a = X_{l_i M_a} + \frac{\frac{1}{2}\sum_{i=1}^{N} y_i - \sum_{i=1}^{X_{l_i M_a}} y_i}{\sum_{i=X_{l_i M_a}}^{X_{l_s M_a}} y_i} \cdot h$$

si se trabaja con el límite inferior $X_{l_i M_a}$

Para el caso de considerar el límite superior del intervalo sera:

$$M_a = X_{l_s M_a} - \frac{\frac{1}{2}\sum_{i=1}^{N} y_i - \sum_{i=N}^{X_{l_s M_a}} y_i}{\sum_{i}^{X_{l_s M_a}} y_i}$$

Ejemplo:

ESTATURA EN CMS. X_i	Nº DE ENFERMOS y_i	VALORES ACUMULADOS DE y_i	VALORES ACUMULADOS DE y_i COMENZANDO DESDE EL FINAL
150-155	1	1	143
155-160	5	6	142
160-165	18	24	137
165-170	34	58	119
170-175	41	99	85
175-180	32	131	44
180-185	10	141	12
185-190	2	143	2
	143		

Método utilizando el límite inferior del intervalo de clase en que se encuentra el valor M_a

$$1°)\ \frac{1}{2}\sum_{i=1}^{N} y_i = \frac{1}{2} \times 143 = 71{,}50$$

$$2°)\ M_a = 170 + \frac{71{,}50 - 58}{41} \times 5 = 171{,}65$$

En el caso de emplear el límite superior del intervalor anterior:

$$M_a = 175 - \frac{71{,}50 - 44}{41} \times 5 = 171{,}65$$

Propiedad de la Mediana

En una distribución simétrica, la mediana coincide con la media aritmética y con el modo.

En una distribución simétrica, si llamamos y_r a la frecuencia máxima, resulta:

$$y_1 = y_N$$
$$y_2 = y_{N-1}$$
$$\cdots\cdots\cdots$$
$$y_{r-1} = y_{r+1}$$

sumando ordenadamente, miembro a miembro y relacionando estos resultados con el de agregar a cada uno el valor y_r se tiene:

$$\sum_{i=1}^{r-1} y_i = \sum_{i=r+1}^{N} y_i < \frac{\sum_{i=1}^{N} y_i}{2} < \sum_{i=1}^{r} y_i = \sum_{i=r}^{N} y_i$$

El valor M_a ha de ser uno del grupo de frecuencia y_r, el central de ellos, si todos son distintos, o uno cualquiera de ellos si todos son iguales.

Hemos demostrado ya que en este caso ese valor es el modo y que coincide con la media aritmética, podemos entonces escribir:

$$M_a = M_o = \bar{X}$$

Observaciones

La gran flexibilidad de que goza este valor de permite adaptarse a todos los tipos de distribuciones de frecuencias, como medida de su tendencia central, incluso en las distribuciones muy irregulares.

No influyen sobre la mediana las variaciones externas, dando por ello, buen resultado su aplicación en casos en que en los valores extremos se presentan intensidades excepcionales que sobre la media aritmética pesarían demasiado.

La mediana puede localizarse aún cuando los datos sean incompletos o cuando se trate de una distribución cuyos intervalos de clase extremos sean abiertos.

Se recuerda que en este caso las medias aritmética, geométrica y armónica no pueden calcularse.

Desde el punto de vista algebraico, en lo que respecta a la operatoria, es más simple la media aritmética que la mediana.

Estas limitaciones reducen, en cierto modo, las aplicaciones de la mediana a los trabajos en que no sean necesarias consideraciones teóricas.

El cálculo de la mediana de una distribución de frecuencias con intervalos de clase desiguales, no difieren del que hemos explicado.

Cálculo del Modo

Ya vimos una fórmula aproximada para determinar el modo en una distribución de frecuencias.

Puede también determinarse en forma aproximada, el valor del modo, utilizando las relaciones que existen entre los valores de la media, la mediana y el modo.

Sabemos que si la distribución es simétrica coinciden los tres valores; pero cuando tal simetría falta, aparecen alterados los puntos correspondientes a estos promedios.

Si el grado de asimetría es moderado, estos tres valores mantienen una relación constante que se expresa diciendo que la mediana está situada entre la media y el modo, a un tercio de la distancia que a estos dos separa, a contar desde la media.

Esta relación desaparece cuando la asimetría es muy acentuada.

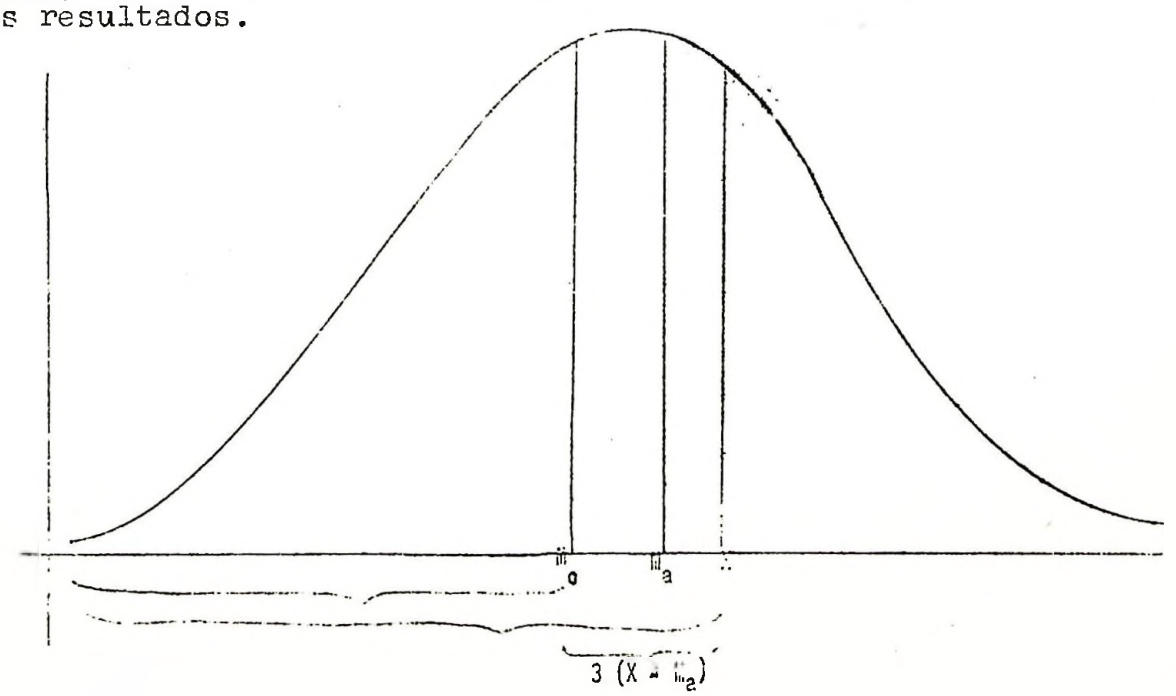
Por lo tanto, cuando dispongamos de dos cualesquiera de estos tres valores en una distribución cuya asimetría sea moderada, podremos hallar el tercero.

Este método se aplica sólo para determinar el modo (para calcular los otros dos valores $\bar{X}$ y M_a deben emplearse las respectivas fórmulas exactas) y sólo cuando no sean aplicables otros procedimientos más convenientes.

La relación expuesta puede expresarse por medio de la siguiente fórmula:

$$M_o = \bar{X} - 3\,(\bar{X} - M_a)$$

Debe observarse que estos valores del modo no ofrecen garantía alguna de exactitud, pues todos los métodos expuestos para la determinación del modo son simplemente métodos aproximados, lo cual nodebe olvidarse al interpretarse y utilizar los resultados.



Est. - A

10a. Conferencia

Cuartiles, deciles y percentiles

Generalidades

A veces es necesario localizar en la escala de la variable de una distribución de frecuencias, valores que dividan en otra forma el número total de frecuencias.

Semejante a la Mediana que divide en dos grupos iguales a la pobalción total o universo son los cuartiles, deciles y percentiles.

Los cuartiles, como su nombre lo indica, son los valores de la variable que dividen al total de frecuencias en cuatro grupos iguales, los deciles los dividen en diez grupos iguales y los percentiles en cien partes iguales.

Es evidente que el primer cuartil corresponde al valor de la variable antes del cual queda una cuarta parte de las frecuencias y despues del cual quedan las tres cuartas partes de ellas.

El segundo cuartil y la mediana tienen valores idénticos.

El tercer decil corresponde al valor antes del cual hay tres décimas del número de frecuencias.

El noveno percentil corresponde al valor antes del cual hay nueve céntimas del número de frecuencias.

Es facil constatar que el quinto decil y el percentil número cincuenta coinciden con la mediana.

Para determinar cualquiera de los valores citados se empezará a contar siempre desde el extremo inferior de la escala.

Los cuartiles, deciles y percentiles no son promedios en sentido nato, pues solo participan de las características de los promedios, en cuanto a que son valores comprendidos dentro de los límites extremos de oscilación de la variable.

Notación y cálculo

En el caso de una serie de valores de la variable con ponderación unitaria, resulta que el primer cuartil es el valor de la variable:

$$X_{\frac{N+1}{4}}$$

el segundo:

$$X_{\frac{2(N+1)}{4}}$$ que es precisamente el de la mediana

y el tercero

$$X_{\frac{3(N+1)}{4}}$$

El primer decilo es el valor $X_{\frac{(N+1)}{10}}$ y el último, ya que son nueve es: $X_{\frac{9(N+1)}{10}}$.

El tercer percentil es el valor $X_{\frac{3(N+1)}{100}}$ y el último $X_{\frac{99(N+1)}{100}}$.

Para las distribuciones de frecuencias el procedimiento de localización es el mismo que el de la mediana y se utiliza la misma fórmula construída en base a los datos de intervalos, en la que se sustituye $\frac{N+1}{2}$ por los subíndices antes indicados.

La fórmula para el primer cuartil es por consiguiente:

$$Q_1 \quad X_{l_i Q_1} + \frac{\frac{1}{4}\sum_{i=1}^{N} f_i - \sum_{i=1}^{X_{Q_1}} f_i}{f_{X_{Q_1}}} \cdot h$$

la del segundo cuartil:

$$Q_2 = X_{l_i Q_2} + \frac{\frac{2}{4}\sum_{i=1}^{N} f_i - \sum_{i=1}^{X_{l_i Q_2}} f_i}{f_{X_{Q_2}}} \cdot h$$

y la del tercero:

$$Q_3 \quad X_{l_i Q_3} + \frac{\frac{3}{4}\sum_{i=1}^{N} f_i - \sum_{i=1}^{X_{l_i Q_3}} f_i}{f_{X_{Q_3}}} \cdot h$$

y la del percentil de orden r es:

$$P_r = X_{l_i P_r} + \frac{\frac{r}{100}\sum_{i=1}^{N} f_i - \sum_{i=1}^{X_{l_i P_r}} f_i}{f_{X_{P_r}}} h$$

Observaciones

El empleo de estos promedios es muy limitado, sin embargo, como representa a daros equidistantes desde el punto de vista de su posición, la mayor o menor desigualdad de las diferencias entre cada dos cuartiles, deciles o percentiles adyacentes, refleja las características de la distribución en cuanto a su regularidad o simetría se refiere, permitiendo conocer los lugares de mayor irregularidad de las mismas.

Determinación gráfica de la mediana, cuartiles, deciles y percentiles.

Para completar el estudio de la curva de frecuencias y de la curva acumulativa de frecuencias, vamos a exponer brevemente el método de determinación gráfica de las medidas de posición antes citadas.

Los valores de la mediana, cuartiles, deciles y percentiles pueden obtenerse gráficamente mediante la curva de frecuencias acumuladas.

Representada gráficamente la curva de frecuencias acumuladas, se puede localizar sobre el eje de ordenadas, el punto que diste del origen $\frac{N}{2}$, se traza por este punto una paralela al eje de abscisas hasta encontrar a la curva, la abscisa del punto de intersección será el valor de la mediana.

Para determinar los cuartiles el procedimiento es igual pero se toman sobre el eje de ordenadas los valores de $\frac{N}{4}$ $\frac{2N}{4}$ y $\frac{3N}{4}$ respectivamente para el primer, segundo y tercer cuartil.

Para los deciles se emplearán los valores $\frac{N}{10}$; $\frac{2N}{10}$; $\frac{3N}{10}$; $\frac{4N}{10}$; ..... $\frac{9N}{10}$ y para los percentiles $\frac{N}{100}$; $\frac{2N}{100}$ ... $\frac{99N}{100}$.

Ejemplo:

Calcular, con los siguientes datos, el 1er. y 3er. cuartil, la mediana, el noveno decil y el percentil número 30.

Determinar sus valores gráficamente:

X_i	y_i	f_i
160-164	5	5
164-168	25	30
168-192	30	60
192-196	36	96
196-200	48	144
200-204	64	208
204-208	100	308
208-212	80	388
212-216	75	463
216-220	70	533
220-224	67	600
	600	

1er. CUARTIL: $\frac{N}{4} = \frac{600}{4} = 150$

$$Q_1 = 200 + \frac{150 - 144}{64} \times 4 = 200 + \frac{6 \times 4}{64} = 200 + 0,375 = 200,375$$

3er. CUARTIL: $\frac{3\,N}{4} = \frac{3 \times 600}{4} = 450$

$$Q_3 = 212 + \frac{450 - 388}{75} \times 4 = 212 + \frac{62 \times 4}{72} = 212 + 3,306 = 215,306$$

MEDIANA: $\frac{N}{2} = \frac{600}{2} = 300$

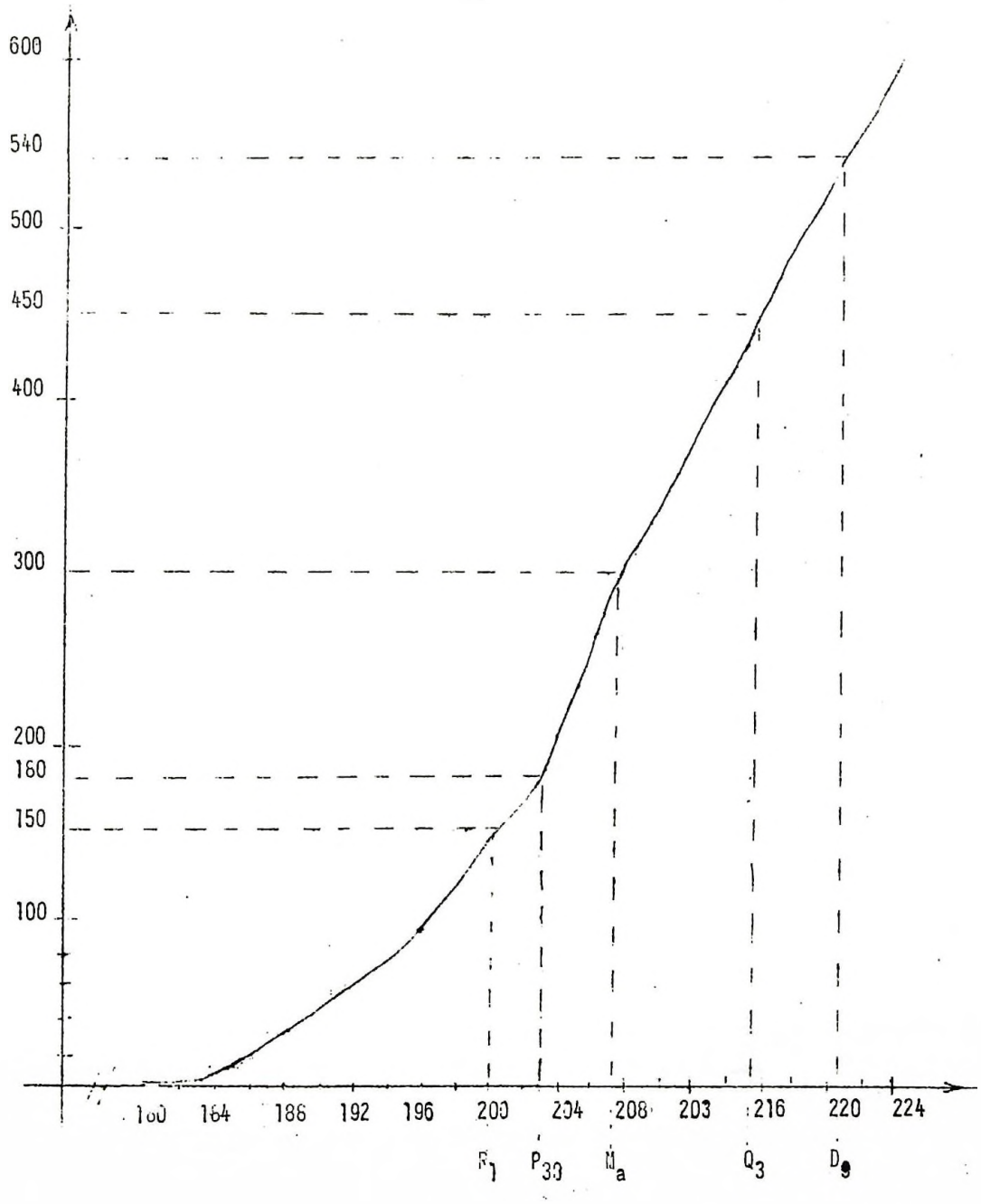
$$M_a = 204 + \frac{300 - 208}{100} \times 4 = 204 + \frac{92 \times 4}{100} = 204 + 3,68 = 207,68$$

9º DECIL: $\frac{9\,N}{10} = \frac{9 \times 600}{10} = 540$

$$D_9 = 220 + \frac{5 + 0-533}{67} \times 4 = 220 + \frac{7 \times 4}{67} = 220 + 0,417 = 220,417$$

30º PERCENTIL: $\frac{30\,N}{100} = \frac{30 \times 600}{100} = 180$

$$P_{30} = 200 + \frac{180 - 144}{64} \times 4 = 200 + \frac{56 \times 4}{64} = 200 + 3,5 = 203,5$$


600
540
500
450
400
300
200
160
150
100
100
164
186
192
196
200
204
208
208
216
220
224
P_{30}
M_a
Q_3
D_9

Est. "A"
11a.Conferencia
24 de marzo de 1955

Medidas de variabilidad y de dispersión

Generalidades

Hasta ahora nos hemos ocupado de los métodos que se emplean para dar a un conjunto de datos cuantitativos, obtenidos de una observación, la forma de una distribución de frecuencia y luego vimos entre los procedimientos que se siguen para su estudio aquellos que cumplen una primer etapa y consisten en obtener un valor en forma de promedio que represente la tendencia central de la distribución.

Pero ya sabemos que un promedio, cualquiera que el sea no representa, por sí mismo la distribución de frecuencias.

Ya dijimos que para conocer la medida de las características principales de una distribución de frecuencias dada, así como para realizar su comparación con otras distribuciones, es necesario disponer de otros tres valores

El primero de ellos es la medida del grado en que los datos incluidos en la distribución se separan o varian del valor central o sea el grado de variabilidad y dispersión.

El segundo es la medida del grado de asimetría que la distribución presenta, del equilibrio o falta de equilibrio entre las dos regiones en que resulta dividida por el valor central.

El tercero es la medida de la kurtosis o grado de concentración de los valores en la clase modal.

Veremos ahora las medidas de variabilidad y dispersión.

Si se apunta exactamente con una pieza de artilleria sobre un blanco y se hacen con ella 100 disparos, se hallará que los impactos que corresponden a estos disparos se dispersan alrededor del verdadero blanco.

Por muy exacta que sea la puntería y por muy ajustada que se halle la pieza con que se dispara, solo, un pequeño tanto por ciento de los disparos, dará exactamente en el blanco

Los impactos se dispersarán sin embargo, en torno del blanco en forma completamente regular .

Si se construye un rectángulo que contenga todos los impactos y se divide dicho rectángulo (o zona de dispersión) en ocho partes iguales, la distribución será la que aparece en el siguiente diágrama:

2	7	16	25	25	16	7	2

En algún caso particular existirán quizá algunas pequeñas divergencias del orden indicado pero al fin prevalecerá la distribución.

La regla indicada puede aplicarse a cualquier clase de cañón.

Cuanto mejor se halle construido el cañón más pequeña será la zona de dispersión, pero la distribución dentro de esa zona es teóricamente la misma en todos los casos.

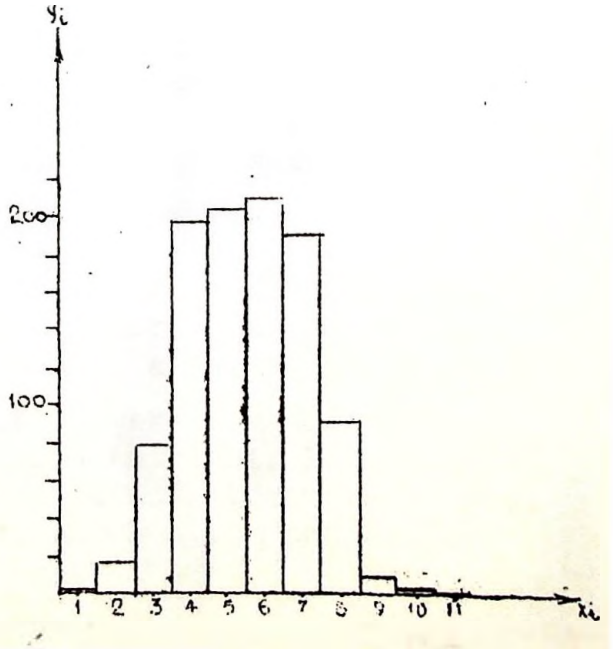
Las reglas de tiro que se usan en balística están basadas precisamente en el hecho que acabamos de señalar.

En el caso que ahora estudiaremos, los resultados se hallan de acuerdo con la distribución teórica.

Ejemplo: distribución de mil disparos hechos con un cañón sobre el centro de un blanco fijo colocado a 100 metros.

El blanco fué dividido horizontalmente en once partes iguales..

División Xi	Impactos registrados Yi.
1	2
2	16
3	79
4	193
5	204
6	212
7	190
8	89
9	10
10	4
11	1
	1000



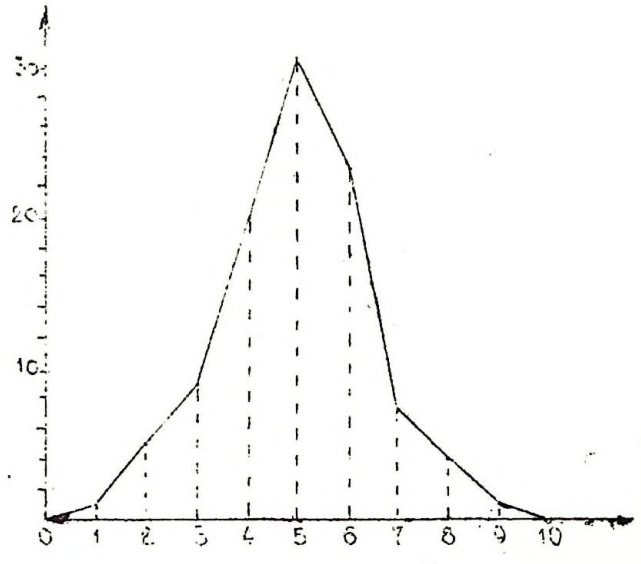
La débil tendencia de concentración que se observa hacia la mitad inferior del blanco es debida, indudablemente, a la imposibilidad de corregir por completo, el efecto de la gravedad al apuntar con el cañón.

Cuando arrojamos monedas al aire se supone que la distribución de cara o cruz está determinada por el azar.

Tenemos la siguiente distribución correspondiente a la frecuencia con que aparece cara en 100 experiencias, consistente, cada una, en arrojar 10 monedas a cara o cruz.

El número de veces que puede aparecer cara en cada experincia varía desde 10 como máximo hasta cero vez

nº de caras	frecuencias
0	0
1	1
2	5
3	9
4	20
5	30
6	23
7	7
8	4
9	1
10	0
	100



Los ejemplos que hmos dado no se relaciona con los hechos económicos pero se consideba que también ellos conducen a distribuciones de frecuencias aunque estas no presentan la simetría y regularidad que caracterizan a las distribuciones correspondientes a datos físicos.

Pero a pesar de tales diferencias, existe evidente semejanza entre las medidas que proceden del campo económico con aquellas observadas en Astronomía, Antropometría, Balística, etc., cuyas principales características vamos a estudiar.

Variación

Aparece en primer lugar, entre los caracteres generales la variación que presentan los valores de las medidas observadas. Algunos autores llaman rango a la variación.

Varía la estatura humana; las diversas medidas astronómicas de una misma cantidad difieren entre si; los proyectiles disparados en condiciones todo lo idénticas que permite la humana naturaleza no dan siempre en el mismo punto; las

rentas varían de un individuo al otro y las cotizaciones cambian de semana en semana, de mes en mes.

Las diferentes observaciones o valores obtenidos en un caso determinado resultan distribuidas a lo largo de una escala entre dos valores extremos, llamado intervalo de variación de la observación

La distribución de estos valores a lo largo de una escala o eje de abscisas x, es tal que yendo de uno de los valores extremos hacia el otro, los casos que ocupan los puntos sucesivos de la escala, o sea los sucesivos grupos de frecuencias, aumentan más o menos regularmente hasta alcanzar un máximo y decrecen después en forma análoga.

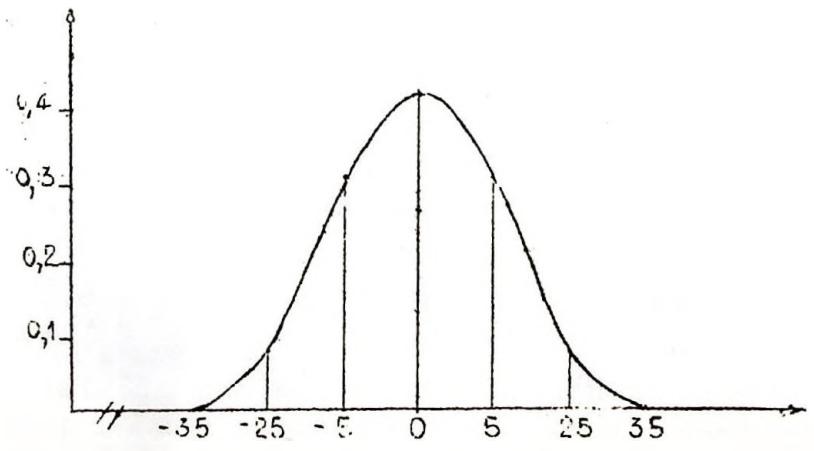
Pero a pesar de esta variación, encontramos una tendencia central, una acumulación de casos en ciertas regiones de la abscisa.

Esta es la segunda característica notable que aparece como un carácter común a todas las distribuciones de frecuencias.

Si medimos la desviación que presenta cada una de las frecuencias de clase a partir del punto de máxima concentración, se nota que son raros los desvíos extremos y que los observados a uno y otro lado de dicho punto presentan perfecta o casi perfecta igualdad en los ejemplos que nos ofrecen las ciencias físicas y los que proceden de fenómenos en que interviene solamente el azar, mientras que dichos desvíos son tan solo aproximadamente iguales en las distribuciones económicas.

Frecuentemente se encuentran excepciones a esta regla, como la que se observa en el ejemplo de la distribución de rentas.

El gráfico siguiente:



representa una curva que se denomina curva de probabilidad o curva normal de los errores.

Es esta curva un tipo fundamental al que se aproximan estrechamente algunos de los ejemplos anteriores y del cual son desviaciones más o menos acentuados otros ejemplos conocidos.

Hay que reconocer que existen numerosos e importantes casos que se separan de este tipo que, sin embargo adquiere gran importancia en los trabajos estadísticos como forma - fundamental, pues hasta sus variedades más importantes se le asemejan de tal modo que basta esta semejanza para justificar el empleo de un método general en el estudio de la distribución de frecuencias.

Las distribuciones de datos cuantitativos varían y las diferencias que existen entre ellas y ciertos tipos standard son de importancia máxima, pues a pesar de esas diferencias,hay semejanzas.

Prácticamente, toda colección de datos cuantitativos que provengan de mediciones, realizadas en el campo social,en el biológico o en el económico, viene caracterizada por su variabilidad, por las diferencias cuantitativas que se observan entre los diferentes valores que forman la colección.

La variabilidad biológica constituye uno de los factores fundamentales en el proceso de la evolución.

Ninguna medida relativa a las propiedades físicas de un grupo racial, como por ejemplo, la estatura, es completa si no la acompaña otra medida de variabilidad media de la propiedad a que se refiera.

El promedio de las rentas de un país determinado tiene quizás menor importancia que la variabilidad de esas rentas, esto es, que las diferencias, entre las rentas percibidas por personas pertenecientes a las diversas clases de la sociedad.

Un promedio tiene por sí mismo escaso valor significativo, a menos que se conozca el grado de variabilidad de la distribución de frecuencias que aquel representa.

Si la variabilidad es tan grande que no aparece acusada en ninguna tendencia central, el promedio carece de significado.

Al descender el grado de variabilidad aumenta dicho significado.

Se estudie una sola distribución de frecuencias, o bien se la compare con otras distribuciones, deberá añadirse siempre a la medida de la tendencia central la medida de la variabilidad.

Medidas de variación absoluta

La variación o dispersión puede expresarse en función de las unidades de medida en que están expresados los datos o en números abstractos en forma de porcentajes independientemente de dichas unidades.

En el primer caso se habrá medido la variabilidad absoluta; en el segundo la variabilidad relativa que es más conveniente que la absoluta para establecer comparaciones.

Estudiaremos en primer término las medidas de variabilidad absoluta.

Oscilación:

Una medida grosera de la variación nos la proporciona la oscilación que es la diferencia absoluta que existe entre el dato de menor valor y de el valor más elevado entre las que forman la distribución.

La tabla siguiente contiene la distribución de tipos de cambio mensuales Londres-New York durante el período 1882-1913

X_i	Y_i
4.8275-4.8224	18
4.8425-4.8574	68
4.8575-4.8724	110
4.8725-4.8874	129
4.8875-4.9024	57
4.9025-4.9174	2
	384

El dato menor entre los incluídos en dicha tabla es 4.8275 dollares; el valor más elevado es 4.9174 y la oscilación será , por tanto:

$$4.9174 - 4.8275 = 0.0899 \text{ dollares}$$

Una distancia que representa en la escala el valor 0.0899 comprenderá, por tanto, a todos los datos.

Es evidente que el valor de la oscilación depende tan sólo de los valores que tengan los datos extremos y el cambio de uno sólo de estos datos bastaría, por lo tanto, para hacer variar su valor.

Como la oscilación presenta un valor poco estable y nada representativo de la verdadera distribución de los datos, es poco usada en los trabajos estadísticos.

En efecto, podemos obtener un mismo valor para el campo de variación de una curva de frecuencias simétrica que en una curva en forma de "J", y es evidente que no se puede deducir de ello que estas dos distribuciones tengan una misma dispersión.

La oscilación se usa, en cambio, a menudo como una medida indicadora de las fluctuaciones que presenta el mercado de valores aunque sea discutible su conveniencia para dicho fin.

Desviación media

Encontramos una medida más aproximada de la dispersión de los datos en torno a un valor central midiendo la desviación que presenta cada uno de los datos respecto de este valor y promediando estas desviaciones.

El método de cálculo de la desviación media puede verse en el siguiente ejemplo:

	X_i	Y_i	$x_i = \lvert \bar{X} - X_i \rvert$	$X_i\ Y_i$	$x_i\ y_i$
2 - 4	3	1	6	3	6
5 - 7	6	6	3	36	18
8 - 10	9	7	0	63	0
11 - 13	12	4	3	48	12
14 - 16	15	2	6	30	12
		20		180	48

$$\bar{X} = \frac{180}{20} = 9 \qquad D.M. = \frac{\sum_{i=1}^{N} |X_i| Y_i}{\sum_{i=1}^{N} Y_i} = \frac{48}{20} = 2{,}4$$

La media aritmética es 9.-

Se sumarán los valores absolutos del producto de los desvíos respecto de la media aritmética, multiplicados por las respectivas frecuencias de clase y se divide esta suma por el número total de observaciones.

El procedimiento queda expresado en la fórmula:

$$D.M. = \frac{\sum_{i=1}^{N} /x_i/\ y_i}{\sum_{i=1}^{N} y_i} \qquad (1)$$

En general, la desviación media en una distribución de frecuencias es la media aritmética de sus desvíos, ponderados por las frecuencias, abstracción hecha de los signos respecto de un valor promedio, bien sea este la media o la mediana, resultando en la práctica pequeña diferencia entre ambos resultados.

Teóricamente debe preferirse el uso de la mediana porque así se obtiene un valor mínimo para la desviación media. En caso de calcular los desvíos existentes entre los valores de la variable y la mediana se procede luego como antes se indicó, es decir, se multiplican los valores absolutos de estos desvíos por las frecuencias de clase correspondientes y su suma se divide por el número total de observaciones.

En general que el cálculo con los desvíos verdaderos resulte muy laborioso por intervenir, a veces, en él cantidades fraccionarias.

Para facilitar las operaciones conviene medir los desvios respecto de un valor arbitrario, sumarlas y corregir la suma en una cantidad igual al error cometido al tomar los desvios con relación a un valor que no es el verdadero.

Se simplificará más aún el cálculo midiendo los desvios, tomando como unidad el propio intervalo, tal como hemos hecho al calcular la media aritmética.

Desvío standart o desviación típica o desvío medio cuadrático

Generalidades.

El método que hemos seguido para calcular la desviación media carece de rigor algebraico puesto que no se tienen en cuenta los signos de los desvíos.

En el cálculo del desvío standard se evita esta causa de error, obteniéndose una medida cuyo significado es más preciso.

Para representar el desvío standard se usa la letra σ griega

Para calcular esta medida se eleva al cuadrado cada uno de los desvíos, se halla la media aritmética del producto de estos desvíos multiplicados por las respectivas frecuencias y se extrae la raíz cuadrada.

El desvío standard es, por tanto, la raíz cuadrada de la media aritmética de los cuadrados de los desvíos y se denomina también desvío medio cuadrático (desviación típica).

Esta denominación es muy conveniente porque da idea del procedimiento que se sigue para el cálculo del desvío standard.

Los desvíos se medirán siempre respecto de la media aritmética, puesto que el valor que buscamos resultará mínimo en estas condiciones.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 \, y_i}{\sum_{i=1}^{N} y_i}} \qquad (1)$$

Cálculo de la desviación típica o desvío estandard

Retomando el ejemplo anterior tenemos

X_i	Y_i	$x_i = X_i - \bar{X}$	$X_i Y_i$	x_i^2	$x_i^2 Y_i$
3	1	-6	3	36	36
6	6	-3	36	9	54
9	7	0	63	0	0
12	4	3	48	9	36
15	2	6	30	36	72
	20		180	90	198

$$\bar{X} = \frac{180}{20} = 9$$

$$x_i = X_i - \bar{X} = X_i - 9$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 y_i}{\sum_{i=1}^{N} y_i}} = \sqrt{\frac{198}{20}} = \sqrt{9,9} = 3,146$$

$$\sigma^2 = 9,9$$

Varianza o desvío medio cuadrático

Elevando al cuadrado la expresión (1) resulta

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 \, y_i}{\sum_{i=1}^{N} y_i} \qquad (2)$$

llamada varianza o desvío medio cuadrático.

Observaciones

Cuando se trabaje con los datos de una muestra en vez de tratar con el universo o población, simbolizaremos a la varianza con s^2 y por tanto la desviación típica con s .

Pero en vez de definirse la desviación típica por la fórmula (1) y la varianza por la (2); se tiene en el caso de una muestra que la desviación típica es:

$$s = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 \, y_i}{N-1}} \qquad (3)$$

y por tanto la varianza

$$s^2 = \frac{\sum_{i=1}^{N} x_i^2 \, y_i}{N-1} \qquad (4)$$

En estas expresiones se ha reemplazado el denominador $\sum_{i=1}^{N} y_i = N$ por N-1.

Esta sustitución tiene su razón de ser en el hecho de que aún cuando hay N cuadrados en los segundos miembros de (3) y (4), la suma de estos cuadrados se reduce realmente al conocimiento de la suma de N-1 cuadrados en virtud de la propiedad estudiada por la cual la suma de los cuadrados de los desvíos es igual a cero.

Por tanto, conociendo los valores de N-1 desvíos estamos en condición de determinar el valor del enésimo desvío. Es decir que para calcular (3) o (4) tenemos N-1 grados de libertad.

Por ejemplo supongamos que de una observación que consta de seis datos, conocemos cinco de sus desvíos, a saber:- -3; -1; 2; 5; 7.
En virtud de la propiedad antes citada el sexto desvío será:

$$0-(-3-1+2+5+7) = -10$$

En el caso considerado, contando con 6 observaciones tenemos 6-1 = 5 grados de libertad.

La propiedad de que goza la suma de los desvíos respecto de la media aritmética, se utiliza con ventaja en el caso de tratar con una muestra, especialmente cuando la muestra se compone de pocas observaciones, (como ocurre cuando se trata de controles de calidad, por ejemplo,) en cuyos casos se tienen 4 ó 5 elementos que analizar donde naturalmente el hecho de dividir por 4 en vez de dividir por 5 tiene gran gravitación sobre el resultado de la varianza s^2 y de la desviación típica s.

Cuando se trabaja con el universo o población total la diferencia de una unidad en el denominador no tiene influencia sobre los resultados de σ y σ^2.

Simplificación en el cálculo de la varianza y la desviación típica.

Cuando tratamos del método simplificado para calcular la media aritmética vimos que mediante un cambio de origen lográbamos gran economía en las operaciones.

El cambio de origen consistió en la siguiente operación: elegido como valor medio arbitrario Xj y siendo la amplitud de un intervalo de clase a otro constante e igual a h, hicimos

$$Xi' = \frac{Xi - Xj}{h} \qquad (1)$$

donde despejando Xi resulta:

$$Xi = Xj + h\, Xi' \qquad (2)$$

haciendo para simplificar

$$Xj = a$$
$$h = b$$

se tiene que (2) puede escribirse

$$Xi = a + b\, Xi' \qquad (3)$$

sumando todos los valores de Xi y de Xi' multiplicados por su ponderación, tenemos que la (3) se transforma en:

$$\sum_{i=1}^{N} Xi\, yi = a \sum_{i=1}^{N} yi + b \sum_{i=1}^{N} Xi'\, yi$$

dividiendo ambos miembros por $\sum_{i=1}^{N} y_i$

$$\frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} y_i} = a + b \frac{\sum_{i=1}^{N} X'_i y_i}{\sum_{i=1}^{N} y_i}$$

de donde resulta

$$\boxed{\bar{X}i = a + b \bar{X}'i} \qquad (4)$$

es decir que la media aritmética de los valores observados X es igual al valor $a = Xj$ que es un valor medio arbitrario convenientemente elegido entre los valores de X, más la media aritmética de los valores de la nueva variable de cálculo X' multiplicada por la amplitud del intervalo de clase $h = b$.

Restamos miembro a miembro de la expresión (4), la expresión (3)

$$\bar{X}i - Xi = (a + b \bar{X}'i) - (a + b X'i)$$

simplificando

$$\bar{X}i - Xi = b (\bar{X}'i - Xi)$$

elevamos ambos miembros al cuadrado, sumando y multiplicando por yi

$$\sum_{i=1}^{N} (\bar{X}i - Xi)^2 Yi = b^2 \sum_{i=1}^{N} (\bar{X}'i - Xi)^2 Yi$$

dividiendo por $\sum_{i=1}^{N} yi$ ambos miembros se tiene:

$$\frac{\sum_{i=1}^{N} (\bar{X}i - Xi)^2 yi}{\sum_{i=1}^{N} yi} = b^2 \frac{\sum_{i=1}^{N} (\bar{X}'i - Xi)^2 yi}{\sum_{i=1}^{N} yi}$$

es decir

$$\boxed{\sigma^2_x = b^2 \sigma^2_{x'}} \qquad (5)$$

Por tanto la varianza de los valores de X es igual a la varianza de los valores de cálculo X' multiplicada por el cuadrado de la amplitud del intervalo $b = h$.

Extrayendo la raíz cuadrada de ambos miembros de (5) resulta:

$$\sigma_x = b\,\sigma_{x'} \qquad (6)$$

es decir que la desviación típica de los valores de X, es igual a la desviación típica de los valores de cálculo X', multiplicada por la amplitud del intervalo de clase b = h

Si se trabajara con una muestra y recordando que la varianza se define en este caso por:

$$s^2 = \frac{\sum_{i=1}^{N} (\overline{X}_i - X_j)^2 Y_i}{\sum_{i=1}^{N} Y_i - 1}$$

procediendo en igual forma a la anterior resultará:

$$s_x^2 = b^2 s_{x'}^2 \qquad (7)$$

y la desviación típica:

$$s_x = b\, s_{x'} \qquad (8)$$

Ejemplo

Determinar los valores de $\overline{X}$, σ_x^2 y σ_x en el siguiente caso:

X_i	Y_i	$X_i' = \frac{X_i - 4.870}{0,005}$	$X_i' Y_i$	X_i^2	$X_i^2 Y$
4.830	1	-8	-8	64	64
4.835	6	-7	-42	49	294
4.840	11	-6	-66	36	396
4.845	21	-5	-105	25	525
4.850	23	-4	-92	16	368
4.855	24	-3	-72	9	216
4.860	25	-2	-50	4	100
4.865	40	-1	-40	1	40
4.870	45	0	-475	0	0
4.875	49	1	49	1	49
4.880	35	2	70	4	140
4.885	45	3	135	9	405
4.890	33	4	132	16	528
4.895	16	5	80	25	400
4.900	8	6	48	36	288
4.905	1	7	7	49	49
4.910	1	8	8	64	64
	384		529		3926
			54		

Elegimos como valor medio arbitrario Xj = 4.870; siendo h = 0,005. Trabajando con la nueva variable X_i hallamos los valores:

$$\bar{X}' = \frac{54}{384} = 0,14$$

$$\sigma_{x'}^2 = \frac{3926}{384} = 10,2239$$

$$\sigma_{x'} = 3,197$$

Por medio de las fórmulas (4) (5) y (6) determinaremos los valores de $\bar{X}$, σ_x^2 y σ_x conociendo $\bar{X}'$, $\sigma_{x'}^2$ y $\sigma_{x'}$ cuyos valores terminamos de calcular. En efecto

$$\bar{X} = 4.870 + 0,005 \times 0,14 = 4.870 + 0,0007 = 4,8707$$

$$\sigma_x^2 = \overline{0,005}^2 \sigma_{x'}^2 = 0,000025 \times 10,2239 = 0,000255$$

$$\sigma_x = 0,005 . \sigma_x = 0,005 \times 3,197 = 0,01598$$

Desviación cuartil

Localizados los cuartiles, según hemos visto, puede definirse con exactitud, el grado y carácter de la variabilidad que presenta una distribución de frecuencias.

Aunque su conocimiento nos ofrece una imagen de la distribución, no es tan útil para el estudio y comparación de aquélla como el de la desviación media y la varianza.

Pero debemos advertir que en cuanto se refiere a la mayor facilidad del cálculo y a la interpretación inmediata, la desviación cuartil presenta considerables ventajas.

Dentro de la oscilación propia de los cuartiles se halla incluída, desde luego, la mitad de todas las medidas.

Cuanto mayor sea la concentración, tanto menor será este intervalo, por lo cual la relación entre ambos cuartiles nos proporcionará una medida de dispersión bastante aproximada.

El desvío cuartil, es la oscilación intercuartil que es la mitad de la distancia entre los cuartiles primero y tercero.

Por tanto si representamos por D.Q. el desvío cuartil, por Q1 el primer cuartil y por Q3 el tercer cuartil, se tendrá:

$$D\,Q = \frac{Q3 - Q1}{2}$$

Designando por $K = \frac{Q3 + Q1}{2}$ un punto situado sobre la escala [illegible] la mitad de la distancia que haya entre los citados cuartiles, la mitad de los datos de una distribución de frecuencias se hallará comprendida en la oscilación

$$K \pm D.Q.$$

Este resultado, unido a una exposición de los valores promedios del cambio (que pueden ser la media aritmética, la mediana o el modo) constituye una descripción muy útil de la distribución que se estudia.

En una distribución rigurosamente simétrica, el valor de K coincidirá con el valor de la mediana, esto es, la mediana estará situada en el punto medio de la distancia que separa en la escala Q1 de Q3 .

Ejemplo

Xi	yi	fi
1.125 - 1.225	51	51
1.225 - 1.325	118	169
1.325 - 1.425	144	313
1.425 - 1.525	181	494
1.525 - 1.625	137	631
1.625 - 1.725	11	642
	642	

$$\frac{N}{4} = \frac{642}{4} = 160,5$$

$$Q_1 = 1.225 + \frac{160,5 - 51}{118}.0,1$$

$$= 1.225 + 0,093 = 1.318$$

$$\frac{3N}{4} = \frac{3 \times 642}{4} = 481,5$$

$$Q_3 = 1.425 + \frac{481,5 - 313}{181}.0,1$$

$$= 1.425 + 0,093 = 1,518$$

$$D.Q. = \frac{1.518 - 1.318}{2} = \frac{0,2}{2} = 0,1$$

$$K = \frac{1.518 + 1.318}{2} = \frac{2.836}{2} = 1,418$$

$$K \pm DQ \begin{cases} K + D.Q. = 1.418 + 0,1 = 1.518 \\ K - D.Q. = 1.418 - 0,1 = 1.318 \end{cases}$$

Dijimos que en el caso de simetría $K = Ma$; calculemos la Ma para confrontar su valor con el hallado para K.

$$\frac{N}{2} = \frac{642}{2} = 321$$

$$Ma. = 1.425 + \frac{321 - 313}{181}.0,1 = 1.425 + 0,004 = 1.429$$

$$Ma > K$$

luego hay leve asimetría.

Error probable

Los resultados que ofrecen las medidas dadas por las observaciones de caracteres físicos, obtenidos por diferentes investigadores al medir una misma cantidad constante, no son iguales.

Estos resultados variables se hallan distribuidos, sin embargo, en forma definida y cuando se representan gráficamente dan lugar a una curva del tipo de la curva normal de los errores.

En tales casos se advierte la perentoria necesidad de disponer de una medida de variabilidad que sirva de índice del grado de confianza que pueden merecer tales resultados.

Cuando los resultados obtenidos por varios observadores o por el mismo observador en diferentes ocasiones, varían ampliamente, no pueden concedérseles confianza alguna, mientras que

ocurre lo contrario si la variabilidad que presentan es pequeña.

La medida de dispersión que se emplea en estos casos se denomina error probable y se la define como la cantidad que, en un caso dado, queda superada por los errores de una mitad de las observaciones.

Puesto que el valor más probable en una serie de observaciones es su media aritmética, el error probable se medirá siempre respecto a esta medida.

Se debe el nombre de esta medida al hecho de que la probabilidad para que una observación dada difiera de la media de todas las observaciones en una cantidad mayor que el error probable es exactamente $\frac{1}{2}$

Se deduce, por tanto, que cuando las observaciones dan lugar a una distribución de frecuencias, cada una de las mitades del total de los casos quedará incluída dentro del resultado de agregar o quitar a la media aritmética una cantidad igual al error probable.

Esta medida de variabilidad se viene extendiendo en su aplicación a campos muy distintos de aquel para el que fué creada y entonces la denominación error probable puede resultar impropia, por lo que debe pensarse en tales casos que se trata de un desvío probable que es una distancia contada a partir de la media aritmética, inferior a la mitad del total de los desvíos.

El error probable es una medida de dispersión que solo tiene significado completo cuando se refiere a distribuciones que siguen la ley normal de los errores.

En estos casos adquiere una significación clara y precisa, lo que no ocurre cuando se aplica a distribuciones asimétricas, no siendo, por lo tanto, conveniente emplearla en tales casos.

En cambio resulta apropiado el uso del desvío cuartil, cuyo valor es igual al del error probable en toda distribución normal, porque tiene significado más directo que éste, cuando se trata de estudiar una distribución asimétrica.

Más adelante explicaremos el uso que puede hacerse del error probable como medida de la exactitud de los resultados estadísticos.

El valor del error probable puede expresarse, siempre que se trate de una distribución normal, por medio del que tiene la desviación típica ya que existe entre ambos la relación siguiente:

$$E.P. = 0{,}6745\ \sigma$$

Relaciones entre las diferentes medidas de variabilidad

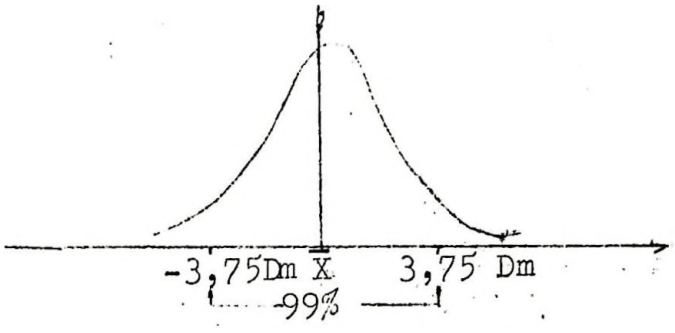
Para comprender facilmente la significación propia de cada una de las medidas de dispersión ya definidas, conviene hacer una comparación general y un encadenado resumen de las relaciones que entre ellas existen.

La oscilación es una distancia tomada sobre el eje de abscisas, que comprende todas las observaciones.

La desviación cuartil y oscilación intercuartil es una distancia tomada sobre el eje de abscisas, que está situada a ambos lados del punto medio entre los dos cuartiles, comprende la mitad del total de observaciones.

El desvío medio con respecto a la media aritmética en una distribución simétrica o ligeramente asimétrica es aproximadamente igual a los 4/5 de la desviación típica σ

Una oscilación que sea igual a $7\frac{1}{2}$ veces la desviación media y cuyo centro sea la $\bar{X}$, comprenderá aproximadamente el 99% de los casos.



Cuando en una distribución normal o levemente asimétrica, se sitúa a cada lado de la $\bar{X}$, una distancia igual a la desviación típica, quedan comprendidos los 2/3 de los casos aproximadamente, es decir el 66,66 %

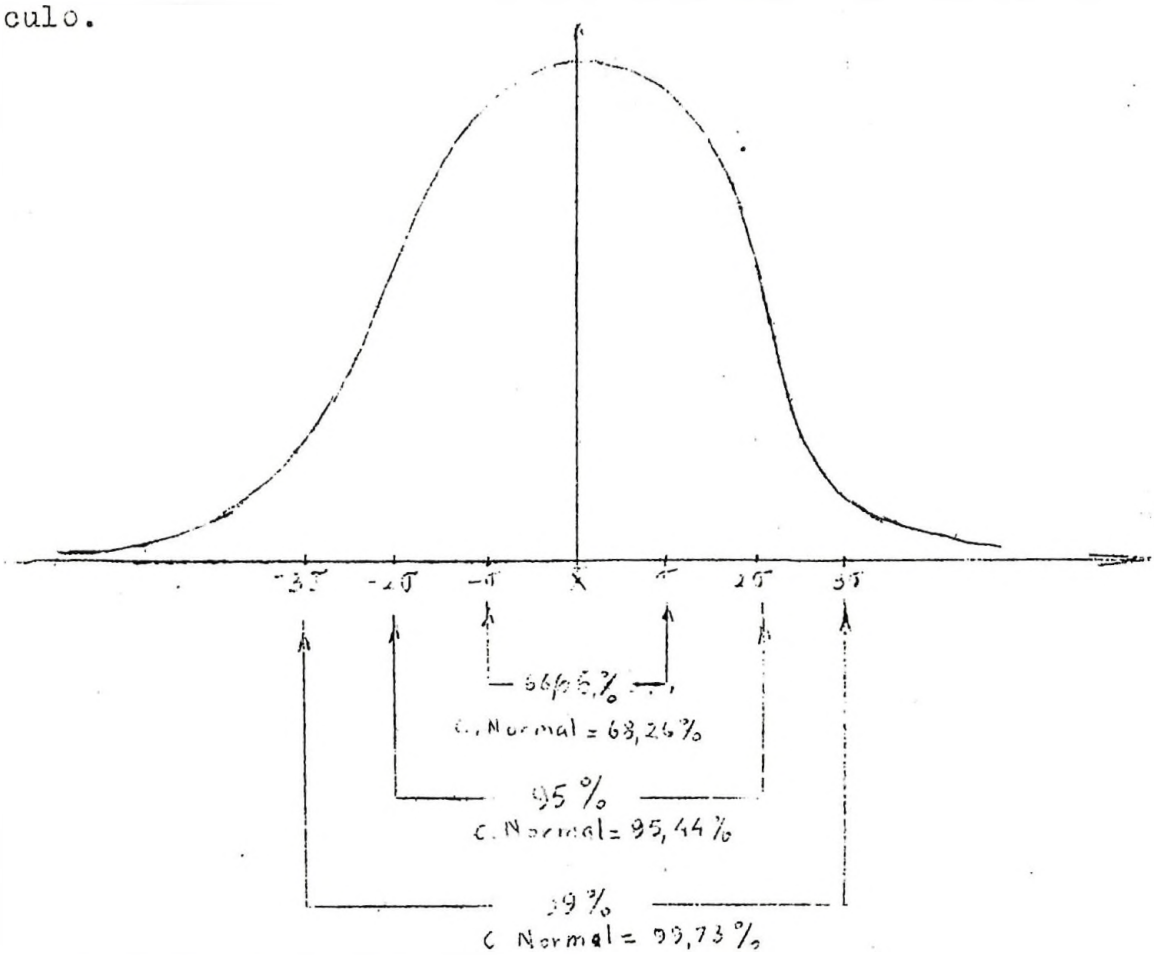
Si la distribución es normal, quedará incluído el 68, 26% de los casos.

Situando a cada lado de la media aritmética una distancia igual al doble de la desviación típica queda incluído el 95 % de los casos aproximadamente.

En la distribución normal, el tanto por ciento de inclusión será 95,46% exactamente.

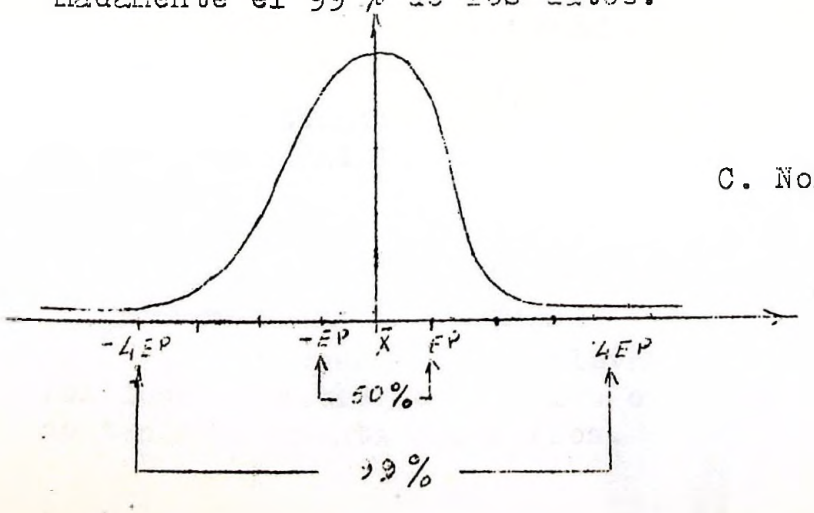
Cuando la distancia, que se toma es tres veces la desviación típica, el número de los casos incluídos es el 99% aproximadamente y en la distribución normal es 99,73% exactamente.

La regla general que dice que una oscilación igual a 6σ centrada en la media aritmética, comprende aproximadamente el 99 % de los casos, proporciona una comprobación del cálculo.



El error probable en una distribución normal es igual a $0{,}6745\ \sigma$.

Una oscilación de dos veces el error probable cuyo centro sea la media aritmética, comprenderá el 50 % de los datos y una oscilación de 8 E. P. con igual centro abarcará aproximadamente el 99 % de los datos.



C. Normal: E. P. = $0{,}6745\ \sigma$

Propiedades de las principales medidas de variabilidad

Oscilación

La oscilación puede calcularse e interpretarse fácilmente y resulta útil considerada como una medida aproximada del grado de variabilidad.

El valor de la oscilación viene determinado por el que tienen los dos casos extremos.

Es por tanto una medida sumamente inestable, cuyo valor puede cambiar por la adición o sustracción de un solo valor de la variable.

Es una medida que no proporciona indicación alguna acerca del carácter de la distribución entre los dos casos extremos.

Desviación cuartil

Es una medida de dispersión que puede calcularse e interpretarse fácilmente y resulta superior a la oscilación como medida aproximada de la variabilidad.

La desviación cuartil no es una medida de variabilidad relacionada con un promedio determinado.

Esta medida no resulta afectada por la distribución de los datos dentro del 1er. y 3er. cuartiles, ni por su distribución fuera de estos cuartiles.

Como no se halla afectada por las desviaciones de cada uno de los datos, no puede aceptarse como una medida segura de variabilidad.

No se puede aplicar el cálculo algebraico a la desviación cuartil.

Desviación media

La desviación media se halla influenciada por el valor de cada uno de los datos y tiene un significado preciso como diferencia promedio entre cada uno de los datos y la media aritmética o la mediana.

La desviación media está menos influenciada por las desviaciones extremas que la desviación típica.

Matemáticamente considerada, la desviación media no es tan lógica ni tan conveniente como la desviación típica.

Varianza

La varianza se halla influenciada por los valores de cada uno de los datos.

La operación de elevar al cuadrado los valores de los desvíos antes de sumarlos evita la falta que se comete al no tener en cuenta los signos.

La varianza tiene clara significación matemática y puede aplicarse a él el cálculo algebraico.

La varianza se halla, en general menos influenciado por las fluctuaciones de los datos que las demás medidas de dispersión.

El resultado a que ha conducido el estudio de la curva normal de los errores, por medio de la varianza, aumenta considerablemente la utilidad de esta medida.

Error probable

Tiene clara significación en el caso de una distribución que sigue una ley normal y carece de ella en otras distribuciones a cuyo estudio no debe aplicarse.

Para aquellas distribuciones a que se adapta el error probable, resulta éste de gran utilidad.

Halla su aplicación más interesante cuandose le emplea como índice de la confianza que merezcan las medidas cuantitativas.

La relación definida que existe en toda distribución normal entre el error probable y la varianza es sin embargo, la más general de todas, y deberá recurrirse a ella cuando se desee obtener gran exactitud.

El error probable es, sencillamente una parte fraccionaria de la varianza que tiene un campo de aplicación definido, al par que restringido.

MEDIDAS DE VARIABILIDAD RELATIVA

Nos hemos ocupado, hasta ahora, solo de la variabilidad absoluta y las diferentes medidas de dispersión obtenidas expresan la variabilidad de los datos considerando unidades absolutas de medida.

Así la varianza en los cambios Londres-París, viene dada en francos; la de la producción de hierro en lingotes, en toneladas, etc.

Cuando se trata de estudiar una sola distribución de frecuencias conviene que prevalezca el uso de la unidad con que se miden los datos; pero si tratamos de comparar las medidas de variabilidad de dos distribuciones distintas, se tropieza con algunas dificultades.

Esto es evidente si unas y otras medidas son heterogeneas; pero aun siendo homogéneas persisten las mismas dificultades.

Así las medidas de variabilidad que se refieren al peso de los perros y al de los caballos pueden calcularse en kilos y por el hecho que la varianza del peso de los caballos sea mayor que el correspondiente al peso de los perros no puede deducirse que la variabilidad sea mayor en el primer caso que en el segundo.

Toda medida de variabilidad absoluta tiene significado solamente con relación al promedio respecto del cual se midieron los desvíos.

Su uso aislado, esto es, sin que se refiera al promedio correspondiente, carece de sentido.

Para efectuar comparaciones será preciso, por lo tanto, dar a estas medidas forma relativa, expresándolas, como porcentajes del promedio que sirvió para hallar las — desviaciones.

Resultará así un número abstracto, una medida de la variabilidad relativa de los datos que se estudian, que puede compararse con valores similares procedentes de otras distribuciones.

Por ejemplo:

$$\frac{\text{Desvío medio}}{\text{Media aritmética}} \qquad \frac{\text{Desvío medio}}{\text{Mediana}} \qquad \frac{\text{Varianza}}{\text{Media aritmética}}$$

Coeficiente de variabilidad

La medida de variación relativa empleada más generalmente es la introducida por Pearson con la denominación de coeficiente de variabilidad, que se representa por la letra V.

Esta medida es simplemente la varianza expresada en tanto por ciento de la media aritmética.

Por tanto:

$$V = \frac{\sigma}{\bar{X}} \cdot 100$$

Pueden obtenerse índices de variabilidad análogos a éste coeficiente sin más que expresar cualquiera de las otras medidas de desviaciones en tanto por ciento del promedio con respecto del cual se han medido las desviaciones.

Sin embargo, el coeficiente de Pearson, es el que se adopta generalmente y el único cuyo uso se ha generalizado.

Medidas de asimetría

Hemos estudiado los métodos que se siguen para la determinación de la tendencia central de una distribución de frecuencias, así como para medir el grado de concentración en torno a una tendencia central.

Pero necesitamos disponer, además de una medida que exprese el grado de asimetría que presenta una distribución dada.

Para ello es esencial saber si las observaciones aparecen dispuestas simétricamente en torno del valor central o si se dispersan en forma irregular y asimétrica.

Cuando dispongamos de esta medida nos será posible resumir las características de una distribución de frecuencias en tres números que son: un promedio, una medida de dispersión y una medida de asimetría.

Hemos visto ya los promedios y las medidas de dispersión, nos ocuparemos ahora de las medidas de asime -

tría.

En toda curva de frecuencias perfectamente simétrica coinciden la media, la mediana y el modo.

A medida que la distribución deja de ser simétrica, estos tres valores se van separando, siendo cada vez mayor la diferencia que existe entre la media y el modo.

Por lo tanto, podremos usar esta diferencia para medir la asimetría.

Es conveniente, en este caso, como se hizo al medir la variabilidad relativa, obtener un índice en forma abstracta que puede compararse con los valores análogos obtenidos en otras distribuciones.

Con este fin propuso Pearson que se usará el número que resulta de dividir la diferencia entre la media aritmética y el modo, por la varianza de la distribución dada.

La fórmula será pues:

$$sk(\text{asimetría}) = \frac{\bar{X} - M_o}{\sigma} \quad (\text{skewness})$$

Claro está que en toda distribución simétrica en la que $\bar{X} = M_o$, el valor de esta medida será igual a cero.

El valor de sk puede ser negativo o positivo según la posición que tengan en la escala los dos promedios de que depende.

Cuando la distribución sea moderadamente asimétrica, el grado de su asimetría puede calcularse más fácilmente por la fórmula:

$$sk = \frac{3(\bar{X} - M_d)}{\sigma}$$

Este valor no se apartará mucho del que se obtiene con la fórmula anterior, puesto que en toda distribución moderadamente asimétrica la mediana se halla situada a un tercio de la distancia que separa la media aritmética del modo, a contar de aquella.

Como resulta difícil localizar el modo valiéndonos de métodos sencillos, conviene calcular en algunos casos

una medida de asimetría más fácil de obtener que la de Pearson.

Bowley propuso con dicho fin un método basado en la relación que existe entre el primer y tercer cuartil y la mediana.

Si la distribución es simétrica entre dos cuartiles equidistarán de la mediana pero si no lo es, no ocurrirá así.

Por tanto, si llamamos q_2 a la diferencia entre el tercer cuartil y la mediana y q_1 a la diferencia que existe entre la mediana y el primer cuartil, podremos emplear la fórmula:

$$sk = \frac{q_2 - q_1}{q_2 + q_1} = \frac{(Q_3 - M_a) - (M_a - Q_1)}{Q_3 - M_a) - (M_a - Q_1)} = \pm 1 \quad \begin{cases} = -1 \\ = 0 \\ = 1 \end{cases}$$

para obtener una medida de asimetría que variará entre ± 1. Si la simetría es perfecta se verificará que:

$$q_2 = q_1$$

y por tanto:

$$sk = 0$$

es decir que no existe asimetría.

Si la asimetría es tan acentuada que la mediana y uno de los cuartiles coinciden, resultará que si:

$q_2 = 0$ es $sk = 1$

$q_1 = 0$ es $sk = -1$

Bowley señala que el valor de 0,1 indica un grado moderado de asimetría, mientras que un valor igual a 0,3 indicará una asimetría pronunciada.

Debemos advertir que los resultados que esta medida proporciona no son comparables, desde luego, con los que se obtengan por la aplicación de la fórmula de Pearson para medir la asimetría.

Ejemplo:

Calcular la asimetría con la fórmula de Pearson y de Bowley

	x_i	y_i	f_i	$x'_i = \frac{x_i - 9}{3}$	$x'_i y_i$	$(x'_i)^2$	$(x'_i)^2 y_i$
1,5- 4,5	3	1	1	- 2	- 2	4	4
4,5- 7,5	6	6	7	- 1	- 6	1	6
7,5-10,5	9	7	14	0	- 8	0	0
10,5-13,5	12	4	18	1	4	1	4
13,5-16,5	15	2	2	2	4	4	8
		20			8		22
					0		

$$\bar{x}' = \frac{0}{20} = 0$$

$$\bar{x} = 9 + 3 \times 0 = 9$$

$$\sigma^2_{x'} = \frac{22}{20} = 1,1$$

$$\sigma^2_x = 3^2 \times 1,1 = 9 \times 1,1 = 9,9$$

$$\sigma_x = \sqrt{9,9} = 3,146$$

$$Q_1 = 4,5 + \frac{5-1}{6} \cdot 3 = 4,5 + 2 = 6,5$$

$$Q_3 = 10,5 + \frac{15 - 14}{4} \cdot 3 = 10,5 + 0,75 = 11,25$$

$$M_a = 7,5 + \frac{10 - 7}{7} \cdot 3 = 7,5 + 1,29 = 8,79$$

$$M_o = 7,5 + \frac{4}{6+4} \cdot 3 = 7,5 + 1,2 = 8,7$$

PEARSON

$$sk = \frac{\bar{x} - M_o}{\sigma} = \frac{9 - 8,7}{3,146} = \frac{0,3}{3,146} = 0,095$$

BOWLEY

$$sk = \frac{(Q_3 - M_a) - (M_a - Q_1)}{(Q_3 - M_a) + (M_a - Q_1)} = \frac{(11,25 - 8,79) - (8,79 - 6,5)}{(11,25 - 8,79) + (8,79 - 6,5)}$$

$$= \frac{2,46 - 2,29}{2,46 + 2,29} = \frac{0,17}{4,65} = 0,036$$

Curtosis

Ya hemos hablado de una cuarta característica de las distribuciones de frecuencias que también puede ser medida , nos referimos al grado de aplanamiento en relación con la curva normal

Veremos más adelante como se mide la curtosis que es la denominación técnica de esta medida

Est. - A

15a. Conferencia

MOMENTOS DE LAS DISTRIBUCIONES DE FRECUENCIAS

Al considerar la media aritmética hemos utilizado la expresión:

$$\frac{\sum_{i=1}^{N} X_i\, y_i}{\sum_{i=1}^{N} y_i}$$

podríamos tener expresiones de la forma:

$$\frac{\sum_{i=1}^{N} X_i^2\, y_i}{\sum_{i=1}^{N} y_i} \qquad \frac{\sum_{i=1}^{N} X_i^3\, y_i}{\sum_{i=1}^{N} y_i}$$

llamadas media cuadrática y media cúbica respectivamente.

En general la relación:

$$M_s = \frac{\sum_{i=1}^{N} X_i^s\, y_i}{\sum_{i=1}^{N} y_i} \qquad (2)$$

se define como momento de orden s respecto del origen natural. También se lo designa con el nombre de momento bruto de orden s .-

En particular para s = 0, 1, 2, 3, ...

$$Mo = \frac{\sum_{i=1}^{N} X_i^0\, y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i} = 1$$

(momento de orden cero)

$$M_1 = \frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} y_i} = \bar{X} \quad \text{(momento de primer orden)}$$

$$M_2 = \frac{\sum_{i=1}^{N} X_i^2 y_i}{\sum_{i=1}^{N} y_i} \quad \text{(momento de segundo orden)}$$

y así sucesivamente.

El término momento es una expresión corriente en el campo de la Mecánica y representa la medida de una fuerza en relación con su tendencia a producir rotación.

La intensidad de la tendencia depende de la distancia que exista entre el origen y el punto de aplicación de dicha fuerza.

En Estadística se usa dicha expresión en sentido análogo, considerando los grupos de frecuencias como las fuerzas en cuestión.

La distancia desde el punto medio de cada intervalo de clase al origen es aquí de importancia primordial.

Los momentos de una distribución de frecuencias respecto a cualquier origen pueden calcularse multiplicando la frecuencia de cada grupo por una potencia dada de las distancias que a lo largo del eje de abscisas las separan del origen, luego se suman los productos resultantes y se divide esta suma por el número de casos observados.

Si se trata de obtener el primer momento, se empleará la primera potencia de la distancia sobre el eje de abscisas; si se trata del cuarto momento, la cuarta potencia de dicha distancia y así sucesivamente.

Ejemplo.

Calcular los momentos M_0, M_1, M_2 y M_3 de:

la siguiente distribución de frecuencias.

X_i	Y_i	$X_i Y_i$	X_i^2	$X_i^2 Y_i$	X_i^3	$X_i^3 y$
3	1	3	9	9	27	27
6	6	36	36	216	216	1296
9	7	63	81	567	729	5103
12	4	48	144	576	1728	6912
15	2	30	225	450	3375	6750
	20	180		1818		20088

$$M_0 = 1$$

$$M_1 = \frac{180}{20} = 9$$

$$M_2 = \frac{1818}{20} = 90,9$$

$$M_3 = \frac{20088}{20} = 1004,4$$

Momentos respecto de un origen arbitrario

Hemos visto en oportunidad de calcular la media aritmética, varianza y desviación media, la necesidad de operar respecto de un origen arbitrario a fin de simplificar las operaciones.

En el caso de los momentos también conviene tomar un origen arbitrario o provisorio para m cálculo y recordando que el cambio de variable responde a la expresión:

$$X_i' = \frac{X_i - m}{h} \qquad (1)$$

donde

X_i es la variable natural

X_i' es la nueva variable o variable de cálculo

m es el origen provisorio o arbitrario

h es la amplitud de clase de la distribución.

Los sucesivos momentos respecto del origen provisorio X_i quedan definidos en general así.

$$\frac{\sum_{i=1}^{N} (X_i')^r y_i}{\sum_{i=1}^{N} y_i} = \qquad (2)$$

Resultará para:

$s = o$ $$\mu'_o = \frac{\sum_{i=1}^{N} (X'_i)^o \, y_i}{\sum_{i=1}^{N} y_i} = 1$$

$s = 1$ $$\mu'_1 = \frac{\sum_{i=1}^{N} (X'_i) \, y_i}{\sum_{i=1}^{N} y_i}$$

$s = 2$ $$\mu'_2 = \frac{\sum_{i=1}^{N} (X'_i)^2 \, y_i}{\sum_{i=1}^{N} y_i}$$

y así sucesivamente.

Como los momentos μ'_s son más fáciles de calcular pero los que necesitamos para operar son los momentos naturales M_s, resulta conveniente tener el repertorio de fórmulas que nos dan estos últimos en función de los μ'_s

En virtud de (1) y despejando X_i se tiene:

$$X_i = X'_i \, h + m \qquad (3)$$

Reemplazando este valor de X_i en (2) resulta:

$$M_s = \frac{\sum_{i=1}^{N} (X'_i \, h + m)^s \, y_i}{\sum_{i=1}^{N} y_i}$$

$$M_s = \frac{\sum_{i=1}^{N} \left[(x_i')^s h^s + s\, m (x_i')^{s-1} h^{s-1} + \frac{s(s-1)}{2!} m^2 (x_i')^{s-2} h^{s-2} + \cdots + m^s \right] y_i}{\sum_{i=1}^{N} y_i}$$

$$M_s = h^s \frac{\sum_{i=1}^{N} (x_i')^s y_i}{\sum_{i=1}^{N} y_i} + s\, m\, h^{s-1} \frac{\sum_{i=1}^{N} (x_i')^{s-1} y_i}{\sum_{i=1}^{N} y_i} + \left(\frac{s}{2}\right) m^2 h^{s-2} \frac{\sum_{i=1}^{N} (x_i')^{s-2} y_i}{\sum_{i=1}^{N} y_i} + \cdots + m^s$$

$$M_s = h^s \mu_s' + s\, h^{s-1} \mu_{s-1}' m + \left(\frac{s}{2}\right) h^{s-2} \mu_{s-2}' m^2 + \cdots + m^s$$

Dando valores sucesivos a s se tiene:

$$M_0 = \mu_0' = 1$$

$$M_1 = h\, \mu_1' + m$$

$$M_2 = h^2 \mu_2' + 2m\, h\, \mu_1' + m^2$$

$$M_3 = h^3 \mu_3' + 3 h^2 \mu_2' m + 3 h \mu_1' m^2 + m^3$$

$$M_4 = h^4 \mu_4' + 4 h^3 \mu_3' m + 6 h^2 \mu_2' m^2 + 4 h \mu_1' m^3 + m^4$$

Ejemplo:

Retomando el ejemplo anterior calcular los momentos M_0; M_1; M_2 y M_3 en función de los momentos μ_0', μ_1', μ_2' y μ_3'

X_i	Y_i	$X_i' = \frac{X_i - 9}{3}$	$X_i' Y_i$	$(X_i')^2$	$(X_i')^2 y_i$	$(X_i')^3$	$(X_i')^3 y_i$
3	1	-2	-2	4	4	-8	-8
6	6	-1	-6	1	6	-1	-6
9	7	0	0	0	0	0	-14
12	4	1	4	1	4	1	4
15	2	2	4	4	8	8	16
	20		8		22		20
			8				8

$\mu'_0 = 1$

$\mu'_1 = \frac{0}{20} = 0$

$\mu'_2 = \frac{22}{20} = 1,1$

$\mu'_3 = \frac{6}{20} = 0,3$

$M_1 = 1$

$M_1 = 3x0+9 = 9$

$M_2 = 3^2x1,1+2x9x3x0+9^2$
$= 9,9+0+81 = 90,9$

$M_3 = 3^3x0,3+3x3^2x1,1x9+0+9^3$
$= 8,1+267,3+729 = 1004,4$

Momentos con origen en la media aritmética

Los momentos respecto de la media aritmética, tienen especial importancia en Estadística y se representan por la letra μ

Definimos al momento de orden s respecto de la media aritmética por:

$$\mu_s = \frac{\sum_{i=1}^{N} x_i^s \, y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} \left(x_i - \bar{X}\right)^s y_i}{\sum_{i=1}^{N} y_i} \qquad (1)$$

donde:

$$x_i = X_i - \bar{X} \qquad (2)$$

Es decir que cuando la variable se considera referida a un origen coincidente con la media aritmética, los valores que así se obtienen no son otra cosa que los desvíos.

Resultan para:

$s = 0 \qquad \mu = \frac{\sum_{i=1}^{N} x_i^0 \, y_i}{\sum_{i=1}^{N} y_i} = 1$

$s = 1 \qquad \mu_1 = \frac{\sum_{i=1}^{N} x_i \, y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} \left(x_i - \bar{X}\right) y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} X_i \, y_i}{\sum_{i=1}^{N} y_i} - \bar{X} \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i} = \bar{X} - \bar{X} = 0$

$s = 2 \qquad \mu_2 = \frac{\sum_{i=1}^{N} x_i^2 \, y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 y_i}{\sum_{i=1}^{N} y_i} = \sigma^2$

y así sucesivamente.

El valor de μ_1 era ya conocido, pues al tratar las propiedades de la media aritmética vimos que la suma de los desvíos respecto de ella era igual a cero. Al considerar la varianza, la definimos por:

$$\frac{\sum (X_i - \bar{X})^2 y_i}{\sum y_i} = \sigma^2 = \mu_2$$

Ejemplo:

Calcular los momentos con origen en la media aritmética: μ_0, μ_1, μ_2 y μ_3 en el siguiente ejemplo:

X_i	y_i	$X_i y_i$	$x_i = X_i - \bar{X}$	$x_i y_i$	x_i^2	$x_i^2 y_i$	x_i^3	$x_i^3 y_i$
3	1	3	-6	- 6	36	36	-216	-216
6	6	36	-3	-18	9	54	- 27	-162
9	7	63	0	-24	0	0	-243	-378
12	4	48	3	12	9	36	27	108
15	2	30	6	12	36	72	216	432
	20	180		24		198	243	540
				0			0	162

$$\bar{X} = \frac{180}{20} = 9$$

$$\mu_0 = 1$$

$$\mu_1 = 0$$

$$\mu_2 = \frac{198}{20} = 9{,}9$$

$$\mu_3 = \frac{162}{20} = 8{,}1$$

Momentos respecto de la media aritmética en función de los momentos respecto de un origen provisorio

Recordemos que:

$$X_i' = \frac{X_i + m}{h}$$

luego:

$$X_i = X_i' \, h + m \qquad (3)$$

Siendo:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} (X_i' h + m)\, y_i}{\sum_{i=1}^{N} y_i} = \frac{h \sum_{i=1}^{N} X_i' y_i}{\sum_{i=1}^{N} y_i} + \frac{m \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i}$$

$$\bar{X} = h \mu_1' + m \qquad (4)$$

donde μ_1' es la media aritmética de los desvíos de la variable de cálculo o también el momento de primer orden respecto de un origen arbitrario o provisorio.

Por definición tenemos que:

$$\mu_s = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^s\, y_i}{\sum_{i=1}^{N} y_i}$$

sustituyendo X_i y $\bar{X}$ por las expresiones indicadas en (3) y (4) resulta:

$$\mu_s \frac{\sum_{i=1}^{N} (X_i' h + m - h \mu_1' - m)^s\, y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} (X_i' h - h \mu_1')^s\, y_i}{\sum_{i=1}^{N} y_i} =$$

$$= \frac{h^s \sum_{i=1}^{N} (X_i' - \mu_1')^s\, y_i}{\sum_{i=1}^{N} y_i}$$

Desarrollando la potencia del binomio se tiene:

$$\mu_s = \frac{h^s \sum_{i=1}^{N} \left\{ (X_i')^s - s (X_i')^{s-1} \mu_1' + \left(\frac{s}{2}\right) (X_i')^{s-2} (\mu_1')^2 - \ldots + (-1)^s (\mu_1')^s \right\} y_i}{\sum_{i=1}^{N} y_i}$$

$$\mu_s = h^s \left[\frac{\sum_{i=1}^{N} (X_i')^s y_i}{\sum_{i=1}^{N} y_i} - s \mu_1' \frac{\sum_{i=1}^{N} (x_i')^{s-1} y_i}{\sum_{i=1}^{N} y_i} + \left(\frac{s}{2}\right) (\mu_1')^2 \frac{\sum_{i=1}^{N} (x_i')^{s-2} y_i}{\sum_{i=1}^{N} y_i} - \cdots + (-1)^s (\mu_1')^s \right]$$

$$\mu_s = h^s \left[\mu_s' - s \mu_1' \mu_{s-1}' + \left(\frac{s}{2}\right) (\mu_1')^2 \mu_{s-2}' - \ldots + (-1)^s (\mu_1')^s \right]$$

para los sucesivos valores de s resulta:

$r = 0 \qquad \mu_0 = \mu'_0 = 1$

$r = 1 \qquad \mu_1 = h\left[\mu'_1 - \mu'_0 \mu'_1\right] = 0$

$r = 2 \qquad \mu_2 = h^2\left[\mu'_2 - 2\mu'_1\mu'_1 + \mu_0(\mu'_1)^2\right] = h^2\left[\mu'_2 - (\mu'_1)^2\right]$

$r = 3 \qquad \mu_3 = h^3\left[\mu'_3 - 3\mu'_1\mu'_2 + 3(\mu'_1)^2\mu'_1 - (\mu'_1)^3\right]$

$= h^3\left[\mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3\right]$

$r = 4 \qquad \mu_4 = h^4\left[\mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 4\mu'_1(\mu'_1)^3 + (\mu'_1)^4\right]$

$= h^4\left[\mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4\right]$

En la práctica es raro que se necesiten calcular momentos superiores al cuarto.

En realidad los momentos de orden superior al cuarto, aunque importantes en teoría son tan sensibles a las fluctuaciones de las observaciones que los valores calculados para un número regular de observaciones no ofrecen ninguna garantía y casi nunca compensan el trabajo de calcularlos.

Ejemplo:

Calcular los momentos μ_0, μ_1, μ_2 y μ_3 en función de los momentos μ'

X_i	y_i	$X_i' = \frac{X_i - 9}{3}$	$X_i' y_i$	$(X_i')^2$	$(X_i')^2 y_i$	$(X_i')^3$	$(X_i')^3 y_i$
3	1	- 2	- 2	4	4	- 8	- 8
6	6	- 1	- 6	1	6	- 1	- 6
9	7	0	- 8	0	0	0	-14
12	4	1	4	1	4	4	16
15	2	2	4	4	8	16	32
	20		8		22		48
			0				34

$\mu'_0 = 1$

$\mu'_1 = 0$

$\mu'_2 = \frac{22}{20} = 1,1$

$\mu'_3 = \frac{6}{20} = 0,3$

$\mu_0 = \mu'_0 = 1$

$\mu_1 = 3 \left\{ \mu'_1 - \mu'_0 \mu'_1 \right\} = 3 \times 0 = 0$

$\mu_2 = 3^2 \left\{ 1,1 - \right\} = 9 \times 1,1 = 9,9$

$\mu_3 = 3^3 \left\{ 0,3 - 0 + \right\} = 27 \times 0,3 = 8,1$

Est.-A
17a. Conferencia

EMPLEO DE LOS MOMENTOS PARA MEDIR LA ASIMETRIA Y LA KURTOSIS

I - Asimetría

Ciertos parámetros deducidos de los momentos calculados respecto de la media aritmética, tienen especial importancia en los trabajos estadísticos.

a) Tercer momento como medida de asimetría

La raíz cúbica del tercer momento constituye una buena medida de asimetría.

Esto es debido al hecho que si la distribución es simétrica resulta:

$$\mu_3 = \frac{\sum_{i=1}^{N} x_i^3 \, y_i}{\sum_{i=1}^{N} y_i} = 0$$

es decir que se anula el momento de tercer orden respecto de la media aritmética. Pero si la distribución no es simétrica, el tercer momento no es igual a cero y puede tener valor positivo o negativo según que la distribución presente asimetría positiva o negativa respectivamente.

Simbólicamente, la medida de asimetría definida es:

$$sk = \sqrt[3]{\frac{\sum_{i=1}^{N} x_i^3 \, y_i}{\sum_{i=1}^{N} y_i}} = \mu_3^{1/3}$$

Expresada como un coeficiente de asimetría, donde el total de la asimetría se relaciona con el desvíosstandard, resulta:

$$sk = \frac{\mu_3^{1/3}}{\sigma} \qquad (1)$$

Por las mismas razones invocadas para utilizar como medida de asimetría el momento de tercer orden respecto de la media aritmética, podrían emplearse cualquiera de los momentos de grado impar μ_5, μ_7, etc.

El motivo por el que no se recurre a ellos se fundamenta en la complicación de su cálculo.

b) Coeficientes β y γ

Se definen los coeficientes beta, así:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad (2)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad (3)$$

y los coeficientes gama así:

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} \qquad (4) \text{ siendo: } \sigma^2 = \mu_2$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2} \qquad (5)$$

Se observa que estos cuatro coeficientes son todos números abstractos y como tales, independientes de la escala de medida de la variable.

Podemos observar que la asimetría medida por la expresión (1) se deduce de la (2) en la siguiente forma:

$$\sqrt[6]{\beta_1} = \sqrt[6]{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3^{1/3}}{\mu_2^1} = \frac{\mu_3^{1/3}}{\sigma}$$

Con frecuencia se toma también como medida de asimetría al valor dado por la raíz cuadrada de (2) así:

$$\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3} = \alpha_3 = \gamma_1$$

En la curva normal que es simétrica se verifican los siguientes valores para los coeficientes β y γ

$\beta_1 = 0$ (por ser el momento de tercer orden igual a cero)

$\beta_2 = 3$

$\gamma_1 = \sqrt{\beta_1} = 0$

$\gamma_2 = \beta_2 - 3 = 3-3 = 0$

II - Kurtosis

La kurtosis fué descripta por Karl Pearson así:
"Dadas dos distribuciones de frecuencias que tengan la misma "variabilidad medida por el desvío standard, ellas pueden ser "relativamente más o menos elevadas que la curva normal..... "Una distribución de frecuencias puede, en otras palabras ser "simétrica pero puede ser más aplanada y la curva normal no "puede describirla".

Las distribuciones de frecuencias con un sólo modo, pueden ser respecto de la curva, más aplanadas o más agudas.

La kurtosis es una medida que hace posible describir el grado relativocon que esta característica existe con referencia a cualquier distribución de frecuencia.

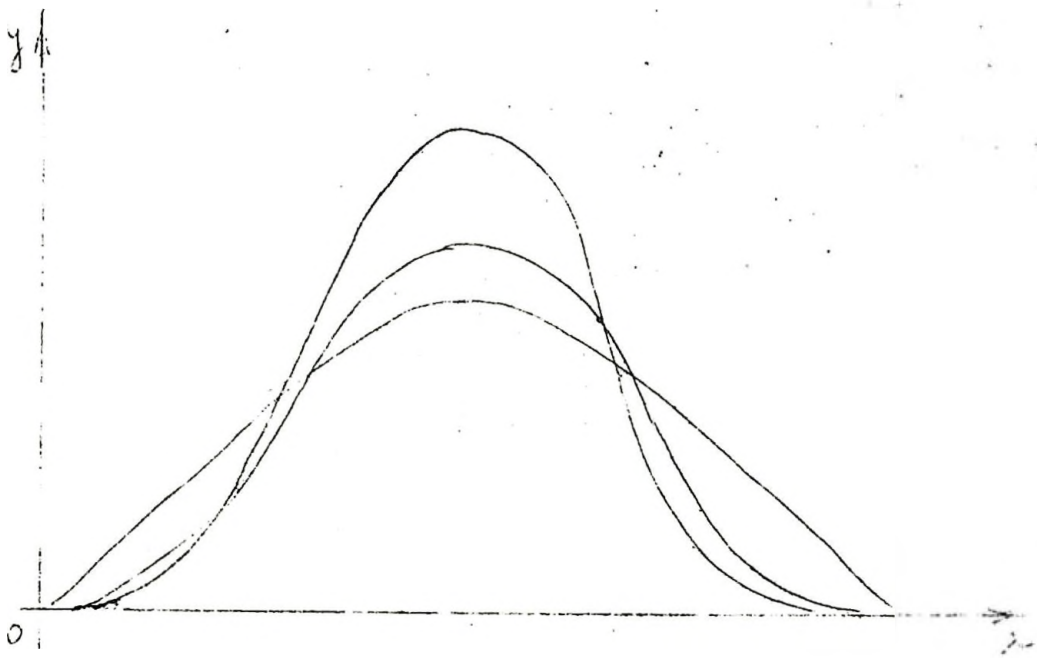
Para la curva normal, la relación entre el segundo y el cuarto momento es la siguiente:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$$

Si fuese mayor que 3 la curva es más aguda y se la designa con el nombre de leptocúrtica (aguda, estrecha) $\beta_2 > 3$

Si fuese menor que 3 la curva resulta aplanada y por eso recibe el nombre de platicúrtica (aplanada) $\beta_2 < 3$

En la distribución normal la kurtosis tiene pues el valor 3; ya que esta distribución es mirada como standard o ideal la cantidad $\beta_2 - 3$ en toda distribución se llama exceso de la kurtosis o brevemente el exceso.



Ejemplo 1º : Distribución simétrica

Determinar la asimetría y la kurtosis mediante los coeficientes β y γ_0 en el siguiente ejemplo:

x_i'	y_i	$x_i' y_i$	$(x_i')^2$	$(x_i')^2 y_i$	$(x_i')^3$	$(x_i')^3 y_i$	$(x_i')^4$	$(x_i')^4 y_i$
-2	2	- 4	4	8	-8	-16	16	32
-1	15	-15	1	15	-1	-15	1	15
0	30	-19	0	0	0	-31	0	0
1	15	15	1	15	1	15	1	15
2	2	4	4	8	8	16	16	32
	64	19		46		31		94
		0				0		

$\mu'_0 = 1$ | $\mu_0 = 1$

$\mu'_1 = 0$ | $\mu_1 = 0$

$\mu'_2 = \frac{46}{64} = \frac{23}{32}$ | $\mu_2 = \sigma^2 = \frac{23}{32}$

$\mu'_3 = 0$ | $\mu_3 = 0$

$\mu'_4 = \frac{94}{64} = \frac{47}{32}$ | $\mu_4 = \frac{47}{32}$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{\mu_2^3} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{47}{32}}{\left(\frac{23}{32}\right)^2} = 2,84312$$

$$\gamma_1 = \sqrt{\beta_1} = 0$$

$$\gamma_2 = \beta_2 - 3 = 2,84312 - 3 = -0,15688$$

Por ser $\beta_1 = 0$ la distribución es simétrica.
Por ser $\beta_2 = 2,84312$ la distribución es platicúrtica (aplanada)
El exceso de kurtosis es negativo por ser la distribución platicúrtica: $\gamma_2 = -0,15688$.

Ejemplo 2 º. Distribución asimétrica

Determinar la asimetría y la kurtosis mediante los coeficientes β y γ en el siguiente ejemplo

x_i	y_i	$x'_i\ y_i$	$(x'_i)^2$	$(x'_i)^2 y_i$	$(x'_i)^3$	$(x'_i)^3 y_i$	$(x'_i)^4$	$(x'_i)^4 y_i$
-2	2	-4	4	8	-8	-16	16	32
-1	6	-6	1	6	-1	- 6	1	6
0	7	-10	0	0	0	-22	0	0
1	4	4	1	4	1	4	1	4
2	1	2	4	4	8	8	16	16
	20	6		22		12		58
		-4				-10		

$\mu'_0 = 1$

$\mu'_1 = -\frac{4}{20} = -\frac{1}{5}$

$\mu'_2 = \frac{22}{20} = 1,1$

$\mu'_3 = -\frac{10}{20} = -\frac{1}{2}$

$\mu'_4 = \frac{58}{20} = 2,9$

$\mu_0 = 1$

$\mu_1 = 0$

$\mu_2 = 1,1 - \left(\frac{1}{5}\right)^2 = \frac{26,5}{25} = 1,06$

$\mu_3 = -\frac{1}{2} + 3.\frac{1}{5}.1,1 + 2\left(-\frac{1}{5}\right)^3 = \frac{18}{125} = 0,144$

$\mu_4 = 2,9 + 4.\frac{1}{5}.1,1 - 6\left(\frac{1}{5}\right)^2.\frac{1}{2} - 3\left(\frac{1}{5}\right)^4 = \frac{4572}{625} = 7,315$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\left(\frac{18}{125}\right)^2}{\left(\frac{26,5}{25}\right)^3} = \frac{18^2}{(26,5)^3} = 0,0178$ $\qquad \gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{18^2}{(26,5)^3}} = 0,1335$

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{4572}{625}}{\left(\frac{26,5}{25}\right)^2} = \frac{4572}{(26,5)^2} = 66,04$ $\qquad \gamma_2 = \beta_2 - 3 = 66,04 - 3 = 63,04$

Ejemplo Nº 3

1) Calcular los momentos $\mu_0, \mu_1, \mu_2, \mu_3, \mu_4$ en función de los momentos con origen provisorio.

2) Hallar el valor de los siguientes coeficientes de dispersión, de variabilidad, de asimetría y de exceso.

Estatura media en cms.	y_i	x'	$(x')^2$	$(x')^3$	$(x')^4$	$x'y_i$	$(x')^2y_i$	$(x')^3y_i$	$(x')^4y_i$
151	2	-5	25	-125	625	-10	50	-250	1250
3	6	-4	16	- 64	256	-24	96	-384	1536
5	12	-3	9	- 27	81	-36	108	-324	972
7	25	-2	4	- 8	16	-50	100	-200	400
9	50	-1	1	- 1	1	-50	50	- 50	50
161	61	0	0	0	0	170	0	1208	0
3	70	1	1	1	1	70	70	70	70
5	40	2	4	8	16	80	160	320	640
7	25	3	9	27	81	75	225	675	2025
9	8	4	16	64	256	32	128	512	2048
171	1	5	25	125	625	5	25	125	625
	300					262	1012	1702	9616
						92		494	

$\mu'_0 = 1$

$\mu'_1 = \frac{92}{300} = 0,3066$

$\mu'_2 = \frac{1012}{300} = 3,3733$

$\mu'_3 = \frac{494}{300} = 1,1466$

$\mu'_4 = \frac{9616}{300} = 32,0533$

$\mu_0 = 1$

$\mu_1 = \mu'_1 - \mu'_0\mu'_1 = 0$

$\mu_2 = h^2(\mu'_2 - [\mu'_1]^2) = 2^2\left[3,3733 - (0,3066)^2\right]$
$= 4\ (3,3733 - 0,0940) = 4 \times 3,2793$
$= 13,1172$

$\mu_3 = h^3\left[\mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3\right]$
$= 8\left[1,1466 - 3x3,3733x0,3066 + 2(0,3066)^3\right]$
$= -11,1917$

$\mu_4 = h^4\left[\mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4\right]$
$= 510,5641$

$\bar{X} = 161 + 2 \times \frac{92}{300} = 161,6133$

$\sigma = \sqrt{\mu_2} = \sqrt{13,1172} = 3,621$

$V = \frac{\sigma}{\bar{X}} \times 100 = \frac{3,62}{161,61} \times 100 = 2,24$

$\alpha_3 = \frac{\mu_3}{\sigma^3} = \sqrt{\beta_1} = -\frac{11,1917}{3,621^3} = -0,235$

$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{510,5641}{(3,621)^4} = 2,969$

Est. - A

18a. Conferencia
5 de abril de 1955

PROBABILIDADES

Hemos tratado hasta aquí de describir una muestra dada en mediciones cuantitativas mediante: I) Un valor de concentración , (promedio o medida de posición): II) La dispersión respecto del valor de concentración; III) El valor de la asimetría; IV) El valor de la kurtosis.

Ahora deseamos completar esta descripción con inferencias sobre las poblaciones de que proceden las muestras, es decir, nos interesa establecer conclusiones acerca de los intervalos, en los cuales caerán probablemente las medias, desviaciones típicas y otros parámetros de la población. Para lograrlo debemos saber algo acerca de como varían, de una muestra a otra, aquellos valores , cuando se han extraído de una misma población muestras repetidas de un tamaño establecido.

Puede abordarse el estudio de las fluctuaciones de muestra a muestra en forma experimental o matemática.

La primera forma resulta lenta y costosa por lo que se la reemplaza por la segunda.

La aproximación matemática al estudio de las leyes del muestreo se fundamenta en la teoría de probabilidades.

Probabilidad: La probabilidad de un suceso A de presentarse un número de veces m sobre el total de casos observados n, está dado por el cociente m/n. Es decir que el número de casos favorables al número considerado sobre el número de casos posibles nos da la probabilidad buscada siempre que todos los casos sean igualmente posibles.

Designaremos por p la probabilidad favorable a un suceso, así:

$$p = \frac{m}{n}$$

Cuando todos los casos son favorables a un acontecimiento, su probabilidad se convierte en la certeza y su expresión se hace igual a la unidad. Cuando todos son desfavorables, su probabilidad es igual a cero.

Así, pues, en la evaluación de la probabilidad de un suceso necesitaremos conocer dos números: el número de casos posibles y el número de casos favorables o sea aquellos en que el suceso pueda ocurrir.

Ejemplos:

I) En una tirada de un dado cuál es la probabilidad de que el número de puntos obtenidos no exceda de 4? Cuál es la probabilidad de sacar un número par?

a) Al Nº 4 le anteceden el 1,2, y 3 que con el 4 forman 4 casos favorables, luego :

$$p = \frac{4}{6} \quad \frac{2}{3}$$

b) Los Nºs pares del dado son 2, 4 y 6 luego:

$$p = \frac{3}{6} = \frac{1}{2}$$

II) Si sacamos una carta de un mazo de barajas españolas ¿cuál es la probabilidad de no obtener una espada? ¿de obtener el 2 o el 5 de bastos ? ¿de obtener un 2 o un 5 de un cierto palo? ¿de obtener una carta con un 2 o un 5?

a) no espada ..... = 30 $\therefore p = \frac{30}{40} = \frac{3}{4}$

b) 2 ó 5 de bastos = 2 $\therefore p = \frac{2}{40} = \frac{1}{20}$

c) 2 ó 5 de un palo = 1 $\therefore p = \frac{2}{40} = \frac{1}{20}$

d) 2 ó 5 hay 8..... $\therefore p = \frac{8}{40} = \frac{1}{5}$

III) Se arrojan dos dados ¿cuál es la probabilidad de obtener un total 7 puntos?

Total de casos distintos 6x6=36
Casos favorables a 7 puntos = 6

primer dado	segundo dado
1	6
2	5
3	4
4	3
5	2
6	1

luego:

$$p = \frac{6}{36} = \frac{1}{36}$$

IV) Una moneda se arroja al azar tres veces sucesivamente, 1) Cuál es la probabilidad de obtener dos caras? 2) Cuál es la probabilidad de obtener cruces por lo menos una vez?

1) Primer tiro: dos casos posibles + ó x 2 — 2
Segundo tiro: cuatro casos posibl. + + ó x + — 2x2= 4
x x
Tercer tiro: ocho casos posibl. + + + x
+ + x
+ x +
+ x x
x + + — 2x2x2= 8
x + x
x x +
x x x

Luego siendo casos favorables a dos caras = 4
Luego siendo casos posibles = 8

$$p = \frac{4}{8} = \frac{1}{2}$$

2) Casos favorables a una cruz = 7
Casos posibles = 8 $\Big\{ p = \frac{7}{8}$

V) Se sacan al azar dos cartas de un mazo de cartas francesas bien barajado Cuál es la probabilidad de que las dos cartas que se han sacado sean ases?

La 1ª carta puede sacarse de 52 maneras distintas.
La 2ª carta puede sacarse de 51 maneras distintas.
Luego casos posibles 52 x 51 .

Casos favorables

4 maneras de sacar el primer as en la primera tirada.
3 maneras de sacar el primer as en la segunda tirada.

Por lo tanto, casos favorables 4x3.

$$p = \frac{4 \times 3}{52 \times 51} = \frac{1}{221}$$

VI) Dos cartas se sacan de un mazo completo, volviéndose a poner en el mismo la primera carta antes de sacar la segunda. Cuál es la probabilidad de que ambas cartas extraídas pertenezcan a un mismo palo?

Casos posibles: 1ª = 52, 2ª = 52 $\Big\}$ 52 x 52

Casos favorables: 1ª = 13, 2ª = 13 $\Big\}$ 13 x 13

$$\Big\} p = \frac{13 \times 13}{52 \times 52} \quad \frac{1}{16}$$

VII) Una urna contiene tres bolillas blancas y cinco negras. Si saca una bolilla Cuál es la probabilidad de que sea negra?

Casos posibles = 8 | $p = \frac{5}{8}$
Casos favorables = 5 |

En los ejemplos dados la enumeración de los casos no presenta dificultades. Ahora veremos como analizar algunos problemas en los que la enumeración no es tan obvia pero puede ser simplificada haciendo uso del análisis combinatorio.

Problema:

Una urna contiene a bolillas blancas y b bolillas negras. Si se sacan de la urna $\alpha + \beta$ bolillas, encontrar la probabilidad de que entre ellas haya exactamente α blancas y β negras.

Sin distinguir el orden, el número total de maneras distintas de obtener $\alpha + \beta$ bolillas del total a + b, se expresa $C^{\alpha+\beta}_{a+b}$ siendo éste el número de casos posibles.

Casos favorables $\alpha = C^{\alpha}_{a}$ | $C^{\alpha}_{a} \cdot C^{\beta}_{b}$
$\beta = C^{\beta}_{b}$ |

luego

$$p = \frac{C^{\alpha}_{a} \cdot C^{\beta}_{b}}{C^{\alpha+\beta}_{a+b}}$$

Si en la urna hay 15 bolillas blancas y 10 bolillas negras ¿Cuál es la probabilidad de sacar tres bolillas blancas y dos bolillas negras?

$$p = \frac{C^{3}_{15} \cdot C^{2}_{10}}{C^{5}_{25}} = \frac{\frac{15 \times 14 \times 13}{1 \times 2 \times 3} \cdot \frac{10 \times 9}{1 \times 2}}{\frac{25 \times 24 \times 23 \times 22 \times 21}{1 \times 2 \times 3 \times 4 \times 5}} = \frac{39}{253}$$

Est. - A

19a. Conferencia

Teoremas fundamentales

A veces la enumeración de los casos se hace difícil y la aplicación directa de la definición resulta muy complicada.

Esta complicación se evita utilizando dos teoremas conocidos con los nombres de Teorema de las probabilidades totales y Teorema de las probabilidades compuestas.

Veamos antes de entrar a considerar el primer teorema lo que entendemos por acontecimientos "mutuamente excluyentes" o "incompatibles". Los acontecimientos se llaman mutuamente excluyentes o incompatibles si la presentación de uno de ellos excluye la presentación de los otros.

Así por ejemplo los cuatro sucesos que conciernen al número de puntos de los dados 1, 2, 3, 4 son excluyentes puesto que si se presenta uno de ellos no puede presentarse el otro.

Por el contrario, los sucesos son compatibles si es posible que se produzcan simultaneamente. Por ejemplo, la obtención de cinco puntos en un dado y de cinco puntos en otro dado, son compatibles, puesto que al arrojar dos dados, es posible obtener cinco puntos en cada uno.

Es decir, que dos sucesos dados pueden verificarse del modo siguiente:

a) el uno o el otro y
b) el uno y el otro.

En el primer caso los sucesos se excluyen, en el segundo son compatibles.

El primer caso nos lleva al teorema de las probabilidades totales y el segundo al de las probabilidades compuestas.

I) Teorema de las probabilidades totales

Supongamos que se tiene una urna con bolillas:

coloradas	c	en número igual a (c)
blancas	b	" " " a (b)
azules	a	" " " a (a)

Extraemos al azar una bolilla y queremos saber cuál es la probabilidad de que sea a o b

Casos favorables = (a) + (b)
Casos posibles = (a) + (b) + (c) = N

luego

$$P_{a\ ó\ b} = \frac{(a)+(b)}{(a)+(b)+(c)} = \frac{(a)}{N} + \frac{(b)}{N}$$

Pero $\frac{(a)}{N}$ y $\frac{(b)}{N}$ son las probabilidades iniciales correspondientes a los sucesos a y b respectivamente de modo que el Teorema de las probabilidades totales dice:

"La probabilidad de que se produzca uno de los acontecimientos a, b, c,... mutuamente excluyentes es igual a la suma de las probabilidades de cada uno de los acontecimientos"

$$\boxed{P_{a\ ó\ b} = p_a + p_b}$$

Razonando en forma análoga se tendrá:

$$P_{a\ ó\ b\ ó\ c} = p_a + p_b + p_c = 1$$

este resultado se utiliza como control de los cálculos.

II) Teorema de probabilidad compuesta

Consideremos dos urnas con bolillas a, b, c,...,la primera contiene $(a)_1$, $(b)_1$, $(c)_1$,... y la segunda $(a)_2$, $(b)_2$, $(c)_2$... siendo:

$$(a)_1 + (b)_1 + (c)_1 + \ldots = N_1$$
$$(a)_2 + (b)_2 + (c)_2 \ldots = N_2$$

Extraemos una bolilla de la primera urna y otra de la segunda, Cuál es la probabilidad de obtener el resultado a y b?

Casos posibles $N_1 \times N_2$
Casos favorables $(a)_1 \times (b)_2$

$$P_{a\ y\ b} = \frac{(a)_1 \times (b)_2}{N_1 \times N_2} = \frac{(a)_1}{N_1} \cdot \frac{(b)_2}{N_2} = p_a^{(1)} \times p_b^{(2)}$$

luego

$$\boxed{P_{a\ y\ b} = p_a^{(1)} \times p_b^{(2)}} \qquad (1)$$

Si no nos interesa que la bolilla a se extraiga de la primera urna, el resultado favorable podría presentar las formas a, b ó bien b, a.

Indicando con P_{ab} esta probabilidad de obtener a y b

en cualquiera de los dos modos será

$$P_{ab} = P_a \text{ y } b + P_b \text{ y } a \qquad (2)$$

pero

$$P_b \text{ y } a = \frac{(b)_1 \times (a)_2}{N_1 \times N_2} = p_b^{(1)} \times p_a^{(2)} \qquad (3)$$

de donde en (2) será en virtud de (1) y (3)

$$P_{ab} = p_a^{(1)} \times p_b^{(1)} + p_b^{(2)} \times p_a^{(2)}$$

El teorema de la probabilidad compuesta dice:

La probabilidad de que ocurran simultáneamente los sucesos a y b está dada por el producto de la probabilidad no condicional del acontecimiento a por la probabilidad condicional de b, supuesto que a haya ocurrido realmente.

Ejemplos:

1) Cuál es la probabilidad de obtener una suma superior a 6 tirando dos dados una sola vez?

Casos posibles: $6 \times 6 = 36$

Casos favorables:

a7 = 6
a8 = 5
a9 = 4
a10 = 3
a11 = 2
a12 = 1

$$P = p_7 + p_8 + p_9 + p_{10} + p_{11} + p_{12}$$

$$= \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$

$$= \frac{21}{36} = \frac{7}{12}$$

2) Tres urnas contienen respectivamente una bolilla blanca y dos negras; tres blancas y una negra; dos blancas y tres negras. Se saca una bolilla de cada urna Cuál es la probabilidad de que entre las bolillas sacadas haya dos blancas y una negra?

1b 2n	3b 1n	2b 3n

Casos posibles $= 3 \times 4 \times 5$

Casos favorables

b b n
b n b
n b b

$$P_{2b\ y\ n} = p_b^{(1)} \cdot p_b^{(2)} \cdot p_n^{(3)} + p_b^{(1)} \cdot p_n^{(2)} \cdot p_b^{(3)} + p_n^{(1)} \cdot p_b^{(2)} \cdot p_b^{(3)}$$

$$= \frac{1}{3} \cdot \frac{3}{4} \cdot \frac{3}{5} + \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{2}{5} + \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{2}{5}$$

$$= \frac{3}{20} + \frac{1}{30} + \frac{1}{5} = \frac{9 + 2 + 12}{60} = \frac{23}{60}$$

3) Un bolillero contiene dos bolillas blancas, una negra y siete rojas, se pregunta:

I) Cuál es la probabilidad de sacar una bolilla blanca y luego una roja a) reponiendo la bolilla extraída, b) sin reponerla.

II) Cuál es la probabilidad de sacar una bolilla blanca y una roja sea cual fuere el órden de presentación; a) con reposición b) sin reposición.

I) $p_{b\ y\ r} = p_b \times p_r$

a) $p_b = \frac{2}{10} = \frac{1}{5}$; $p_r = \frac{7}{10}$

$$p_{b\ y\ r} = \frac{2}{10} \cdot \frac{7}{10} = \frac{7}{50}$$

b) $p_b = \frac{2}{10} \quad \frac{1}{5}$; $p_r = \frac{7}{9}$

$$p_{b\ y\ r} = \frac{1}{5} \cdot \frac{7}{9} = \frac{7}{45}$$

II) $p_{br} = p_{b\ y\ r} + p_{r\ y\ b}$

a) $p_{b\ y\ r} = \frac{7}{50}$; $p_{r\ y\ b} = \frac{7}{50}$

$$p_{br} = \frac{7}{50} + \frac{7}{50} = \frac{14}{50} = \frac{7}{25}$$

b) $p_{b\ y\ r} = \frac{7}{45}$; $p_{r\ y\ b} = \frac{7}{10} \cdot \frac{2}{9} = \frac{7}{45}$

$$p_{br} = \frac{7}{45} + \frac{7}{45} = \frac{14}{45}$$

Est. - A

20a. Conferencia

Variables aleatorias

Al arrojar un dado se pueden obtener seis valores dis-distintos con las probabilidades

$$p_1,\ p_2,\ p_3,\ p_4,\ p_5,\ p_6 = \frac{1}{6}$$

Es una magnitud aleatoria de 1. orden, definida por los seis pares de valores (x_i, p_i)

x_i	$p_i = f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

donde x_i son los valores que toma una variable aleatoria $\underline{x}$ y $\underline{p_i}$ son las probabilidades correspondientes.

Si se tiran simultáneamente dos dados y se efectúa la suma de los puntos, esta suma es una variable aleatoria de 11º orden definida totalmente por los once pares de valores siguientes:

x_i	$p_i = f(x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Si se juega a cara o cruz tirando una moneda y un jugador ganara $ 1 cada vez que sale cruz y nada si sale cara, en cada jugada, se define una variable aleatoria de 2º orden, la ganancia del jugador, mediante los dos pares de valores:

x_i	$p_i = f(x)$
0	$\frac{1}{2}$
1	$\frac{1}{2}$.

Si se conviene jugar 15 veces la ganancia total al final de las 15 jugadas es una variable aleatoria de 15º orden definida por los 15 pares de valores:

x_i	$p_i = f(x)$
0	p_0
1	p_1
2	p_2
3	p_3
4	p_4
5	p_5
6	p_6
7	p_7
8	p_8
9	p_9
10	p_{10}
11	p_{11}
12	p_{12}
13	p_{13}
14	p_{14}

Definición de variable aleatoria

Variable aleatoria X es aquella variable que puede tomar valores reales en correspondencia con sucesos incompatibles entre sí y con probabilidades cuya suma es igual a la unidad.

El orden de la variable aleatoria queda determinado por el número de pares de valores que es necesario para definirla totalmente.

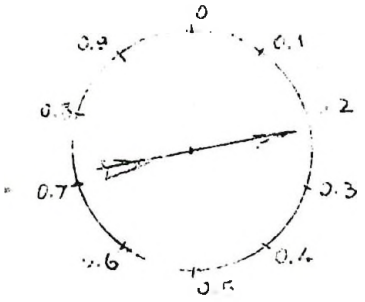
Una variable aleatoria X queda definida entonces si 1) se da la serie de sus posibles valores x_i y 2) tambien se da la probabilidad p que tiene la variable X al tomar cada valor particular x

El conjunto de pares (x_i, p_i) da la ley de repartición o la distribución de probabilidad de la magnitud X considerada.

Variable aleatoria discreta y continua

Cuando X puede tomar solamente ciertos valores aislados de un intervalo diremos que X es una variable aleatoria discreta. Así vimos que en el problema de los dados o de las monedas, X solo puede tomar valores aislados.

La variable aleatoria será continua cuando pueda tomar cualquier valor de un intervalo. Por ej. consideremos una escala circular de 360° de longitud unidad, con una aguja que se mueve libremente sobre un pivote en el centro.

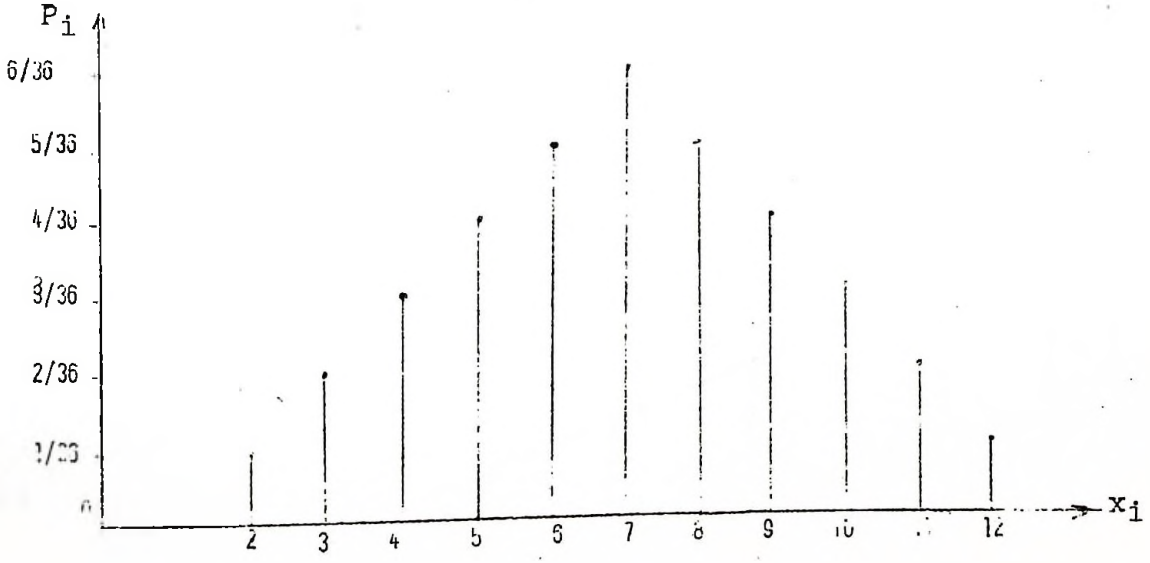


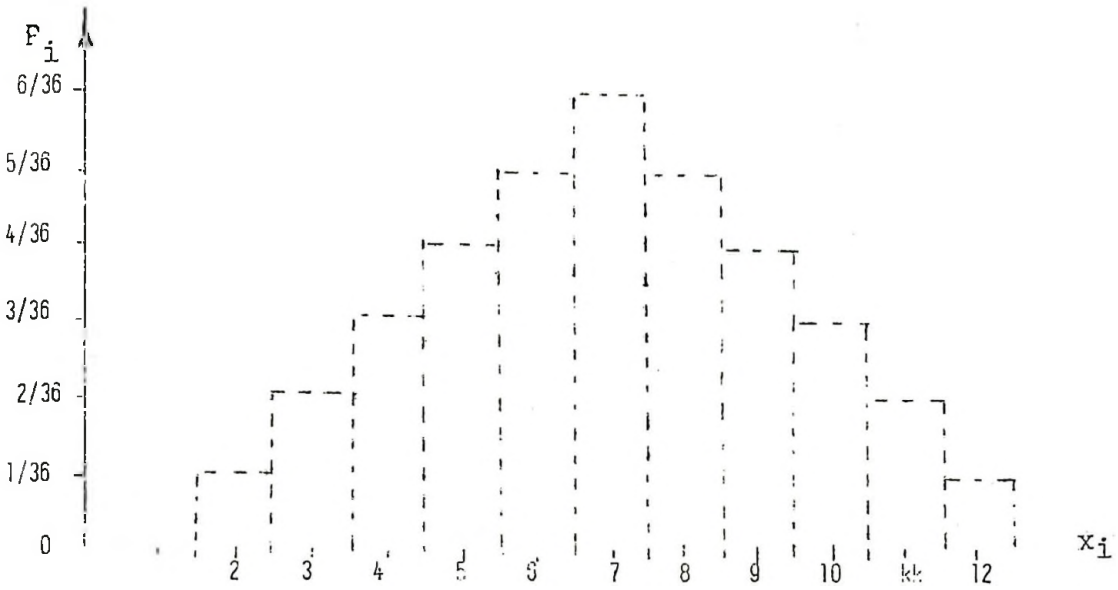
Cuando la aguja se detiene después de girar libremente, es "igualmente probable" que se detenga en cualquier sitio. En este caso midiendo la distancia desde el punto cero, sea X una variable aleatoria continua ya que puede tomar cualquier valor comprendido entre 0 y 1.

Representación de la distribución de probabilidad.

Los valores dados en los ejemplos contenidos en las tablas anteriores, pueden representarse gráficamente mediante un diagrama de barras de probabilidad o por medio de un histograma de probabilidad.

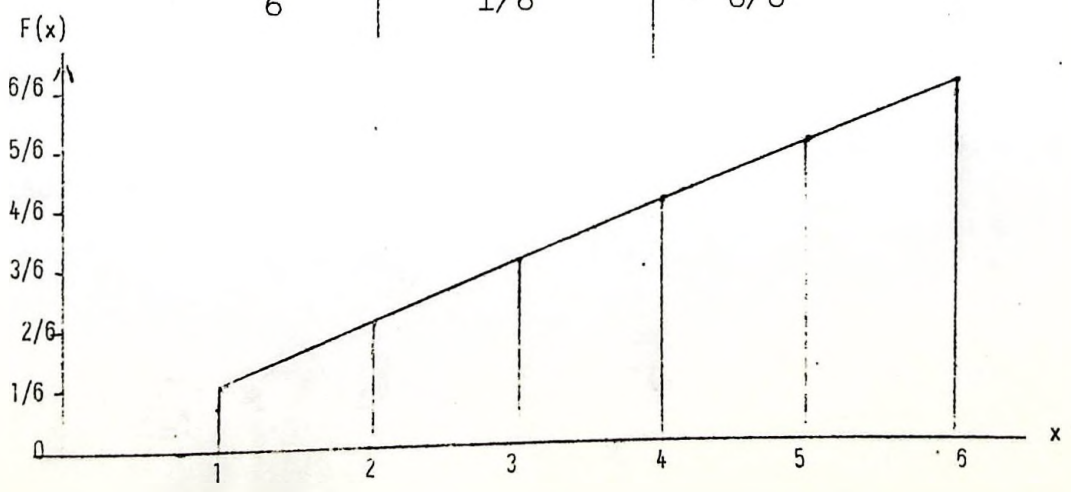
Representando en ambas formas los pares de valores de la tabla segunda se tiene:





Es útil representar las probabilidades acumulativas obtenidas de la tabla de probabilidades acumulativas F(x). Trabajando con el ejemplo primero, tendríamos:

x_i	$p_i = f(x)$	F(x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6



Puede tratarse, para su análisis a una distribución de probabilidades como a una distribución de frecuencias relativas. En tal virtud podrá calcularse, en una distribución de probabilidades, valores medios, varianzas, desviaciones típicas y demás parámetros que se determinaban en las distribuciones de frecuencias relativas.

Media aritmética o esperanza matemática

La media aritmética de una variable aleatoria discreta X que tiene una distribución de probabilidad f(x) está dada por:

$$\mu = x_1 f(x_1) + x_2 f(x_2) + x_3 f(x_3) + \ldots + x_n f(x_n)$$

$$\mu = \sum_{i=1}^{n} x_i f(x_i) \qquad (1)$$

que suele escribirse más brevemente:

$$\mu = E(x) \qquad (2)$$

μ es la media aritmética o esperanza matemática de X cuya distribución de probabilidad es f(x).

Varianza

La varianza σ^2 de una variable aleatoria discreta X que tiene una distribución de probabilidad f(x) es igual a:

$$\sigma^2 = (x_1 - \mu)^2 f(x_1) + (x_2 - \mu)^2 f(x_2) + \ldots + (x_n - \mu)^2 f(x_n)$$

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) \qquad (3)$$

Una forma más conveniente para σ^2 se tiene desarrollando el cuadrado en (3)

$$\sigma^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\mu + \mu^2) f(x_i)$$

$$= \sum_{i=1}^{n} x_i^2 f(x_i) - 2\mu \underbrace{\sum_{i=1}^{n} x_i f(x_i)}_{\mu} + \mu^2 \underbrace{\sum_{i=1}^{n} f(x_i)}_{1}$$

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2 \qquad (4)$$

o en virtud de (2)

$$\sigma^2 = E(x^2) - [E(x)]^2$$

Desviación típica

Está dada por la raíz cuadrada de la varianza; es decir

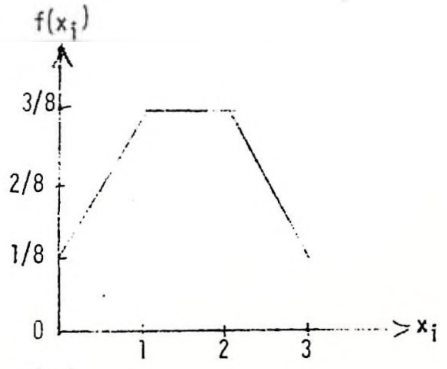
$$\sqrt{\sigma^2} = \sigma$$

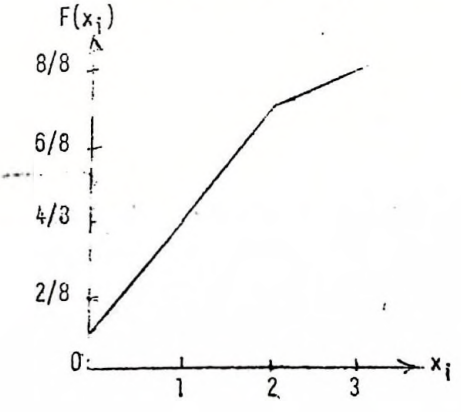
Ejercicios

1.- Se lanzan tres monedas. Sea X la variable aleatoria que representa el Nº de caras obtenidas. Preséntese la distribución de probabilidad de X en forma de tabla. Dibújese un diagrama de barras y un gráfico de probabilidad acumulativa para la distribución. Determínese la media y la varianza de X.

+ + +
+ + x
+ x +
+ x x
x + +
x + x
x x +
x x x

x_i	$f(x_i)$	$F(x_i)$
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8





x_i	$f(x_i)$	$x_i f(x_i)$	x_i^2	$x_i^2 f(x_i)$
0	1/8	0	0	0
1	3/8	3/8	1	3/8
2	3/8	6/8	4	12/8
3	1/8	3/8	9	9/8
		12/8		24/8
		3/2=1,5		3

$\sigma^2 = 3$

Est. - A

2la. Conferencia

Esperanza matemática

Llamamos esperanza matemática de $\underline{x}$ al valor

$$E(x) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i \qquad (1)$$

$$= x_1 f(x_1) + x_2 f(x_2) + \ldots + x_n f(x_n) = \sum_{i=1}^{n} x_i f(x_i)$$

Esperanza matemática de una suma

Teorema: La esperanza matemática de la suma de varias variables es igual a la suma de sus esperanzas matemáticas

$$E(x + y + z \ldots + w) = E(x) + E(y) + E(z) + \ldots + E(w)$$

Consideremos primero el caso de la suma de dos variables. Supongamos que $\underline{x}$ tome los valores numéricos diferentes $x_1, x_2 \ldots x_m$ y que la variable $\underline{y}$ tome los valores numéricos diferentes $y_1, y_2, y_3, \ldots y_n$. Respecto a la suma $x + y$ podemos distinguir $m.n$ casos mutuamente excluyentes; esto, cuando $\underline{x}$ toma un valor definido $\underline{x_i}$ e $\underline{y}$ otro valor definido $\underline{y_j}$ mientras que $\underline{i}$ y $\underline{j}$ están comprendidos respectivamente entre $\underline{1}$ y $\underline{m}$ y $\underline{1}$ y $\underline{n}$, pudiendo ser cualquiera de ellos.

Si representamos por p_{ij} la probabilidad de coexistencia de $x = x_i$, $y = y_j$, tenemos en virtud de la definición de esperanza matemática que

$$E(x + y) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} (x_i + y_j)$$

o bien

$$E(x + y) = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} x_i + \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} y_j \qquad (2)$$

Como la variable $\underline{x}$ toma un valor definido $\underline{x_i}$ de $\underline{m}$ maneras mutuamente excluyentes es evidente que la suma

$$\sum_{j=1}^{n} p_{ij} = p_i$$

De igual modo resulta : $\sum_{i=1}^{m} p_{ij} = p_j$

por lo tanto en (2) se tiene:

$$E(x+y) = \sum_{i=1}^{m} x_i p_i + \sum_{j=1}^{n} y_j p_j$$

de donde resulta:

$$E(x+y) = E(x) + E(y) \qquad (3)$$

Si consideramos ahora tres variables $x + y + z$ podemos encarar el problema como si se tratase en primer lugar de la suma de $x+y$ y z, entonces aplicando el teorema anterior:

$$E(x+y+z) = E(x+y) + E(z)$$

reemplazando $E(x+y)$ por su igual (3) se tiene:

$$E(x+y+z) = E(x) + E(y) + E(z)$$

De igual manera se demuestra el teorema para la suma de cualquier número de variables.

Esperanza matemática de un producto

Podemos establecer un teorema análogo al considerado anteriormente referente a esperanza matemática del producto de variables aleatorias sólo en condiciones muy restringidas.

Varias variables aleatorias se llaman "independientes" si la probabilidad de cualquiera de ellas de adquirir un valor determinado no depende de los valores que han tomado las variables restantes. Así por ejemplo, si las variables son los números de puntos de un dado, pueden considerarse como independientes.

Si más de dos variables son independientes de acuerdo con la definición anterior, es evidente que las de cualquier par de ellas son independientes entre sí.

Para dos variables independientes se verifica el siguiente teorema:

Teorema: La esperanza matemática del producto xy de dos variables independientes x e y es igual al producto de sus esperanzas individuales.

$$E(xy) = E(x) \cdot E(y)$$

Sean

$x_1, x_2 \ldots x_m$ los valores de x

e

$y_1, y_2 \ldots y_n$ los valores de y

Representando la probabilidad de x_i por p_i
y " " " " y_j por q_j

Por el teorema de la probabilidad compuesta, la verificación simultánea de los acontecimientos

$$x = x_i \quad e \quad y = y_j$$

que son independientes, tiene la probabilidad $p_i\, q_j$. Además por la definición de esperanza matemática

$$E(xy) = \sum_{i=1}^{m} \sum_{j=i}^{n} p_i\, q_j\, x_i\, y_j$$

realizando la suma primero respecto de j manteniendo i constante, tenemos

$$\sum_{j=1}^{n} p_i\, q_j\, x_i\, y_j = p_i\, x_i \sum_{j=1}^{n} q_j\, y_j = p_i\, x_i\, E\,(y)$$

y también

$$E\,(xy) = \sum_{i=1}^{m} p_i x_i\, E\,(y) = E(y) \sum_{i=1}^{m} p_i x_i$$

$$E\,(xy) = E\,(x)\, E\,(y)$$

De manera análoga se generaliza el teorema a cualquier número de factores independientes en su totalidad.

Así, si x, y, z, son independientes, es obvio que también lo son xy y z. Por lo tanto:

$$E(xyz) = E\,(xy)\, E\,(z)$$

y además

$$E(xyz) = E\,(x)\, E\,(y)\, E\,(z)$$

Ejemplo:

Se tiran dos dados: uno común y el otro en el que se han marcado los siguientes puntos: 1, 3, 5, 7, 9, 11. Calcular la esperanza matemática del producto de los puntos que aparecerán en ambos dados al ser arrojados simultáneamente:

$x_1 = 1, 2, 3, 4, 5, 6$

$p_1 = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$$E(x) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$$

$y_1 = 1, 3, 5, 7, 9, 11$

$p_y = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

$$E(y) = \frac{1}{6}(1 + 3 + 5 + 7 + 9 + 11) = 6$$

luego

$$E(x.y) = E(x) \cdot E(y) = \frac{7}{2} \times 6 = 21$$

El cálculo directo nos daría:

y \ x	1	2	3	4	5	6
1	1	2	3	4	5	6
3	3	6	9	12	15	18
5	5	10	15	20	25	30
7	7	14	21	28	35	42
9	9	18	27	36	45	54
11	11	22	33	44	55	66

xy	f(x) f(y)	xy f(x) f(y)
1	1/36	1/36
2	1/36	2/36
3	2/36	6/36
4	1/36	4/36
5	2/36	10/36
6	2/36	12/36
7	1/36	7/36
9	2/36	18/36
10	1/36	10/36
11	1/36	11/36
12	1/36	12/36
14	1/36	14/36
15	2/36	30/36
18	2/36	36/36
20	1/36	20/36
21	1/36	21/36
22	1/36	22/36
25	1/36	25/36
27	1/36	27/36
28	1/36	28/36
30	1/36	30/36
33	1/36	33/36
35	1/36	35/36
36	1/36	36/36
42	1/36	42/36

44	1/36	44/36
45	1/36	45/36
54	1/36	54/36
55/	1/36	55/36
66	1/36	66/36
	36/36	756/36 = 21

Luego:

$$E\,(xy) = 21$$

Est. - A

22a. Conferencia

Distribuciones teóricas - Distribució binomial - Distribución Normal - Distribución de Poisson.

Las distribuciones de frecuencia con que hemos trata do hasta aquí, han sido el resultado de observaciones o expe - riencias.

Pero también, es posible, partiendo de cierta hipóte sis de carácter general, deducir matemáticamente como deben ser las distribuciones de frecuencias de ciertos sucesos. Estas dis tribuciones se llaman teóricas.

Hay tres tipos de distribuciones teóricas clásicas: la binomial, la normal y la de Poisson. Existen otras distribu ciones teóricas, pero dependen en cierto modo de las propieda- des de las tres primeras y sobre todo de la normal.

Distribución binomial

Supongamos que tenemos una moneda para la cual la pro babilidad de obtener cara es p y la probabilidad de obtener cruz es $q = 1 - p$.

Preguntamos ¿ Cuál es la probabilidad de obtener m caras al lanzar dicha moneda al aire n veces?

1) Consideremos el caso de que tiramos la moneda dos veces. Los cuatro casos posibles son:

++ ; + x ; x + ; xx

Suponemos que los sucesos son independientes, entonces en virtud del teorema de la probabilidad compuesta, resulta que las probabilidades de estos cuatro acontecimientos son:

pp ; pq ; qp ; qq

Los acontecimientos indicados en 2º y 3er término , presentan cada uno, una cara y una cruz y por el teorema de la probabilidad total, es la probabilidad de obtener una cara, igual a la suma de las dos probabilidades o sea

$$pq + qp = 2pq$$

Luego, las probabilidades de obtener 0, 1; 2 caras al tirar dos veces, una moneda son:

$$q^2 \ , \quad 2pq, \quad p^2$$

respectivamente.

Si tirásemos tres veces la moneda, los acontecimientos posibles son:

+ + + ;	+ + x ;	+ x x ;	x x x
	+ x +	x + x	
	x + +	x x +	

y sus probabilidades

$$p^3; \quad 3p^2q; \quad 3pq^2; \quad q^3$$

Generalizando y de acuerdo a la pregunta formulada , deseamos calcular la probabilidad de obtener en n tiradas de la moneda m veces cara y por lo tanto $n - m$ veces cruz. Si se presentara primero m veces el suceso favorable o sea cara y luego $n - m$ veces el suceso contrario o sea cruz se tendría

$$p^m \ q^{n-m}$$

No interesa a nuestro problema el orden de presentación de cara o cruz sino el número de veces total que cada una de ellas debe presentarse y se sabe que el total de órdenes que pueden formarse con n elementos tomándolos de m en m es : C_n^m . Aplicando entonces el teorema de la probabilidad total se tiene que la probabilidad buscada es igual a:

$$P_m = C_n^m \ p^m \ q^{n-m} \qquad (1)$$

$$= \frac{n!}{m!\,(n-m)!} \ p^m \ q^{n-m}$$

Por ejemplo deseamos conocer la probabilidad de que arrojando 100 veces una moneda se obtenga 40 veces cara.

En este caso la probabilidad elemental de que se presente cara p es igual a la presentación del suceso contrario, cruz q es decir:

$$p = q = \frac{1}{2}$$

entonces la probabilidad pedida según la expresión (1) será:

$$P_{40} = C_{100}^{40} \ p^{40} \ q^{100-40}$$

$$= \frac{100!}{40!\ 60!} \left(\frac{1}{2}\right)^{40} \left(\frac{1}{2}\right)^{60}$$

Observamos en la expresión (1) que P_m es el término general del desarrollo del binomio de Newton $(q+p)^n$, se tiene en efecto que

$$(q+p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{2!}q^{n-2}p^2 + \ldots + \frac{n(n-1)\ldots n-(n-1)}{n!}p^n$$

$$= C_n^0 q^n + C_n^1 q^{n-1}p + C_n^2 q^{n-2}p^2 + \ldots + C_n^m q^{n-m}p^m + \ldots + C_n^n p^n \quad (2)$$

Resulta en virtud de (1) que podemos escribir:

$$(q+p)^n = P_0 + P_1 + P_2 + \ldots + P_m + \ldots + P_n = \sum_{m=0}^{n} P_m \quad (3)$$

siendo $q + p = 1$; es $(q+p)^n = 1$; de donde:

$$1 = \sum_{m=0}^{n} P_m$$

El primer término del segundo miembro de (2) da la probabilidad de que en n tiradas aparezca cara cero vez; el segundo, de que aparezca una vez cara; el tercero de que aparezca dos veces cara y así sucesivamente. Estos resultados pueden presentarse en una tabla de distribución de probabilidad.

Tabla de probabilidad para la distribución binomial

m	$P_m = f(x)$
0	q^n
1	$C_n^1 pq^{n-1}$
2	$C_n^2 p^2 q^{n-2}$
.	.
.	.
.	.
m	$C_n^m p^m q^{n-m}$
.	.
.	.
.	.
n	p^n

Si en vez de referirnos específicamente a una moneda, generalizamos y hablamos del acontecimiento "E" y del acontecimiento "no E" podemos decir

Si un suceso "A" tiene la probabilidad p de presentarse en cada una de n pruebas independientes, la probabilidad para que en dichas n pruebas tal suceso se presente exactamente m veces es igual a:

$$P_m = C_n^m p^m q^{n-m} \qquad (4)$$

donde, $q = 1 - p$

La distribución de probabilidad (4) se llama distribución de probabilidad binomial o distribución binomial. Para el caso particular de que

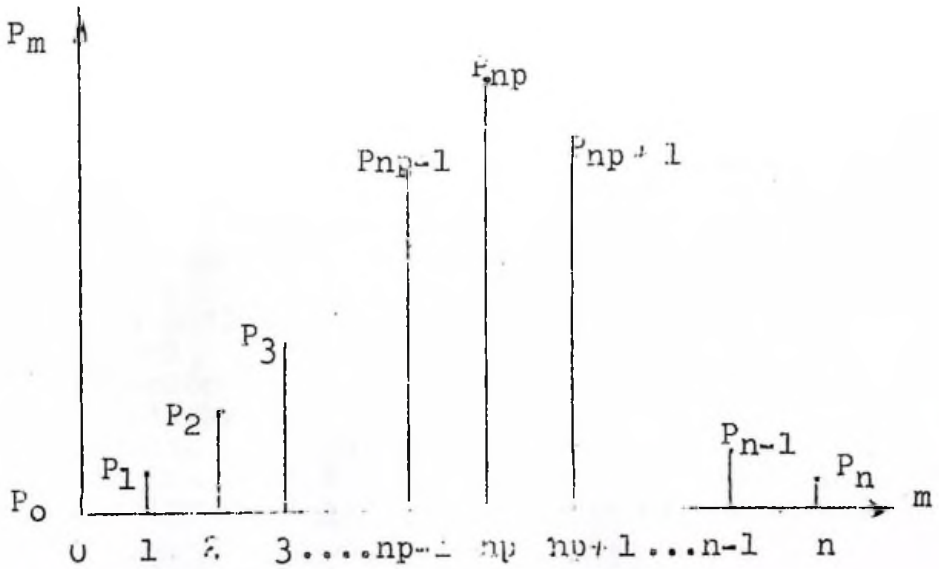
$$p = q = \frac{1}{2}$$

como en el ejemplo de la moneda, (4) resulta:

$$P_m = C_n^m p^n = C_n^m \left(\frac{1}{2}\right)^n$$

Representando gráficamente los valores de la tabla anterior para el caso de $p=q=\frac{1}{2}$ se tendría un gráfico binomial simétrico con respecto a la ordenada correspondiente a la probabilidad $m=n/2$, porque las alturas de las ordenadas estarían dadas por los valores de los números combinatorios correspondientes a C_n^m cuando m toma los $(n+1)$ valores entre 0 y n, multiplicados siempre por p^n; y ya sabemos que para valores de m equidistantes de los extremos, los números combinatorios son iguales, en virtud de la conocida propiedad:

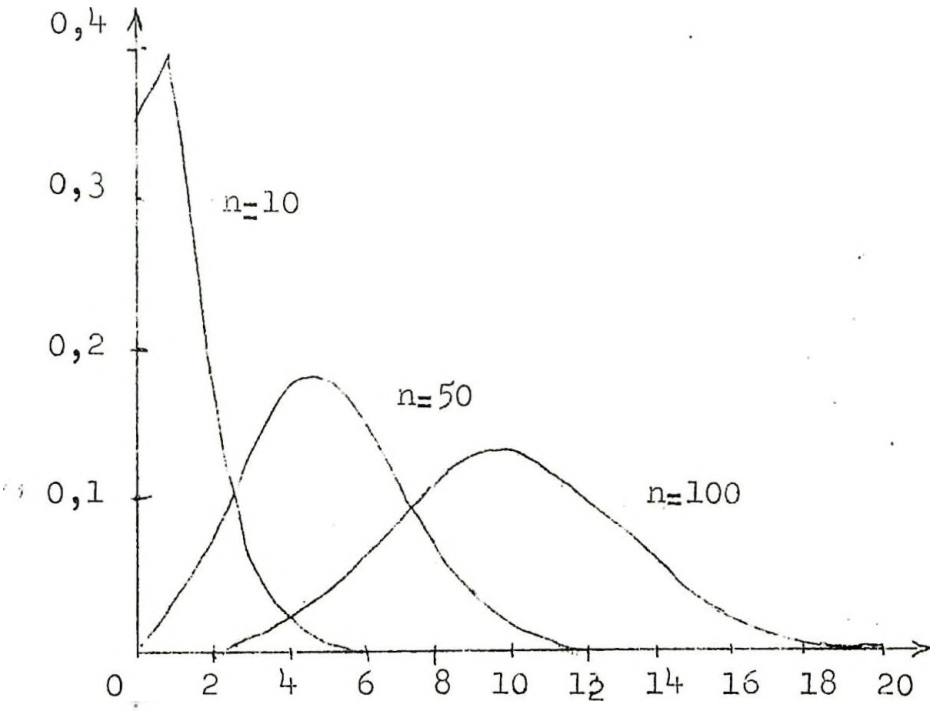
$$C_n^m = C_n^{n-m}$$



La simetría que presenta el gráfico anterior desaparece cuando las probabilidades p y q toman valores distintos y también según los valores que toma n . Si $p > q$ la asimetría es positiva y si $p < q$ la asimetría es negativa.

En el caso de ser $p=q$ al crecer n sube el valor de la media aritmética y aumenta la dispersión; pero si $p \neq q$, al crecer n no sólo sube el valor de la media aritmética y aumenta la dispersión sino que disminuye la asimetría.

Así para la potencia del binomio $(0,9 + 0,1)^n$, dando a n los valores 10, 50 y 100 se tendrán los siguientes gráficos:



La distribución binomial tiene su origen en el esquema de Bernouilli quien consideró una urna de composición conocida en la que se efectúan extracciones sucesivas con reposición de la bolilla extraída cada vez, de modo que tanto la probabilidad favorable al acontecimiento esperado como la probabilidad contraria fueran constantes de prueba a prueba.

La utilización práctica del esquema de Bernoulli podemos concretarla así: si observamos la representación gráfica de una distribución de frecuencias de un fénomeno cualquiera que , pretendemos describir equiparándolo a aquel esquema, pero, para el que desconocemos sus probabilidades elementales, podemos llegar a tener una idea cualitativa de la mayor o menor diferencia entre los valores de p y q. En efecto la mayor asimetría del grá

fico indicará una mayor diferencia entre los valores de ambas probabilidades y cuanto más simétrico sea el gráfico más próximo a ser iguales a ½ estarán p y q.

En virtud de las observaciones efectuadas respecto al comportamiento de la distribución binomial cuando n aumenta, veremos que la ley conocida como distribución de probabilidad normal o de Gauss da una buena aproximación para las probabilidades de la distribución binomial.

Est. - A

23a. Conferencia

Probabilidad máxima

Veremos ahora para cuál de los $n+1$ valores posibles de m, la probabilidad P_m es máxima. Si fuese m el valor de la variable para el que P_m es máxima, podemos escribir estas dos condiciones inmediatas:

$$P_m > P_{m-1} \qquad y \qquad P_m > P_{m+1}$$

$$\frac{P_m}{P_{m-1}} > 1 \quad (1) \; ; \qquad \frac{P_{m+1}}{P_m} < 1 \quad (2)$$

Reemplazando en (1) a P_m y P_{m-1}, por sus valores, se tiene

$$\frac{C_n^m p^m q^{n-m}}{C_n^{m-1} p^{m-1} q^{n-m+1}} =$$

$$\frac{\frac{n(n-1)(n-2)\ldots(n-m+1)}{m!}\, p^m q^{n-m}}{\frac{n(n-1)(n-2)\ldots(n-m+2)}{(m-1)!}\, p^{m-1} q^{n-m+1}} > 1$$

Simplificando se tiene:

$$\frac{n-m+1}{m} \cdot \frac{p}{q} > 1$$

de donde:

$$(n-m+1)\,p > mq$$
$$np + p > m(p+q)$$
$$np + p > m \qquad \therefore \qquad m < np + p$$

En (2)

$$\frac{C_n^{m+1} p^{m+1} q^{n-m-1}}{C_n^m p^m q^{n-m}} = \frac{\frac{n(n-1)(n-2)\ldots(n-m)}{(m+1)!}\, p^{m+1} q^{n-m-1}}{\frac{n(n-1)\ldots(n-m+1)}{m!}\, p^m q^{n-m}} < 1$$

Simplificando:

$$\frac{(n-m)}{(m+1)}\,\frac{p}{q} < 1$$

$$np - mp < mq + q$$

$$np - q < m(p+q) \quad \therefore \quad \boxed{m > np - q}$$

Resulta así que:

$$np + p > m > np - q$$

pero siendo

$$np + p - (np - q) = np + p - np + q$$
$$= p + q = 1$$

el valor buscado m dentro de un intervalo de amplitud 1

Siendo nuestra variable aleatoria discontinua, cuando np sea entero, este será el valor de m y si fuera fraccionario, tomaremos el mayor entero contenido en np.

El valor (o uno cualquiera de los dos valores consecutivos) de m, al que corresponde la probabilidad máxima, o sea el término máximo (o los dos términos máximos consecutivos iguales) en el desarrollo del binomio, es el valor más probable o normal. Podemos entonces concluir diciendo que en una serie de n pruebas, el caso más probable es que se presente np veces el suceso favorable y, n-np = nq veces el suceso desfavorable y que si np no es entero, puede sustituirse por uno cualquiera de los dos enteros próximos, siempre que n sea muy grande.

Ejemplos:

1) Se tira una moneda 10.000 veces consecutivas ¿Cuál es el número de veces que tiene mayor probabilidad de presentarse cara?

$p = \frac{1}{2}$

$q = \frac{1}{2}$

$n = 10.000$

$$\frac{10.000}{2} + \frac{1}{2} > m > \frac{10.000}{2} - \frac{1}{2}$$

$$m = 5.000$$

2) Se tira un dado 500 veces consecutivas ¿Cuál es el número de veces que tiene mayor probabilidad de presentarse el 6?

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

$q = \frac{5}{6}$

$n = 500$

$$\frac{500}{6} + \frac{1}{6} > m > \frac{500}{6} - \frac{5}{6}$$

$$\frac{501}{6} > m > \frac{495}{6}$$

$$83,5 > m > 82,5$$

$$m = 83$$

3) Un bolillero contiene dos bolillas blancas, tres verdes y una roja. Se hacen 100 extracciones reponiendo cada vez la bolilla extraída. ¿Cuál es el número de veces que tiene mayor probabilidad de aparecer bolilla blanca?

$p = \frac{2}{6} = \frac{1}{3}$

$q = \frac{4}{6} = \frac{2}{3}$

$n = 100$

$$\begin{cases} \frac{100}{3} + \frac{1}{3} > m > \frac{100}{3} - \frac{2}{3} \\ \frac{101}{3} > m > \frac{98}{3} \\ 33\frac{2}{3} > m > 32\frac{2}{3} \end{cases}$$

$m = 33$

4) Se tira una moneda cinco veces consecutivas. Determinar cual es el número de veces que tiene mayor probabilidad de presentarse cara?

$p = \frac{1}{2}$

$q = \frac{1}{2}$

$n = 5$

$$\begin{cases} \frac{5}{2} + \frac{1}{2} > m > \frac{5}{2} - \frac{1}{2} \\ \frac{6}{2} > m > \frac{4}{2} \\ 3 > m > 2 \end{cases}$$

$\therefore \begin{cases} m = 3 \\ m = 2 \end{cases}$

5) Se tira un dado cinco veces consecutivas ¿Cuál es el número de veces que tiene mayor probabilidad de presentarse el mismo?

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

$q = \frac{5}{6}$

$n = 5$

$$\begin{cases} \frac{5}{6} + \frac{1}{6} > m > \frac{5}{6} - \frac{5}{6} \\ \frac{6}{6} > m > \frac{0}{6} \\ 1 > m > 0 \end{cases}$$

$\therefore \begin{cases} m = 1 \\ m = 0 \end{cases}$

Est. - A

24a. Conferencia

La media aritmética, la varianza y el desvío standard de la distribución binomial

Media aritmética

La media aritmética como en el caso de toda distribución de probabilidad se simboliza por μ

De acuerdo a la definición conocida de media aritmética, tendremos para la distribución binomial que:

$$\mu = \sum_{x=0}^{n} x\, P_x \qquad (1)$$

reemplazando el valor de P_x por la expresión antes hallada, resulta:

$$\mu = \sum_{x=0}^{n} x \left[C_n^x \, p^x \, q^{n-x} \right] \qquad (2)$$

Mediante un sencillo artificio hallaremos el valor de μ
Vamos a considerar el desarrollo de:

$$(q + tp)^n = C_n^0 q^n (tp)^0 + C_n^1 q^{n-1}(tp)^1 + \ldots + C_n^x q^{n-x}(tp)^x + \ldots + C_n^n q^0 (tp)^n$$

donde t es una variable auxiliar introducida para facilitar la demostración.

Derivando ambos miembros de la igualdad respecto de t se tiene:

$$n\,(q+tp)^{n-1} = 1\left[C_n^1 p^1 q^{n-1}\right] t^0 + 2\left[C_n^2 p^2 q^{n-2}\right] t + \ldots + x\left[C_n^x p^x q^{n-x}\right] t^{x-1} + \ldots + n\left[C_n^n p^n q^0\right] t^{n-1} \qquad (3)$$

para $t = 1$ y recordando que $(q+p) = 1$ resulta:

$$np = 1\left[C_n^1 p^1 q^{n-1}\right] + 2\left[C_n^2 p^2 q^{n-2}\right] + \ldots + x\left[C_n^x p^x q^{n-x}\right] + \ldots + n\left[C_n^n p^n q^0\right]$$

pero el segundo miembro no es otra expresión que el desarrollo de la fórmula (2) por lo tanto

$$\boxed{\mu = np} \qquad (4)$$

Es decir que en el caso de la distribución binomial la media aritmética es igual al producto de n (número de pruebas) por p (probabilidad favorable al acontecimiento esperado)

Varianza:

La fórmula para la varianza en el caso de una distribución de probabilidad era:

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 \, p_i - \mu^2$$

y que para la distribución binomial toma la forma:

$$\sigma^2 = \sum_{x=0}^{n} x^2 \, P_x - \mu^2 = \sum_{x=0}^{n} x^2 \left\{ C_n^x \, p^x \, q^{n-x} \right\} - \mu^2 \qquad (5)$$

Con el propósito de hallar una expresión más sencilla multiplicamos por t los dos miembros de la expresión (3)

$$npt\,(q+tp)^{n-1} = \left\{ C_n^1 \, p^1 \, q^{n-1} \right\} t + 2 \left\{ C_n^2 \, p^2 q^{n-2} \right\} t^2 + \ldots + x \left\{ C_n^x \, p^x \, q^{n-x} \right\} t^x + \ldots + n \left\{ C_n^n \, p^n q^0 \right\} t^n$$

derivamos ambos miembros respecto de t y tenemos:

$$np\,(q+tp)^{n-1} + n(n-1)\,p^2 t (q+tp)^{n-2} = C_n^1 \, p^1 q^{n-1} + 2^2 \left\{ C_n^2 \, p^2 q^{n-2} \right\} t + \ldots + x^2 \left\{ C_n^x \, p^x q^{n-x} \right\} t^{x-1} + \ldots + n^2 \left\{ C_n^n \, p^n q^0 \right\} t^{n-1}$$

haciendo t=1 y siendo (q + p) = 1, queda:

$$np + n(n-1)p^2 = 1^2 C_n^1 \, p^1 q^{n-1} + 2^2 C_n^2 p^2 q^{n-2} + \ldots + x^2 C_n^x p^x q^{n-x} + \ldots + n^2 C_n^n p^n q^0$$

la expresión del segundo miembro no es otra cosa que el desarrollo de la sumatoria indicada en (5), de donde reemplazando la nos queda:

$$\sigma^2 = np + n\,(n-1)\,p^2 - \mu^2$$

Sustituimos μ por su expresión indicada en (4)

$$\sigma^2 = np + n\,(n-1)\,p^2 - (np)^2$$
$$= np + n^2 p^2 - np^2 - n^2 p^2$$
$$= np\,(1 - p)$$

$$\boxed{\sigma^2 = npq} \qquad (6)$$

La varianza en la distribución binomial es igual al producto del número de pruebas por las probabilidades favorable y contraria .

Desvío Standard:

Extrayendo la raíz cuadrada de la varianza tenemos determinado el desvío o desviación típica o standard. Operando en

(6) se tiene:

$$\sigma = \sqrt{npq} \qquad (7)$$

o sea la expresión de σ en función de n, p y q.

Ajuste de una distribución binomial a una distribución de frecuencias.

Dijimos que observando el comportamiento de las frecuencias relativas de una observación podemos decidir si es posible considerarlas vinculadas a una distribución binomial teórica.

En muchos problemas no se puede determinar previamente el valor de p. Entonces estimaremos a p según el procedimiento que indicaremos a continuación.

Supongamos el siguiente problema planteado por Wilks: ¿Cuál es la probabilidad p de que al tirar una chinche ordinaria caiga con la punta hacia arriba? Por el simple examen de la chinche no se puede predecir, pero puede estimarse arrojándola un gran número de veces, o tirando al suelo varias chinches iguales, un gran número de veces. Por ejemplo, en 100 tiradas de cinco chinches, obtuvo Wilks el resultado que se consigna en las dos primeras columnas del cuadro siguiente:

x_i	y_i	$\frac{y_i}{\sum y_i} = f_i$	$x_i f_i$	Distribución binomial	Frec. teórica	Frec. acum. obs.	Frec. acum. teór.
0	2	0,02	0	$C_0^5(0,568)^0\,(0,432)^5 = 0,015$	1,5	2	1,5
1	14	0,14	0,14	$C_1^5(0,568)^1(0,432)^4 = 0,099$	9,9	16	11,4
2	20	0,20	0,40	$C_2^5(0,568)^2(0,432)^3 = 0,260$	26.-	36	37,4
3	34	0,34	1,02	$C_3^5(0,568)^3(0,432)^2 = 0,342$	34,2	70	71,6
4	22	0,22	0,88	$C_4^5(0,568)^4(0,432)^1 = 0,225$	22,5	92	94,1
5	8	0,08	0,40	$C_5^5(0,568)^5(0,432)^0 = 0,059$	5,9	100	100,0
	100	1	2,84	1	100		

¿Cómo estimar la probabilidad p de que la chinche caiga con la punta hacia arriba?

Teóricamente podríamos determinar la distribución de

probabilidad para x, usando la probabilidad desconocida p, con la expresión:

$$P_x = C_5^x \, p^x \, q^{5-x} \qquad (8)$$

para x = 0, 1, 2, 3, 4, 5

Estimando que la observación realizada corresponde a una distribución binomial, para determinar el valor de p, igualaremos esta probabilidad teórica a la media aritmética de la distribución de frecuencias relativas obtenida de la observación.

Siendo n = 5 la media de la distribución teórica, resulta:

$$\mu = np = 5p \qquad (9)$$

y la media de la distribución observada:

$$\bar{X} = \sum_{i=0}^{5} x_i \, f_i = 2,84 \qquad (10)$$

igualando (9) y (10) resulta:

$$5p = 2,84 \qquad (11)$$

obteniendo que

$$p = \frac{2,84}{5} = 0,568 \qquad (12)$$

Observamos que si tirásemos la chinche 500 veces resultará por (11) que 284 veces caería con la punta hacia arriba y en virtud de la definición de probabilidad será:

$$p = \frac{284}{500} = 0,568$$

Este es el resultado estimado de p, basado en un experimento. Si se realizaran otros experimentos se tendrían resultados probablemente un poco diferentes, pero próximos a 0,56 o a 0,57.

Reeplazando el valor de p dado por (12) en la fórmula (8) resulta:

$$P_x = C_5^x \, (0,568)^x \, (1-0,568)^{5-x}$$

$$= C_5^x \, (0,568)^x \, (0,432)^{5-x}$$

que puede considerarse una aproximación de la distribución de frecuencias relativas observadas del número de chinches que caen con la punta hacia arriba. Los resultados que aplicando esta

fórmula se consignan en el cuadro anterior en la quinta columna se multiplicaron por 100 a fin de obtener las frecuencias teóricas, o probabilidades binomiales ajustadas, de la columna 6.

Puede observarse la concordancia entre frecuencias relativas y probabilidades binomiales ajustadas y también entre frecuencias observadas y frecuencias teóricas.

La misma indicación puede hacerse respecto de las frecuencias acumuladas de valores observados y teóricos.

Est. A

25a. Conferencia

DISTRIBUCION DE POISSON

La distribución de Poisson también llamada ley de Poisson es una ley límite del cálculo de probabilidades. En efecto al aplicar la distribución binomial cuya expresión matemática está representada por la fórmula:

$$P_m = C_n^m \, p^m \, q^{n-m} \qquad (1)$$

se presentan casos en que la probabilidad favorable al suceso p es muy pequeña, menor que 1/10 y n grande, mayor que 50 de modo que el valor medio np oscila entre 0 y 10.

En tales casos, se utiliza la distribución de Poisson, es decir en el estudio de aquellos fenómenos que se revelan con frecuencias muy pequeñas, llamados "sucesos raros" como ser: nacimiento de mellizos, de albinos, suicidios de menores, etc.

La expresión (1) podemos escribirla también:

$$P_m = \frac{n(n-1)(n-2)\ldots(n-m+1)}{m!} \cdot p^m (1-p)^{n-m} \qquad (2)$$

El suceso se considera raro cuando se verifica que:

$$\lim_{n \to \infty} (np) = k \qquad (3)$$

mientras que en el caso de no ser la probabilidad p muy pequeña, el límite de np cuando $n \to \infty$ será también infinito.

Estadísticamente se dice que un suceso es raro cuando para el mismo corresponden frecuencias pequeñísimas y el total de casos observados es sensiblemente constante, a pesar de aumentar el número de experiencias.

Por tanto si el suceso es raro, m número de veces que se verifica, debe ser un número pequeño. En un conjunto grande de experiencias, podremos despreciar el valor de m frente al de n; es decir, aceptar la siguiente igualdad aproximada:

$$n - m \cong n \qquad (4)$$

Si, dentro de nuestra clase de experiencias suficientemente grandes, tomamos

$$np = K \qquad (5)$$

podemos hallar un valor estimativo de p, que será:

$$p = \frac{K}{n} \qquad (6)$$

de aquí pues la importancia del parámetro K

Reemplazando en (2) los valores expresados en (4) y (6) resulta:

$$P_m = \frac{n(n-1)\ldots(n-m+1)}{m!} \left(\frac{K}{n}\right)^m \left(1-\frac{K}{n}\right)^n \qquad (7)$$

que podemos escribir

$$P_m = \frac{K^m}{m!} \cdot \frac{n(n-1)\ldots(n-m+1)}{n^m} \left(1-\frac{K}{n}\right)^n$$

$$P_m = \frac{K^m}{m!} \cdot 1 \cdot \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)\left(1-\frac{K}{n}\right)^n$$

haciendo tender n a infinito:

$$\lim_{n\to\infty} P_m = \frac{K^m}{m!} \lim_{n\to\infty} \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)\left(1-\frac{K}{n}\right)^n \qquad (8)$$

Cuando $n\to\infty$ los primeros m-1 factores de la expresión (8) tienden a la unidad por lo cual queda:

$$\lim_{n\to\infty} P_m = \frac{K^m}{m!} \lim_{n\to\infty} \left(1-\frac{K}{n}\right)^n$$

Pero sabemos que (*)

$$\lim_{n\to\infty} \left(1-\frac{K}{n}\right)^n = e^{-K}$$

(*) En efecto, haciendo:

$$\frac{K}{n} = -\frac{1}{r}$$

y por tanto $n = -K.r$ podemos escribir que:

$$\left(1-\frac{K}{n}\right)^n = \left(1+\frac{1}{r}\right)^{-rK} = \left[\left(1+\frac{1}{r}\right)^r\right]^{-K} \quad \text{pasando a límite}$$

$$\lim_{n\to\infty} \left(1-\frac{K}{n}\right)^n = \lim_{r\to\infty} \left[\left(1+\frac{1}{r}\right)^r\right]^{-K} \quad \text{pero:}$$

$$\lim_{r\to\infty} \left(1+\frac{1}{r}\right) = e \quad \text{luego:}$$

$$\lim_{r\to\infty} \left[\left(1+\frac{1}{r}\right)^r\right]^{-K} = e^{-K}$$

luego $\lim_{n\to\infty} P_m = \frac{K^m}{m!} \cdot e^{-K}$ (8)

que es la llamada fórmula o función de Poisson y depende de una sola constante, que es K

La función de Poisson depende del parámetro R y de los valores que tome la variable aleatoria m, es pues función de K y m, representando a dicha función por φ resulta:

$$\lim_{n\to\infty} P_m = \varphi(K,m) = \frac{K^m}{m!} e^{-K} \qquad (9)$$

Representación gráfica de la función de Poisson:

Para verificar las variaciones que sufre esta función, fijaremos en tres casos distintos para K los valores 1, 2 y 3 y daremos a m valores que varíen de 0 a 5

K = 1

m	$\varphi(1,m)$
0	$\varphi(1,0) = \frac{1}{e} \cdot \frac{1^0}{0!} = 0,3679$
1	$\varphi(1,1) = \frac{1}{e} \cdot \frac{1}{1} = 0,3679$
2	$\varphi(1,2) = \frac{1}{e} \cdot \frac{1}{2!} = 0,1839$
3	$\varphi(1,3) = \frac{1}{e} \cdot \frac{1}{3!} = 0,0613$
4	$\varphi(1,4) = \frac{1}{e} \cdot \frac{1}{4!} = 0,0153$
5	$\varphi(1,5) = \frac{1}{e} \cdot \frac{1}{5!} = 0,0031$

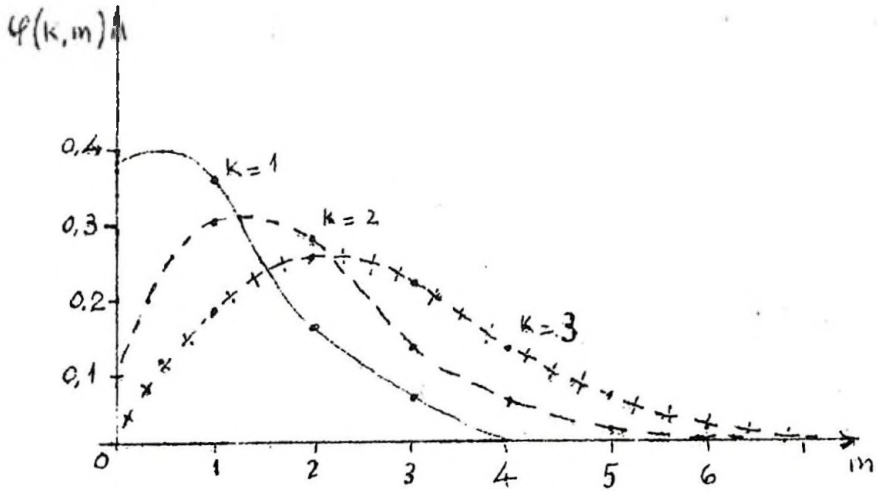
K = 2

m	$\varphi(2,m)$
0	$\varphi(2,0) = \frac{1}{e^2} \cdot \frac{2^0}{0!} = 0,1354$
1	$\varphi(2,1) = \frac{1}{e^2} \cdot \frac{2}{1} = 0,2708$
2	$\varphi(2,2) = \frac{1}{e^2} \cdot \frac{2^2}{2!} = 0,2707$
3	$\varphi(2,3) = \frac{1}{e^2} \cdot \frac{2^3}{3!} = 0,1805$
4	$\varphi(2,4) = \frac{1}{e^2} \cdot \frac{2^4}{4!} = 0,0902$
5	$\varphi(2,5) = \frac{1}{e^2} \cdot \frac{2^5}{5!} = 0,0361$

K = 3

m	$\varphi(3,m)$
0	$\varphi(3,0) = \frac{1}{e^3} \cdot \frac{3^0}{0!} = 0{,}0498$
1	$\varphi(3,1) = \frac{1}{e^3} \cdot \frac{3}{1} = 0{,}1494$
2	$\varphi(3,2) = \frac{1}{e^3} \cdot \frac{3^2}{2!} = 0{,}2241$
3	$\varphi(3,3) = \frac{1}{e^3} \cdot \frac{3^3}{3!} = 0{,}2241$
4	$\varphi(3,4) = \frac{1}{e^3} \cdot \frac{3^4}{4!} = 0\ 1\emptyset 81$
5	$\varphi(3,5) = \frac{1}{e^3} \cdot \frac{3^5}{5!} = 0{,}1008$

Representando gráficamente estos valores, se tiene:



Est. A

26a. Conferencia

VALOR MAXIMO DE LA FUNCION DE POISSON

Al representar gráficamente la función de Poisson vimos que las ordenadas crecen hasta un cierto valor, según sea K, para disminuir luego constantemente. Este hecho se debe a que en esta función el denominador crece más rápidamente que el numerador.

Veamos ahora para qué valor de m, suponiendo k determinado, alcanza la función $\varphi(k,m)$ su valor máximo.

Si suponemos que es m el valor de la variable aleatoria al que le corresponde la ordenada máxima, debe verificarse:

$$\frac{\varphi(k,m)}{\varphi(k,m-1)} > 1$$

o lo que es igual:

$$\frac{\frac{k^m}{m!}e^{-k}}{\frac{k^{m-1}}{(m-1)!}e^{k}} > 1$$

simplificando:

$$\frac{k}{m} > 1$$

de donde resulta que:

$$\boxed{k > m} \qquad (10)$$

Si m es el valor al que corresponde la ordenada máxima, también se verificará que:

$$\frac{\varphi(k,m)}{\varphi(k,m+1)} > 1$$

o lo que es igual:

$$\frac{\frac{k^m}{m!}e^{-k}}{\frac{k^{m+1}}{(m+1)!}e^{-k}} > 1$$

simplificando:

$$\frac{m+1}{k} > 1$$

de donde resulta que:

$$m + 1 > k$$

$$\boxed{m > k - 1} \qquad (11)$$

Es decir, que en virtud de (10) y (11) el valor de m que hace máxima a $\varphi(k,m)$ cumple las condiciones:

$$\boxed{k-1 < m < k} \qquad (12)$$

condiciones que pueden verificarse en los ejemplos que se re presentaron gráficamente, cuando k toma valores enteros.

Cuando k sea un número fraccionario, por ejemplo k = 3,75 el valor máximo de la función $\varphi(k,m)$ se tendrá para el valor de m igual al mayor entero comprendido en k, es decir 3; si fuera k = 0,75 el máximo se verificará para m=0

Media aritmética, varianza y desvío típico de la distribución de Poisson :

La distribución de Poisson tiene la siguiente propiedad simple:

$$\mu = \sigma^2 = k$$

Constataremos estas relaciones, considerando en primer término la determinación del valor de la media aritmética.

Por definición de media aritmética en el caso de la distribución de Poisson tendríamos:

$$\mu = \sum_{m=0}^{\infty} m\ \varphi(k,m) = \sum_{m=0}^{\infty} m \frac{k^m}{m!} e^{-k}$$

Sea el desarrollo de:

$$e^{tk} = 1 + tk + \frac{t^2k^2}{2!} + \frac{t^3k^3}{3!} + \ldots.$$

multiplicando ambos miembros por e^{-k} resulta:

$$e^{tk} \cdot e^{-k} = e^{-k} + tke^{-k} + \frac{t^2k^2}{2!} e^{-k} + \frac{t^3k^3}{3!} e^{-k} + \ldots.$$

derivando respecto a t se tiene:

$$ke^{tk} \cdot e^{-k} = ke^{-k} + \frac{2tk^2}{2!} e^{-k} + \frac{3t^2k^3}{3!} e^{-k} + \ldots. \qquad (13)$$

haciendo t = 1

$$k = ke^{-k} + 2\frac{k^2}{2!} e^{-k} + 3\frac{k^3}{3!} e^{-k} + \ldots.$$

$$k = \sum_{m=0}^{\infty} m \frac{k^m}{m!} e^{-k}$$

Hemos demostrado así que la media aritmética correspondiente a la distribución de Poisson es igual a k, o sea:

$$\boxed{\mu = k} \qquad (14)$$

Est. A

27a. Conferencia

V A R I A N Z A

La varianza por definición será:

$$\sigma^2 = \sum_{m=0}^{\infty} (m - \mu)^2 \, f(k,m)$$

$$\sigma^2 = \sum_{m=0}^{\infty} (m - k)^2 \frac{k^m}{m!} e^{-k}$$

$$\sigma^2 = \sum_{m=0}^{\infty} (m^2 - 2mk + k^2) \frac{k^m}{m!} e^{-k}$$

$$= \sum_{m=0}^{\infty} m^2 \frac{k^m}{m!} e^{-k} - 2k \sum_{m=0}^{\infty} m \frac{k^m}{m!} e^{-k} + k^2 \sum_{m=0}^{\infty} \frac{k^m}{m!} e^{-k}$$

En virtud de (14) resulta:

$$\sigma^2 = \sum_{m=0}^{\infty} m^2 \frac{k^m}{m!} e^{-k} - 2k.k + k^2 \qquad (15)$$

Nos falta determinar el primer término del segundo miembro de la fórmula (15)

Multiplicamos ambos miembros de (13) por t:

$$tke^{tk}e^{-k} = tke^{-k} + \frac{2t^2k^2}{2!} e^{-k} + \frac{3t^3k^3}{3!} e^{-k} + \ldots$$

derivando respecto a t:

$$ke^{tk}e^{-k} + tk^2e^{tk}e^{-k} = ke^{-k} + \frac{2^2tk^2}{2!} e^{-k} + \frac{3^2t^2k^3}{3!} e^{-k} + \ldots$$

$$(k+k^2)\, e^{tk}e^{-k} = ke^{-k} + \frac{2^2tk^2}{2!} e^{-k} + \frac{3^2t^2k^3}{3!} e^{-k} + \ldots$$

haciendo $t = 1$

$$k + k^2 = ke^{-k} + 2^2 \frac{k^2}{2!} e^{-k} + 3^2 \frac{k^3}{3!} e^{-k} + \ldots$$

$$\sum_{m=0}^{\infty} \frac{k^m}{m!} e^{-k} = e^{-k} \sum_{m=0}^{\infty} \frac{k^m}{m!} = e^{-k} \left[1 + \frac{k^2}{1!} + \frac{k^2}{2!} + \frac{k^3}{3!} + \ldots \right] = e^{-k}e^{k} = 1$$

$$k + k^2 = \sum_{m=0}^{\infty} m^2 \frac{k^m}{m!} e^{-k} \qquad (16)$$

Reemplazando el valor de (16) en (15) se tiene:

$$\sigma^2 = k + k^2 - k^2$$

$$\boxed{\sigma^2 = k} \qquad (17)$$

es decir que la varianza en la distribución de Poisson es igual a k

Por tanto el desvío típico se obtendrá extrayendo la raíz cuadrada en (17)

$$\sigma = \sqrt{k}$$

Tomando las propiedades expresadas en las fórmulas (14) y (17) como característica de la función de Poisson, será facil utilizarlas frente a un caso práctico experimental.

Si suponemos que los datos de una observación pueden pertenecer a un suceso raro y decidimos aplicar la distribución de Poisson para representarla, debemos antes de iniciar el trabajo de cálculo, tener la certidumbre de que no realizamos una tarea inútil. Para trabajar sobre seguro calcularemos la media aritmética y la varianza. Si estos valores son iguales o muy próximos entre sí, podemos concluír que el fenómeno será descripto por la ley de Poisson.

EJEMPLOS:

1) Se lanzan cinco monedas 64 veces. ¿Cuáles son las probabilidades de obtener cinco caras 0, 1, 2;... 64 veces?

+ + + + +	$\binom{5}{4} = 5$	$n = 64$
+ x + + +		
+ x x + +	$\binom{5}{3} = 10$	$p = \frac{1}{32}$
+ x x x +	$\binom{5}{2} = 10$	
+ x x x x	$\binom{5}{1} = 5$	$q = \frac{31}{32}$
x x x x x	$= 1$	
	$\overline{32}$	

Valores exactos

$$P_{64} = C_{64}^{m} \left(\frac{1}{32}\right)^{m} \left(\frac{31}{32}\right)^{64-m}$$

para m = o, 1, 2.....64

Valores aproximados de estas probabilidades dados por la distribución de Poisson

$\mu = k = np$ $\qquad np = 64 \frac{1}{32} = 2$

$$\varphi(k,m) = \frac{2^m e^{-2}}{m!} \qquad \text{para } m = 0, 1, 2, \ldots \infty$$

m	P_m	$\varphi(m,k)$
0	0,131	0,135
1	0,271	0,271
2	0,275	0,271
3	0,183	0,180
4	0,090	0,090
5	0,035	0,036
6	0,011	0,012
7	0,003	0,004
8	0,001	0,001
9	0,000	0,000

2) Sea la distribución siguiente obtenida de una observación a la que se pregunta si puede ajustarse una distribución de Poisson. En caso afirmativo realizar los cálculos.

Suicidos de varone menores de 15 años
Rosario 1907-1928

Años	Nº de Suicidios	Años	Nº de Suicidios	Años	Nº de Suicidios
1907	0	1916	2	1926	3
1908	0	1917	0	1927	1
1909	0	1918	1	1928	2
1910	1	1919	1		
1911	1	1920	2		
1912	0	1921	2	Total	20
1913	1	1922	0		
1914	0	1923	0		
1915	2	1924	1		
		1925	0		

n = 22 (Nº de años observados)

m	y_i	m y_i	$(m-\overline{X})$	$(m-\overline{X})^2$	$(m-\overline{X})^2 y_i$
0	9	0	-0,9091	0,8265	7,4385
1	7	7	+0,0909	0,0083	0,0581
2	5	10	+1,0909	1,1901	5,9505
3	1	3	+2,0909	4,3719	4,3719
4					
5					
	22	20			17,8190

$\overline{X} = \frac{20}{22} = 0,9091$ $\qquad \mu = np = k = 0,9091$

$\sigma^2 = \frac{17,8190}{22} = 0,8099$ $\qquad p = \frac{k}{n} = \frac{0,9091}{22} = 0,0413$

$$\varphi(k,m) = \frac{k^m}{m!} e^{-k} = \frac{0,9091^m}{m!} e^{-0,9091}$$

Usaremos logaritmos para calcular los valores de $\varphi(k,m)$

En el cuadro que sigue vemos que la comparación de las dos últimas columnas permite observar una concordancia buena entre los valores observados y los teóricos.

$k = 0{,}9091$ $\quad \log k = \bar{1}{,}9586$ $\quad k \log e = 0{,}3948$

m	mlogk	m!	log m!	m log k -log m! -k loge	$\varphi(k,m)$	$22\varphi(k,m)$	y
0	0	1	0	$\bar{1}.6052$	0,4029	9	9
1	$\bar{1}.9586$	1	0	$\bar{1}.5638$	0,3663	8	7
2	$\bar{1}.9172$	2	0,3010	$\bar{1}.2214$	0.1665	4	5
3	$\bar{1}.8758$	6	0,7782	$\bar{2}.7028$	0.0504	1	1
4	$\bar{1}.8344$	24	1,3802	$\bar{2}.0594$	0.0115	0	0
5	$\bar{1}.7830$	120	2,0792	$\bar{3}.3190$	0.0021	0	0
6	$\bar{1}.7516$	720	2,8573	$\bar{4}.5055$	0.0003	0	0
7	$\bar{1}.7102$	5040	3.7024	$\bar{5}.6130$	0.0004	0	0

Tabla de valores de e^{-k} para resolver ejercicios

k	e^{-k}	k	e^{-k}
0,1	0,9048	1	0,3679
0,2	0,8187	2	0,1353
0,3	0,7408	3	0,0498
0,4	0,6703	4	0,0183
0,5	0,6065	5	0,0067
0,6	0,5488	6	0,0025
0,7	0,4966	7	0,0009
0,8	0,4493	8	0,0003
0,9	0,4066	9	0,0001

3) Se tiran 3 dados 216 veces. Usando la distribución de Poisson como una aproximación, determinar la probabilidad de obtener 3 ases m veces. Calcular la distribución de probabilidad de X con dos cifras decimales. ¿Cuál es aquí la variable aleatoria?

3 dados 216 veces
obtener 3 ases m veces

$+ \; + \; + \quad$ ——— $= 1$

$+ \; + \; (5) \quad$ ——— $5\binom{3}{2} = 15$

$+ \; (5)(5) \quad$ ——— $25\binom{3}{1} = 75$

$(5)(5)(5) \quad$ ——— $= 125$

$\frac{}{216}$

$n = 216$

$p_{3+} = \frac{1}{216}$

$q = \frac{215}{216}$

$$\varphi(k,m) = \frac{k^m}{m!} e^{-m}$$

$A = k = np$

$$k = 216 \times \frac{1}{216} = 1$$

$$\varphi(67,5^m) = \frac{1}{m!} e^{-1} \quad \therefore \log \varphi = 0 - \log m! - \log e$$

$k = 1$. $\log k = 0$ $\log e = 0,4343$

m	m!	log m!	log φ	φ	216 φ
0	1	0	$\bar{1},5657$	0,3678	79
1	1	0	$\bar{1},5657$	0,3678	79
2	2	0,3010	$\bar{1},2647$	0,1840	40
3	6	0,7782	$\bar{2},7875$	0,0613	13
4	24	1,3802	$\bar{2},1855$	0,0153	4
5	120	2,0792	$\bar{3},4864$	0,0031	1
6	720	2,8573	$\bar{4},7804$	0,0005	0
7	5040	3,7024	$\bar{5},8633$	0,0000	0
				0,9998	216

4) La siguiente tabla da la distribución del número de vacantes ocurridas por año en el Tribunal Supremo de E.E.U.U. durante los años 1837 a 1932 (datos Wallis)

Nº de vacantes	y
0	59
1	27
2	9
3	1

ajustar una distribución de Poisson a esta distribución observada.

5) La distribución del número de fallecimientos de soldados de cuerpos de caballería prusiana debidos a coces de caballos en 200 años (Bortkiewiez) es:

Nº	y_i
0	109
1	65
2	22
3	3
4	1

Ajustar una distribución de Poisson a esta distribución.-

Est. A

29a. Conferencia

DISTRIBUCION NORMAL

Hemos visto al tratar la distribución binomial que las probabilidades que corresponden a los distintos valores que puede tomar el número de veces que se presenta el suceso esperado están expresadas por la fórmula:

$$P_m = C_n^m \, p^m \, q^{n-m}$$

y que estas probabilidades están definidas para valores enteros de m comprendidos entre 0 y n.- En otras palabras son probabilidades que varían mientras m toma los valores 0, 1, 2, 3 .... n; o lo que es igual son probabilidades dadas en el campo discontínuo.

Cuando los valores de n no son muy grandes, pueden calcularse los valores de P_m con ayuda de las tablas factoriales. También suele recurrirse a la utilización de la conocida fórmula asintòtica de Moivre-Stiuling, que expresa el valor de un factorial en función de e y π :

$$n! \cong n^n \, e^{-n} \sqrt{2\pi n}$$

Pero aún así los cálculos resultan penosos cuando n y m son grandes. Por esta razón es importante, sobre todo en Estadística, para cuando n es grande, buscar la ley límite de esta ley discontínua.

Es lo que haremos ahora, demostrando como el límite de las probabilidades discontínuas de la distribución binomial o de Bernoulli, es la ley de Laplace-Gaus que corresponde a la distribución normal, cuando la probabilidad favorable al suceso esperado es igual a la probabilidad contraria, o sea cuando:

$$p = q$$

Cuando n se hace grande, cada uno de los términos del desarrollo binomial se hace pequeño y en el caso de: $p = q = \frac{1}{2}$ los términos de la serie son:

$$P_m = C_n^m \, p^m q^{n-m} = C_n^m \left(\frac{1}{2}\right)^m \left(\frac{1}{2}\right)^{n-m} = C_n^m \left(\frac{1}{2}\right)^n$$

$$= \left(\frac{1}{2}\right)^n \left[1 + n + \frac{n(n-1)}{2!} + \frac{n(n-1)(n-2)}{3!} + \ldots\right]$$

La probabilidad para que el suceso favorable se presente m veces es:

$$\left(\frac{1}{2}\right)^n \frac{n!}{m!\,(n-m)!}$$

y la probabilidad para que se presente m + 1 veces se deduce de la fórmula anterior sin más que multiplicarla por $\frac{(n-m)}{m+1}$

Esta última frecuencia será por tanto mayor que la anterior siempre que:

$$n-m > m+1 \qquad \therefore \qquad n > 2m+1$$

o bien

$$m < \frac{n-1}{2}$$

Supongamos para mayor sencillez que n sea par, por ejemplo $n=2k$, entonces, la probabilidad para que el suceso favorable se presente k veces, será la mayor y su valor será:

$$P_o = \left(\frac{1}{2}\right)^{2k} \frac{(2k)!}{k!\,(2k-k)!}$$

$$= \left(\frac{1}{2}\right)^{2k} \frac{(2k)!}{k!\,k!}$$

Representando los valores de y_o para k = 0, 1, 2, ... se observa que el polígono de frecuencias se extiende simétricamente a cada lado de la ordenada máxima.

Consideremos el valor de la probabilidad para que el suceso favorable se presente k + x veces:

$$P_x = \left(\frac{1}{2}\right)^{2k} \frac{(2k)!}{(k+x)!\,(k-x)!}$$

y por tanto:

$$\frac{P_x}{P_o} = \frac{\left(\frac{1}{2}\right)^{2k} \frac{(2k)!}{(k+x)!\,(k-x)!}}{\left(\frac{1}{2}\right)^{2k} \frac{(2k)!}{k!\,k!}}$$

$$\frac{P_x}{P_o} = \frac{k(k-1)\,(k-2)\ldots(k-x+1)}{(k+1)\,(k+2)\ldots(k+x)}$$

expresión que podemos escribir después de dividir numerador y denominador por k^x

$$\frac{P_x}{P_o} = \frac{1-\frac{1}{k}\;\;1-\frac{2}{k}\;\cdots\;1-\frac{x-1}{k}}{1+\frac{1}{k}\;\;1+\frac{2}{k}\;\cdots\;1+\frac{x}{k}} \qquad (1)$$

Calculemos ahora su valor aproximado, suponiendo que k sea muy grande en comparación con x, de modo que pueda despreciarse $\frac{(x)^2}{(k)}$ comparado con $\frac{x}{k}$

Esta hipótesis no implica dificultad alguna, porque no hace falta considerar valores de x mucho mayores que tres veces la desviación típica o sea:

$$3\sqrt{k/2}$$

y el cociente de este valor por k es:

$$\frac{3\sqrt{k/2}}{k} = 3\sqrt{\frac{k}{2k^2}} = \frac{3}{\sqrt{2k}}$$

que necesariamente será pequeño cuando k sea grande.

En esta hipótesis, podemos aplicar el desarrollo de la serie logarítmica

$$\log_e (1+\delta) = \delta + \frac{\delta^2}{2} + \frac{\delta^3}{3} + \frac{\delta^4}{4} + \ldots$$

a cada uno de los paréntesis de la fracción (1) despreciando los términos que siguen al primero. Con esta aproximación (*) tendremos:

$$\log_e \frac{P_x}{P_o} = -\frac{2}{k}\left[1+2+3+\ldots+(x-1)\right] - \frac{x}{k}$$

(*)
$$\log\left(1-\frac{1}{k}\right) = -\frac{1}{k} - \frac{1}{2k^2} - \frac{1}{3k^3} - \ldots$$

$$\log\left(1-\frac{2}{k}\right) = -\frac{2}{k} - \frac{2^2}{2k^2} - \frac{2^3}{3k^3} - \ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\log\left\{1-\frac{x-1}{k}\right\} = -\frac{x-1}{k} - \frac{(x-1)^2}{2k^2} - \frac{(x-1)^3}{3k^3} - \ldots$$

$$-\log\left(1+\frac{1}{k}\right) = -\frac{1}{k} - \frac{1}{2k^2} - \ldots$$

$$-\log\left(1+\frac{2}{k}\right) = -\frac{2}{k} - \frac{2^2}{2k^2}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$-\log\left(1-\frac{x}{k}\right) = -\frac{x}{k} - \frac{x^2}{2k^2} - \ldots$$

luego sumando los primeros términos: $-2\frac{1}{k} - 2\frac{2}{k} - 2\frac{3}{k} - \ldots$

de donde: $$-\frac{2}{k}\left[1+2+\ldots+(x-1)\right] - \frac{x}{k}$$

$$\log_e \frac{P_x}{P_o} = -\frac{2}{k}\,\frac{(x-1)(x-1+1)}{2} - \frac{x}{k}$$

$$= -\frac{x\,(x-1)}{k} - \frac{x}{k}$$

$$= -\frac{x^2}{k} + \frac{x}{k} - \frac{x}{k}$$

$$= -\frac{x^2}{k}$$

Pasando a antilogaritmos resulta:

$$P_x = P_o\, e^{-\frac{x^2}{k}}$$

haciendo: $\frac{k}{2} = \sigma^2 \therefore \quad k = 2\sigma^2$

resulta:

$$P_x = P_o\, e^{-\frac{x^2}{2\sigma^2}}$$

CASO $p \neq q$

Este caso puede tratarse de manera análoga, si bien resulta más complicado.

La probabilidad para que el suceso favorable a parezca m veces era:

$$C_n^m p^m q^{n-m} = \frac{n!}{m!\,(n-m)!} p^m q^{n-m}$$

La probabilidad de $m+1$ aciertos se deduce de la anterior multiplicando por:

$$\frac{n-m}{m+1}\,\frac{p}{q}$$

y por tanto será mayor que la anterior siempre que:

$$\frac{n-m}{m+1}\,\frac{p}{q} > 1$$

$$(n-m)p > (m+1)q$$

$$np - mp > mq + q$$

$$np - q > mq + mp$$

$$np - q > m(q+p)$$

$$\boxed{np - q > m}$$

Supongamos que np sea un número entero. Puesto que n crece tendiendo a ∞, ello no impone restricción alguna a nuestro razonamiento.

La probabilidad máxima es, por tanto:

$$P_o = \frac{n!}{(np)!\,(nq)!} p^{np} q^{nq}$$

La probabilidad que se presenten $pn + x$ veces el suceso favorable es:

$$P_x = \frac{n!}{(np+x)!\,(nq-x)!} p^{np+x} q^{nq-x}$$

relacionando ambas probabilidades:

$$\frac{P_x}{P_o} = \frac{\dfrac{n!}{(np+x)!\,(nq-x)!}\, p^{np+x}\, q^{nq-x}}{\dfrac{n!}{(np)!\,(nq)!}\, p^{np}\, q^{nq}}$$

simplificando:

$$\frac{P_x}{P_o} = \frac{(np)!\,(nq)!}{(np+x)!\,(nq-x)!}\, p^x\, q^{-x} \qquad (1)$$

Por el teorema de Stirling para valores grandes de n se verifica:

$$n! = n^n\, e^{-n}\, \sqrt{2\pi n}$$

reemplazando en (1) se tiene:

$$\frac{P_x}{P_o} = \frac{(np)^{np} e^{-np} \sqrt{2\pi np}\, (nq)^{nq}\, e^{-nq} \sqrt{2\pi nq}\, p^x\, q^{-x}}{(np+x)^{np+x}\, e^{-(np+x)} \sqrt{2\pi(np+x)}\, (nq-x)^{nq-x} e^{-(nq-x)} \sqrt{2\pi(nq-x)}}$$

$$= \frac{n^{n(p+q)}\, e^{-n(p+q)}\, 2\pi n \sqrt{pq}\, p^{np}\, q^{nq}\, p^x\, q^{-x}}{(np+x)^{np+x}\, e^{-(np+x+nq-x)}\, (nq+x)^{nq-x}\, 2\pi \sqrt{(np+x)(nq-x)}}$$

$$= \frac{n^{n+1}\, p^{np+x+\frac{1}{2}}\, q^{nq-x+\frac{1}{2}}}{(np+x)^{np+x+\frac{1}{2}}\, (nq-x)^{nq-x+\frac{1}{2}}}$$

$$= \frac{1}{\left(1+\dfrac{x}{np}\right)^{np+x+\frac{1}{2}} \left(1-\dfrac{x}{nq}\right)^{nq-x+\frac{1}{2}}}$$

logaritmando:

$$\log_e \left(\frac{P_x}{P_o}\right) = -(np+x+\tfrac{1}{2}) \log\left(1+\frac{x}{np}\right) - (nq-x+\tfrac{1}{2}) \log\left(1-\frac{x}{nq}\right)$$

$$np + x + \tfrac{1}{2} + nq - x + \tfrac{1}{2} = np + nq + 1 = n(p+1) + 1 = n + 1$$

$$\log_e\left(\frac{P_x}{P_o}\right) = -(np+x+\tfrac{1}{2})\left\{\frac{x}{np} + \frac{x^2}{2n^2p^2} + \frac{x^3}{3n^3p^3} + \frac{x^4}{4n^4p^4} + \ldots\right\}$$

$$-(nq-x+\tfrac{1}{2})\left\{-\frac{x}{nq} \quad \frac{x^2}{2n^2q^2} \quad \frac{x^3}{3n^3q^3} \quad \frac{x^4}{4n^4q^4} \ldots \ldots\right\}.$$

y realizando operaciones resulta limitándonos a considerar términos hasta de segundo órden:

$$\log_e\left(\frac{P_x}{P_o}\right) = -x - \frac{x^2}{2np} - \frac{x^2}{np} - \frac{x}{2np} - \frac{x^2}{4n^2p^2} \ldots\ldots$$

$$+x + \frac{x^2}{2nq} - \frac{x^2}{nq} + \frac{x}{2nq} + \frac{x^2}{4n^2q^2}$$

reuniendo términos semejantes:

$$\log_e\left(\frac{P_x}{P_o}\right) = -\frac{x^2}{npq} + \frac{x^2(p^2+q^2)}{4n^2p^2q^2} + \frac{p-q}{2npq}x + \frac{q^2-p^2}{6n^2p^2q^2}x^3$$

$$+\text{términos de } \frac{1}{n^3} \text{ y órdenes superiores}$$

Por ser $p+q=1$ y despreciando los términos que contengan $\frac{1}{n^3}$ y órdenes superiores que, para grandes valores de n, son pequeños si se comparan con los demás:

$$q^2 - p^2 = (q-p)(q+p) = q - p$$

$$\log_e\left(\frac{P_x}{P_o}\right) = -\frac{x^2}{2npq} + \frac{x^2(p^2+q^2)}{4n^2p^2q^2} + \frac{q-p}{2npq}\left[-x + \frac{x^3}{3npq}\right]$$

Por ser $npq = \sigma^2$ y si n es grande, el segundo término será pequeño comparado con el primero. Además, veremos que no es necesario considerar valores de x/σ mucho mayores que 3, podemos prescindir de todo el tercer término cuando $\frac{q-p}{npq}$ sea pequeño.

Con estas hipótesis tenemos:

$$\log_e\left(\frac{P_x}{P_o}\right) = -\frac{x^2}{2\sigma^2}$$

o bien:

$$P_x = P_o\, e^{-\frac{x^2}{2\sigma^2}}$$

como antes.

Por tanto, sean n no iguales p y q, la distribución binomial tiende, al crecer n a la forma de la curva continua, por lo menos en la parte sustancial del campo de variación.

Es, en realidad sorprendente la aproximación de la distribución binomial a la curva, hasta para valores relativamente bajos de n, siempre que p y q sean casi iguales.

la curva $P_x = P_o\, e^{\frac{x^2}{2\sigma^2}}$

recibe el nombre de curva normal

Propiedades de la curva normal

La curva normal es simétrica respecto del punto $x=0$ puesto que su ecuación es independiente del signo de x. En ese punto tiene la ordenada su valor máximo.

La media, la mediana y el modo coinciden.

Valuación de y_o (Smith y Duncan)

Por definición es:

$$P_o = \frac{n!}{(np)!\,(nq)!}\, p^{np}\, q^{nq}$$

Utilizando la fórmula de Stirling:

$$y_o = \frac{n^n\, e^{-n}\, \sqrt{2\pi n}\; p^{np}\; q^{nq}}{np^{np}\, e^{-np}\, \sqrt{2\pi np}\;\; nq^{nq}\, e^{-nq}\, \sqrt{2\pi nq}}$$

$$= \frac{n^n\, e^{-n}\, \sqrt{2\pi n}\; p^{np}\, q^{nq}}{n^{np} n^{nq}\, \sqrt{2\pi n}\, \sqrt{p}\; p^{np}\, e^{-np-nq}\, q^{nq}\, \sqrt{2\pi nq}}$$

$$= \frac{1}{\sqrt{2\pi}\;\sqrt{npq}} = \frac{1}{\sigma\sqrt{2\pi}}$$

luego

$$\boxed{P_x = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}}$$

que es la ecuación de la distribución normal.

Est. A

3la. Conferencia

FUNCION DE DENSIDAD

La ecuación correspondiente a la distribución normal era:

$$P_x = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}$$

como estamos ahora tratando con una variable aleatoria continua hemos de cambiar el símbolo Px por el de f(x) que indica más claramente a la función de x en el campo continuo, será entonces:

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

El valor que la función f(x) toma para cada valor particular de x se llama "densidad en el punto x"

Al definir una variable aleatoria debemos suponer conocido su campo de variación que estará representado por un intervalo (a,b) finito o infinito.

Si P(c,d) es la probabilidad de que x se halle comprendida en el intervalo (c,d) se llega a que:

1º) $P(c,d) \geqq 0$

2º) $P(a,b) = 1$

3º) $P(c,d) = P(c,e) + P(e,d)$

supuesto (c,d) dividido en dos subintervalos (c,e) y (e,d)

Si la densidad f(x) se conoce, la masa contenida en cualquier intervalo (c,d) está representada por la integral

$$P(c,d) = \int_c^d f(x)dx$$

que es también la probabilidad correspondiente al intervalo considerado.

Esta expresión satisface las tres condiciones antes indicadas si la densidad de la probabilidad f(x) satisface estas dos condiciones:

1) $f(x) \geqq 0$ para todos los valores de x en(a,b)

2) $\int_a^b f(x)\, dx = 1$

El intervalo que contiene todos los valores posibles de la variable aleatoria, puede ser finito o infinito de acuerdo con la naturaleza de la variable. Sin embargo, en todos los casos podemos tomar el mayor intervalo posible entre $-\infty$ y $+\infty$ para lo cual basta definir la densidad fuera del intervalo originalmente dado como nula.

Por lo tanto, la densidad estará definida para todos los valores reales de x y satisfará las condiciones

1) $f(x) \geq 0$ para todos los valores de x

2) $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Además la probabilidad de x de encontrarse en cualquier intervalo (c,d) estará dado, como hemos visto por:

$$\int_c^d f(x)\, dx \qquad (1)$$

Est. A

32a. Conferencia

FUNCION DE DISTRIBUCION DE LA PROBABILIDAD

Si en (1) tomamos $c = -\infty$ y escribimos t en lugar del límite superior d, tendremos una función que depende exclusivamente de este límite y que representamos por $F(t)$, por tanto será:

$$F(t) = \int_{-\infty}^{t} f(x)\, dx$$

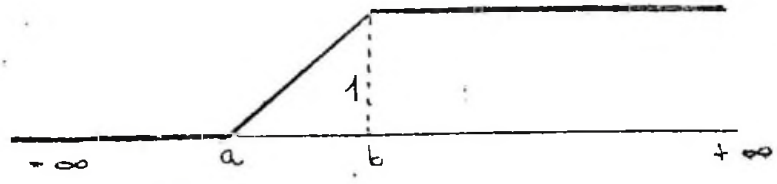
que representa la probabilidad de que x no sobrepase o sea menor que t

Considerada como función de t la función $F(t)$ nunca es decreciente y varía entre los valores:

$$F(-\infty) = 0$$
$$F(+\infty) = 1$$

Esta función se llama la función de distribución de la probabilidad

En el caso de que x posea una distribución uniforme de la probabilidad en el intervalo (a,b) su función de distribución está definida así



Para $t < a$ es $F(t) = 0$

Para $a \leq t \leq b$ es $F(t) = \dfrac{t - a}{b - a}$

Para $t > b$ es $F(t) = 1$

Media aritmética de la distribución normal

La definición de valor medio o esperanza matemática está expresada por:

$$\mu = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-\frac{x^2}{2\sigma^2}}\, dx = 0$$

Sabemos que la media aritmética es el momento de primer órden y por tratarse de una curva simétrica con origen en el valor medio, todos los momentos de orden impar se anulan es decir

$$\mu_{2k+1} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k+1}\, e^{-\frac{x^2}{2\sigma^2}}\, dx = 0$$

Varianza de la distribución normal

Por definición es:

$$\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2\, e^{-\frac{x^2}{2\sigma^2}}\, dx = \mu_2 - \mu_1^2$$

$$\sigma^2 = \mu_2$$

Est. A

33a. Conferencia

MOMENTOS DE LA DISTRIBUCION NORMAL

Ya vimos al tratar del valor medio que todos los momentos de órden impar se anulan.

En lo que respecta a los momentos de órden par, recordando que por definición se tiene:

$$\mu_{2k} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k} \, e^{-\frac{x^2}{2\sigma^2}} \, dx$$

y sin entrar a la resolución de esta integral para cuando k toma los sucesivos valores 0, 1, 2, 3.... haremos solo mención del resultado final que nos da:

$$\mu_{2k} = \frac{(2k)!}{2^{2k}\,\left(\frac{2k}{2}\right)!}\,\sigma^{2k}$$

si $k = 0$ $\quad \mu_0 = 1$

$k = 1$ $\quad \mu_2 = \sigma^2$

$k = 2$ $\quad \mu_4 = \sigma^4$

y así sucesivamente.

En virtud de estos resultados y recordando que los momentos de orden impar son todos iguales a cero, resulta que

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

$$\gamma_1 = \beta_1 = 0$$

$$\gamma_2 = \beta_2 - 3 = 0$$

o sea que la curtosis en la curva normal es cero. Esto explica por que se eligió el valor 3, aparentemente arbitrario para definir las curvas platicúrticas y leptocúrticas.

Est. A

34à. Conferencia

PROPIEDADES DE LA CURVA NORMAL

La representación gráfica de la expresión

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

recibe el nombre de curva normal.

Las aplicaciones de la curva normal no se limitan a las distribuciones del tipo binomial, pero antes de referirnos a sus muchas aplicaciones prácticas y teóricas, expondremos brevemente sus principales propiedades.

1) La curva normal es simétrica respecto del punto $x = 0$ puesto que su ecuación es independiente del signo de x

Para $x = 0$, toma la ordenada su valor máximo

2) La media aritmética, la mediana y el modo coinciden en la curva normaa

3) La curva queda completamente determinada cuando se define la media aritmética (origen de las x) y la desviación típica σ

En la práctica, cuando se trata de ajustar una curva normal a los datos obtenidos de una experiencia, no tenemos directamente el valor de P_o, sino que debemos calcularlo imponiendo la condición de que el área de la curva sea igual, en la escala elegida, al número de observaciones.

Por esta razón nos interesa hallar el área limitada por la curva

$$P_x = P_o\, e^{-\frac{x^2}{2\sigma^2}}$$

Hemos visto que el área de un histograma, es decir el total de observaciones que representa, está expresado por:

$$\sum_{i=1}^{n} X_i\, y_i = \text{Area}$$

donde X_i representa el valor central de cada intervalo de clase, y_i la frecuencia observada para cada intervalo

i = 1, 2, 3 ... n y n el número de intervalos.

Cuando el histograma tiende a la curva continua, la amplitud de los intervalos disminuye y el número de términos de la suma aumenta. En la curva normal, que se extiende indefinidamente a ambos lados de la media, el límite a que tiende la suma al hacerse los intervalos infinitamente pequeños y el número de términos infinitamente grandes, se representa por:

$$\int_{-\infty}^{+\infty} P_o \, e^{-\frac{x^2}{2\sigma^2}} dx \qquad (1)$$

en donde el signo $\int$ es una forma convencional del signo Σ de sumación y dx representa el valor infinitamente pequeño de h

Esta es la notación que se emplea en el cálculo integral y la cantidad expresada en (1) se lee: integral de la curva normal entre $-\infty$ y $+\infty$

Hemos demostrado que:

$$P_o = \frac{1}{\sigma\sqrt{2\pi}}$$

por lo que la curva normal se presenta bajo la expresión:

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

y limita un área igual a la unidad.

En el caso de una experiencia práctica en la que la frecuencia es N, la correspondiente curva normal es:

$$f(x) = \frac{N}{\sigma\sqrt{2\pi}} \; e^{\frac{x^2}{2\sigma^2}}$$

Est. A

35a. Conferencia

ORDENADAS DE LA CURVA NORMAL

La curva normal tiene tanta importancia que se han preparado tablas que dan valores para la llamada función reducida de Laplace-Gauss

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (1)$$

expresión a la que se llega, tomando a σ como unidad de medida.

1) Tenemos así la tabla que nos da la ordenada correspondiente a cada valor de x/σ (variable reducida) o sea el valor de $\varphi(x)$ en la expresión (1)

2) Tabla con las áreas limitadas por la curva a la derecha y a la izquierda de cierta ordenada, esto es, los valores de:

$$\frac{1}{\sqrt{2\pi}} \int_{x_1}^{\infty} e^{-\frac{x^2}{2}} dx \qquad \text{y} \qquad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-\frac{x^2}{2}} dx$$

3) Tabla con los valores de las probabilidades de que una observación se encuentre fuera de los límites $\pm \frac{x}{\sigma}$ en la curva normal de errores. Donde la probabilidad buscada está expresada por:

$$P = 2(1 - A)$$

siendo A el área que se da en la tabla que hemos indicado en el párrafo anterior.

Con aquellas tres tablas tenemos posibilidad además de resolver los problemas directos a que cada una se refiere, la de hallar el área comprendida entre las ordenadas de abscisas x_1 y x_2

Bastará para ello calcular la diferencia entre los siguientes valores:

$$\frac{1}{\sqrt{2\pi}} \int_{x_1}^{\infty} e^{-\frac{y^2}{2}} dx - \frac{1}{\sqrt{2\pi}} \int_{x_2}^{\infty} e^{-\frac{x^2}{2}} dx$$

También tiene amplia aplicación estadística la función de distribución acumulativa F(t) para una distribución normal que tiene μ por media y σ^2 por varianza.

Esta función se expresa así:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \qquad (2)$$

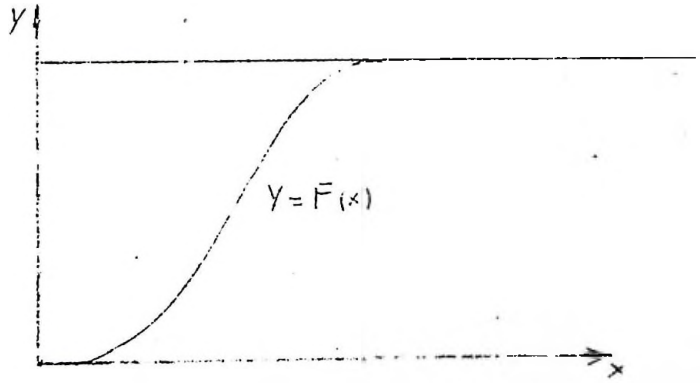
Si X es una variable aleatoria continua que tiene una distribución normal con media μ y σ por desviación típica, para un valor de x, por ejemplo x_1 resulta que $F(x_1)$ es simplemente la probabilidad de que la variable aleatoria X tome valores menores o iguales a x_1, simbólicamente sería:

$$P(X \leq x_1) = F(x_1)$$

Como estamos considerando aquí el caso de una variable aleatoria continua el doble signo $\leq$ puede reemplazarse por el signo $<$ y tendríamos que:

$$P(X < x_1) = F(x_1)$$

El gráfico acumulativo nos permitiría determinar en forma aproximada y directa el valor de esta probabilidad.



Si el problema fuese determinar la probabilidad de que la variable aleatoria continua pueda tomar valores comprendidos entre x_1 y x_2, donde x_1 y x_2 son dos valores particulares de x entonces tendremos:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

que es una forma de expresar la probabilidad buscada como diferencia entre las ordenadas de la curva de probabilidad acumulativa F(x)

Est. A

36a. Conferencia

La función de densidad de probabilidad f(x) que se obtiene diferenciando F(x) dado en (2) es la siguiente:

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

Aún cuando hemos visto que:

$F(x) = 0$ para $x = -\infty$

$F(x) = 1$ para $x = \infty$

prácticamente F(x) varía desde un valor próximo a cero hasta un valor próximo a uno, al recorrer el intervalo $\mu - 3\sigma$ a $\mu + 3\sigma$

Los valores de F(x) para valores de x que oscilan desde $\mu - 3\sigma$ hasta $\mu + 3\sigma$ en intervalos de $0{,}1\sigma$ están dados en la tabla siguiente.

VALORES DE LA DISTRIBUCION NORMAL ACUMULATIVA $F_N(X)$

Z	X	$F_n(X)$	Z	X	$F_N(X)$
-3,0	$\mu-3{,}0\sigma$	0,0013	0,1	$\mu+0{,}1\sigma$	0,5398
-2,9	$\mu-2{,}9\sigma$	0,0019	0,2	$\mu+0{,}2\sigma$	0,5793
-2,8	$\mu-2{,}8\sigma$	0,0026	0,3	$\mu+0{,}3\sigma$	0,6179
-2,7	$\mu-2{,}7\sigma$	0,0035	0,4	$\mu+0{,}4\sigma$	0,6554
-2,6	$\mu-2{,}6\sigma$	0,0047	0,5	$\mu+0{,}5\sigma$	0,6915
-2,5	$\mu-2{,}5\sigma$	0,0062	0,6	$\mu+0{,}6\sigma$	0,7258
-2,4	$\mu-2{,}4\sigma$	0,0082	0,7	$\mu+0{,}7\sigma$	0,7580
-2,3	$\mu-2{,}3\sigma$	0,0107	0,8	$\mu+0{,}8\sigma$	0,7881
-2,2	$\mu-2{,}2\sigma$	0,0139	0,9	$\mu+0{,}9\sigma$	0,8159
-2,1	$\mu-2{,}1\sigma$	0,0179	1,0	$\mu+1{,}0\sigma$	0,8413
-2,0	$\mu-2{,}0\sigma$	0,0227	1,1	$\mu+1{,}1\sigma$	0,8643
-1,9	$\mu-1{,}9\sigma$	0,0287	1,2	$\mu+1{,}2\sigma$	0,8849
-1,8	$\mu-1{,}8\sigma$	0,0359	1,3	$\mu+1{,}3\sigma$	0,9032
-1,7	$\mu-1{,}7\sigma$	0,0446	1,4	$\mu+1{,}4\sigma$	0,9192
-1,6	$\mu-1{,}6\sigma$	0,0548	1,5	$\mu+1{,}5\sigma$	0,9332
-1,5	$\mu-1{,}5\sigma$	0,0668	1,6	$\mu+1{,}6\sigma$	0,9452
-1,4	$\mu-1{,}4\sigma$	0,0808	1,7	$\mu+1{,}7\sigma$	0,9554
-1,3	$\mu-1{,}3\sigma$	0,0968	1,8	$\mu+1{,}8\sigma$	0,9641
-1,2	$\mu-1{,}2\sigma$	0,1151	1,9	$\mu+1{,}9\sigma$	0,9713
-1,1	$\mu-1{,}1\sigma$	0,1357	2,0	$\mu+2{,}0\sigma$	0,9773
-1,0	$\mu-1{,}0\sigma$	0,1587	2,1	$\mu+2{,}1\sigma$	0,9821
-0,9	$\mu-0{,}9\sigma$	0,1841	2,2	$\mu+2{,}2\sigma$	0,9861
-0,8	$\mu-0{,}8$	0,2119	2,3	$\mu+2{,}3\sigma$	0,9893
-0,7	$\mu-0{,}7$	0,2420	2,4	$\mu+2{,}4\sigma$	0,9918
-0,6	$\mu-0{,}6$	0,2742	2,5	$\mu+2{,}5\sigma$	0,9938
-0,5	$\mu-0{,}5\sigma$	0,3085	2,6	$\mu+2{,}6\sigma$	0,9953
-0,4	$\mu-0{,}4\sigma$	0,3446	2,7	$\mu+2{,}7\sigma$	0,9965
-0,3	$\mu-0{,}3\sigma$	0,3821	2,8	$\mu+2{,}8\sigma$	0,9974
-0,2	$\mu-0{,}2\sigma$	0,4207	2,9	$\mu+2{,}9\sigma$	0,9981
-0,1	$\mu-0{,}1\sigma$	0,4602	3,0	$\mu+3{,}0\sigma$	0,9987
0	μ	0,5000			

Es útil a los fines prácticos y para trabajar con la distribución normal, hacer el siguiente cambio de variable

$$Z = \frac{x - \mu}{\sigma}$$

o sea

$$x = \mu + z\sigma$$

y considerar los valores de Z correspondientes a los x elegidos o viceversa.

Estos valores de Z están dados también en el cuadro anterior.

Si es X una variable aleatoria con media μ y desviación típica σ, entonces Z es también una variable aleatoria con media 0 y desviación típica 1

En la práctica esta aproximación es válida corresponda o no X(y Z) a una distribución normal

Por ejemplo: sea X una variable aleatoria que inidica el número total de "caras" que resultan arrojando 10 monedas

Siendo $n = 100$

$p = \frac{1}{2}$

la media de X es: $\mu = np = 5$

y la desviación típica : $\sigma = \sqrt{npq} = 1,581$

Si hacemos: $Z = \frac{X - 5}{1,581}$

será Z una variable aleatoria que tiene:

$$\mu = 0 \quad y \quad \sigma = 1$$

Aún cuando en este ejemplo X y también Z es una variable aleatoria discreta, su probabilidad acumulativa puede ser ajustada aproximadamente por una distribución normal acumulativa como veremos.

En la tabla anterior se observa que los valores de F(x) correspondientes a $\mu \pm \sigma z$ son simétricos respecto de 0,5 y que su suma, por consiguiente es 1

Por ejemplo para:

$x = \mu \pm 1,5\sigma$ o sea para $z = \pm 1,5$

se tiene: $F(x) = 0,5 \pm 0,4332$

la suma de los dos resultados que se obtienen es igual a 1, en efecto:

$$0,5 + 0,4332 = 0,9332$$
$$0,5 - 0,4332 = 0,0668$$
$$1,0000$$

Si a una variable aleatoria X, corresponde una distribución normal, con media μ y desviación típica σ, se puede determinar con ayuda de la tabla anterior la probabilidad de que X caiga en un intervalo dado.

Ejemplo:1) Sea X una variable aleatoria con distribución normal y cuya

$$\mu = 30$$
$$\sigma = 5$$

¿Cuál es la probabilidad de que $26 < X \leq 40$?

Se tendrá:

$$P(26 < X \leq 40) = F(40) - F(26) \qquad (3)$$

Para hallar los valores de F(40) y F(26) se utilizará la relación entre x y Z

$$Z = \frac{x - 30}{5}$$

para $X = 40$ $\quad Z = \frac{40 - 30}{5} = 2 \quad F(40) = 0,9773$

para $X = 26$ $\quad Z = \frac{26 - 30}{5} = -0,8 \quad F(26) = 0,2119$

Reemplazando en (3) estos valores resulta

$$P(26 < X \leq 40) = 0,9773 - 0,2119 = 0,7654$$

Est. A

37a. Conferencia

A veces interesa la probabilidad de que X difiera de la media μ en un cierto múltiplo de σ

Ejemplo 2) ¿Cuál es la probabilidad de que X difiera de μ en menos de σ ?Es decir ¿Cuál es el valor de

$$P(\mu-\sigma \leq x \leq \mu+\sigma)$$

o también:

$$P(|X-\mu| \leq \sigma)\ ?$$

De la tabla se deduce que:

$$P(\mu-\sigma \leq X \leq \mu+\sigma) = P(-1 < Z \leq 1) = F(\mu+\sigma)-F(\mu-\sigma)$$

$$= 0{,}8413 - 0{,}1587 = 0{,}6826$$

Análogamente se observa en la tabla que:

$$P(|X-\mu| \leq 2\sigma) = P(-2 < Z \leq 2) = 0{,}9773 - 0{,}0227$$
$$= 0{,}9546$$

y que:

$$P(|X-\mu| \leq 3\sigma) = P(-3 < Z \leq 3) = 0{,}9987 - 0{,}0013$$
$$= 0{,}9974$$

Estos resultados se expresan así:

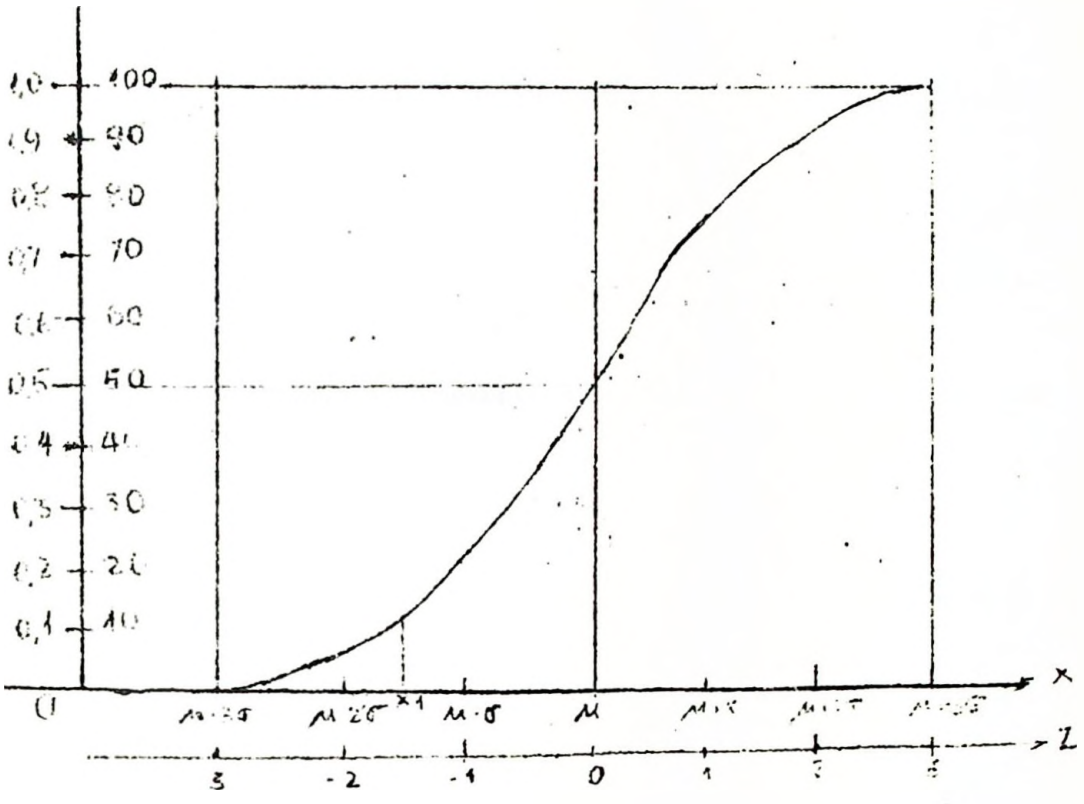
Si los valores de una población están normalmente distribuídos con media μ y desviación típica σ

el 68,26% de los valores difieren de μ menos de 1σ

el 95,46% " " " " " μ " " 2σ

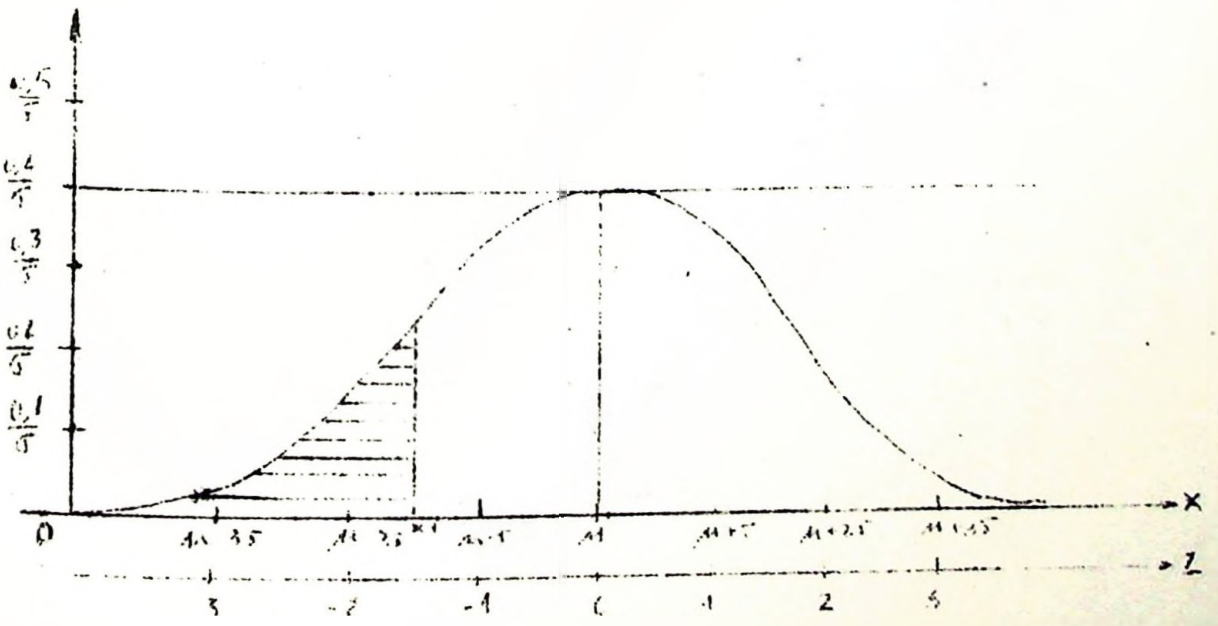
el 99,74% " " " " " μ " " 3σ

El siguiente gráfico corresponde a la representación de la función F(x) donde se indican las dos escalas x y z



En las aplicaciones se eligen los valores convenientes, pero suelen marcarse al menos, los siguientes valores de Z: -3, -2, -1, 0, 1, 2, 3

La representación gráfica de la función f(x) de densidad de probabilidad, es la siguiente:



La relación entre las dos representaciones gráficas es la siguiente:

El valor numérico de la ordenada en el primer gráfico para el valor x_1 es igual al valor numérico del área encerrada bajo la curva del segundo gráfico y situado a la izquierda de x_1.

Prácticamente se utiliza más la función acumulativa.

Est. A

38a. Conferencia

LA DISTRIBUCION DE ERRORES

Llegamos a la distribución normal como un caso límite de la distribución binomial cuando el exponente n es muy grande.

Pero esta no es sino una más de las diferentes formas en que la curva normal aparece en la literatura estadística. Gauss llegó a ella por un razonamiento distinto: por la investigación de la ley de distribución a que obedecen los errores de observación para que la media aritmética de una serie de medidas sea el valor más probable de la "verdadera" magnitud.

LLega a demostrarse que la distribución de los errores x alrededor del valor verdadero (tomado como 0) viene dado por ley

$$y = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{x^2}{2\sigma^2}}$$

En la teoría de errores es más corriente escribir

$$h^2 = \frac{1}{2\sigma^2}$$

con lo cual la distribución será:

$$y = \frac{h}{\sqrt{\pi}} \; e^{-h^2x^2}$$

donde h recibe el nombre de "precisión"

Al aumentar h la curva normal se estrecha y por eso h mide en cierto sentido la "condensación" de la mayoría de las observaciones en torno al verdadero valor.

Est. A

39a. Conferencia

APLICACION DE LA DISTRIBUCION NORMAL

La distribución normal, aún cuando no corresponda a ningún amplio campo de la Naturaleza, no disminuye su importancia en la teoría estadística.

La utilización de esta distribución ha aumentado en los últimos años por varias razones.

1) La curva normal y su integral tienen muchas propiedades matemáticas en virtud de las cuales aquellas resultan útiles y relativamente fáciles de manejar.

Además la curva normal se adapta bastante bien a muchas distribuciones campaniformes (unimodales). Por tanto, cuando se desconozca la naturaleza exacta de una distribución de frecuencias campaniforme o conozcamos su forma, pero esta sea difícil de tratar matemáticamente podemos suponer en primera aproximación, que la distribución es normal y luego comprobar hasta que punto es cierta esta hipótesis. No es raro que, a los fines de una determinada investigación resulte bastante precisa la representación de un colectivo mediante la curva normal.

2) Al considerar las distribuciones obtenidas de muestras, muchos de los colectivos que se presentan siguen la ley normal, bien sea exactamente, o bien con un grado aceptable de aproximación.

En virtud de estas razones la distribución normal se aplica:

a) Para aproximar o ajustar una distribución de medidas de una muestra bajo ciertas condiciones.

b) Aproximar la distribución binomial u otras distribuciones de probabilidad discretas o continuas, que cumplan también ciertas condiciones.

c) Aproximar la distribución de las medias u otros estadígrafos calculados a partir de las muestras especialmente de las grandes muestras.

ercicio 1) Hallar la frecuencia representada por la menor de las areas limitada por la curva

$$y = \frac{10.000}{4\sqrt{2x}} e^{-\frac{x^2}{32}}$$ y la ordenada $x = 7$

$$\sigma = 4$$

$$\frac{x}{\sigma} \equiv \frac{7}{4} = 1,75$$

para 1,7 área mayor = 0,95543

1,8 " " = 0,96407

Interpolando

0,1 ——— 0,00864

0,05 ——— $\frac{0,00864 \times 0,05}{0,1} = 0,004320$

Luego: 0,95543
0,00432
0,95975 área mayor

área menor 1 - 0,95975 = 0,04025

multiplicada por 10.000 nos da la frecuencia buscada: 402,5

Est. A

40a. Conferencia

Ejercicio 2) Se tiran 100 monedas un cierto número de veces. ¿Qué número aproximado de veces es de esperar que se obtengan en 10.000 tiradas?
65 caras exactamente

1) Nº de caras dadas por el desarrollo

$$10.000\ (\tfrac{1}{2} + \tfrac{1}{2})^{1.000}$$

$$\sigma = \sqrt{\tfrac{1}{2}.\tfrac{1}{2}.100} = 5$$

$$\frac{N}{\sigma} = \frac{10.000}{5} = 2.000$$

y el exponente es bastante grande para considerar la distribución como normal.

$np = 100.\frac{1}{2} = 50$ media del Nº de caras

$65 - 50 = 3\sigma$

La frecuencia de una desviación igual a 3σ se halla en la Tabla I y es igual a

$$2.000 \times 0{,}00443 = 8{,}86$$

o sea unas nueve tiradas entre 10.000.

Un resultado de 65 caras es, pues de esperar que ocurra unas nueve veces en 10.000 tiradas

Ejercicio 3) Hallar la ordenada de la curva normal dada por la ecuación:

$$y = \frac{10.000}{4\sqrt{2\pi}}\ e^{-\frac{x^2}{32}}$$

para el valor de la variable $x = 7$

Es $N = 10.000$

$\sigma = 4$

Modificar el valor de σ equivale a modificar la la escala de las x

La ordenada de esta curva para $x = 7$ será la mis

ma que la ordenada de la curva que tenga por desviación típica la unidad, para el valor de $x = 7/4 = 1,75$

Según T. 1 apéndice (pag. 639)

$x = 1,8 \qquad y = 0,07895$

$x = 1,7 \qquad y = 0,09405$

interpolando:

$0,1 \text{ ——— } 0,01510$

$0,05 \text{ ——— } \dfrac{0,01510 \times 0,05}{0,1} = 0,00755$

$x = 1,75 \quad \therefore \quad \begin{array}{r} 0,07895 \\ 0,00755 \\ \hline 0,08650 \end{array}$

La ordenada es igual a este último valor multiplicado por $\frac{10.000}{4}$.

o sea: $\dfrac{0,08650 \times 10.000}{4} = 216$

Est. A

41a. Conferencia

Ejercicio 4) Ajustar los índices cefálicos de alumnas en la Escuela Normal Nº 2 de Rosario.

Observaciones	y_i	x'_i	$x'_i y_i$	$(x'_i)^2$	$(x'_i)^2 y_i$	$x'_i - \bar{x}'$	$(x'_i - \bar{x}')^2$
70 - 72	7	-5	- 35	25	175	-4,5888	21,0571
72	7	-4	- 28	16	112	-3,5888	12,8795
74	45	-3	-135	9	405	-2,5888	6,7019
76	108	-2	-216	4	432	-1,5888	2,5243
78	146	-1	-146	1	146	-0,5888	0,3467
80	151	0		0	0	0,4112	0,1691
82	98	1	98	1	98	1,4112	1,9915
84	54	2	108	4	216	2,4112	5,8139
86	16	3	48	9	144	3,4112	11,6363
88	8	4	32	16	128	4,4112	19,4587
90 - 92	2	5	10	25	50	5,4112	29,2811
	642		-264		1906		

$\frac{(x'_i - \bar{x})^2}{2\sigma^2}$	$\frac{(x'_i - \bar{x}')^2}{2\sigma^2} \log e$ (x)	$\log \frac{1}{\sigma\sqrt{2\pi}}$ -(x)	log.frec. relat.teor.	frec. relat. y_i	frec. absol.
3,7606	1,6332	-2,2559	$\bar{3}$,7441	0,0055	4
2,3002	0,9990	-1,6217	$\bar{2}$,3783	0,0239	15
1,1969	0,5198	-1,1425	$\bar{2}$,8575	0,0720	46
0,4508	0,1958	-0,8185	$\bar{1}$,1815	0,1519	98
0,0619	0,0269	-0,6496	$\bar{1}$,3504	0,2241	144
0,0302	0,0131	-0,6358	$\bar{1}$,3642	0,2313	148
0,3557	0,1545	-0,7772	$\bar{1}$,2228	0,1670	107
1,0383	0,4509	-1,0736	$\bar{2}$,9264	0,0844	54
2,0781	0,9025	-1,5252	$\bar{2}$,4748	0,0298	19
3,4751	1,5092	-2,1319	$\bar{3}$,8681	0,0074	5
5,2293	2,2711	-2,8938	$\bar{3}$,1062	0,0013	1
					641

$$\bar{X}' = m_1 = -\frac{264}{642} = -0,4112$$

$$\sigma^2 = m_2 - m_1^2 = \frac{1906}{642} = (0,4112)^2 = 2,9688 - 0,1691 = 2,7997$$

$$\sigma = \sqrt{2,7997} = 1,6732$$

Reemplazando estos valores en la función normal

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

$$= \frac{1}{1,6732\sqrt{2\pi}}\, e^{-\frac{(x_i'+0,4112)^2}{2(1,6732)^2}}$$

$$= \frac{1}{4,1942}\, e^{\frac{-(x_i'+0,4112)^2}{5,5994}}$$

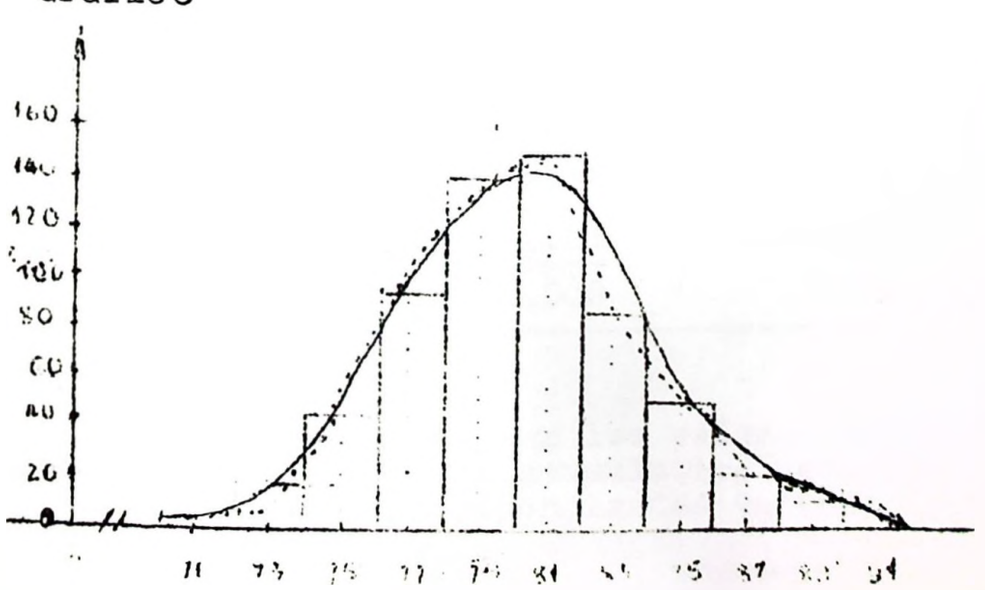
logaritmando resulta:

$$\log f(x) = \log 0,2384 - \frac{(x_i' + 0,4112)^2}{5,5994} \log e$$

$$= \bar{1},3773 - 0,4343\, \frac{(x_i' + 0,4112)^2}{5,5994}$$

Con esta ecuación se opera en la tabla y luego se buscan los antilogari mos para, finalmente, hallar las frecuencias absolutas multiplicando por la población 642

Gráfico



Ejercicio 5) Ajustar una distribución normal acumulativa a una binomial para $n=10$ y $p=\frac{1}{2}$

En este caso es:

$$\mu = np = 10 \cdot \tfrac{1}{2} = 5$$

$$\sigma = \sqrt{npq} = \sqrt{10 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2}} = 1{,}581$$

$$z = \frac{x - 5}{1{,}581}$$

Se calcula la probabilidad binomial por la fórmula:

$$P_x = C_n^x \, p^x \, q^{n-x} = C_{10}^x \, (\tfrac{1}{2})^{10}$$

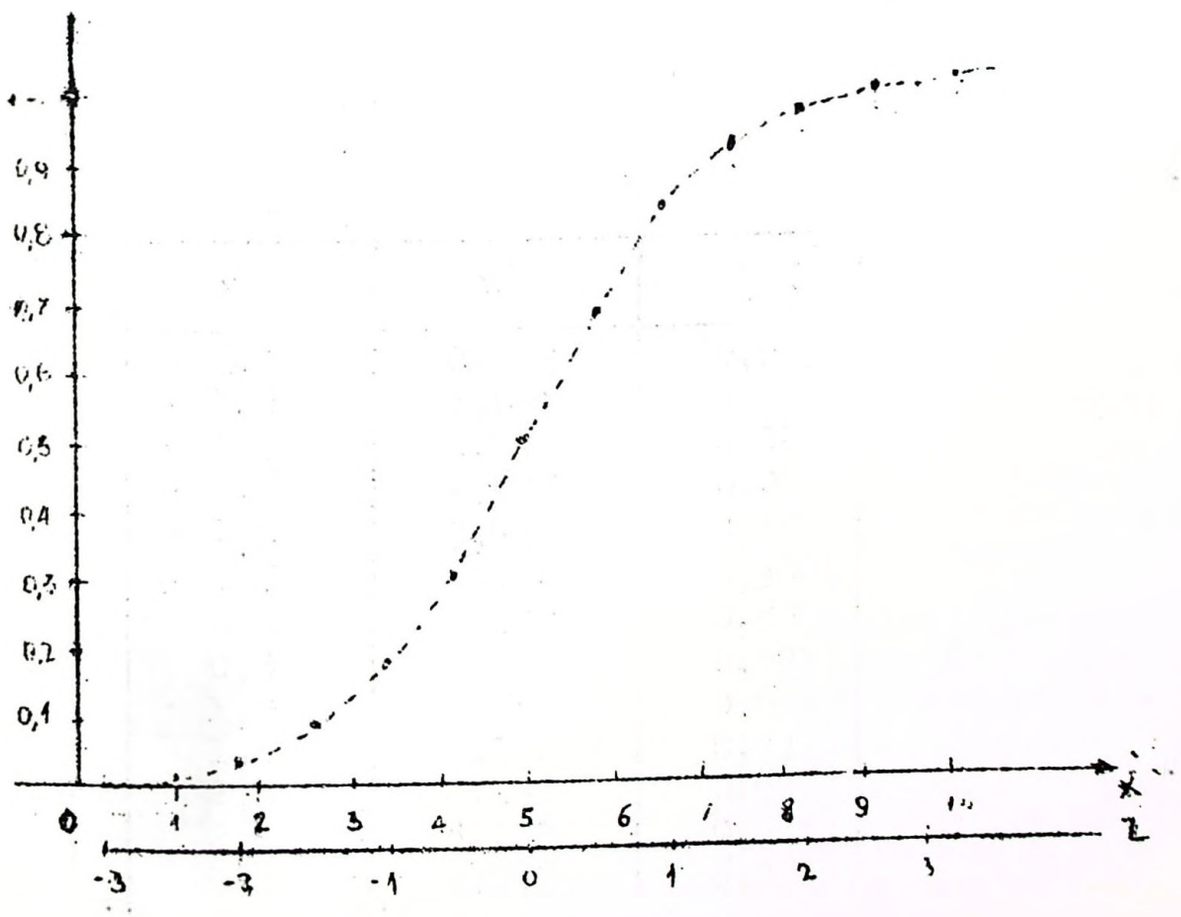
damos a x valores sucesivos 0, 1, 2, ...10 pues sabemos que al tirar 10 veces una moneda pueden salir 0, 1, 2,... ..10 caras.

Los valores obtenidos para la distribución de probabilidad $f_B(x)$ y para la acumulativa $F_B(x)$ son:

x	f_B (x)	F_B (x)
0	0,001	0,001
1	0,010	0,011
2	0,044	0,055
3	0,117	0,172
4	0,205	0,377
5	0,246	0,623
6	0,205	0,828
7	0,117	0,945
8	0,044	0,989
9	0,010	0,999
10	0,001	1,000
	1,000	

Para calcular los valores correspondientes de la distribución normal acumulativa ya hemos determinado z y buscamos los correspondientes valores de x y $F_N(x)$ en la tabla:

z	x	$F_N(x)$
- 3,0	0,257	0,0013
- 2,5	1,048	0,0062
- 2,0	1,838	0,0227
- 1,5	2,628	0,0668
- 1,0	3,419	0,1587
- 0,5	4,210	0,3085
0,0	5,000	0,5000
0,5	5,791	0,6915
1,0	6,581	0,8413
1,5	7,372	0,9332
2,0	8,162	0,9773
2,5	8,953	0,9938
3,0	9,743	0,9987



Est. A

42a. Conferencia

NOCION DE MUESTRA

La investigación estadística que tiene por finalidad principal establecer las características de un universo, población o masa de personas u objetos, tropieza , la mayor parte de las veces, ante la imposibilidad física de poder tomar en cuenta todos los casos.

En tal emergencias se recurre a muestras de la población.

Se llama muestra a la técnica de seleccionar algunos casos del universo, de manera tal que obtengan estadísticas sobre la base de las cuales podamos, por inferencia, construir un panorama de la población derivado de aquellos casos.

La técnica del muestreo se usa porque algunas veces es el 1) único medio posible; a menudo 2) el más práctico y generalmente 3) el más eficiente para estudiar una población o universo de datos.

1) Es el único medio posible en el caso en que debamos realizar un análisis de materiales para examinar su calidad, porque esto puede implicar su destrucción. Por ejemplo, si se trata de constatar la resistencia de un lote de ladrillos, no se sometería a cada uno de ellos a la prueba, sino que se elegiría un número limitado con lo que se está obligado a usar el muestreo.

Los primeros problemas que se plantean son los de decidir cuantos ladrillos deben tomarse de cada lote para analizarlos y como deben ser seleccionados.

Realizada la prueba, se tabulan los resultados y se los analiza con el propósito de inducir y generalizar las características del total de ladrillos en base a lo que se ha visto en la muestra.

2) Es el único método práctico, por que por lo general el universo está compuesto por miles o millones de casos, y mediante la muestra es posible captar un conjunto de aquellos casos para su análisis

Por ejemplo, supongamos que nos proponemos establecer las condiciones del mercado de aceite para automotores. Debercmos preparar una encuesta dirigida a los conductores para saber entre otros detalles, la calidad del aceite utilizado, el motivo para usar tal calidad y el lu

gar donde realiza la compra del aceite. Es natural que no podremos dirigirnos a todos los conductores, por lo cual debemos recurrir a una muestra y hacerla con suficiente cuidado en forma que pueda ser considerada representativa de los hábitos de compra del total de los conductores.

3) Es generalmente el método más eficiente aunque desde un punto de vista físico y financiero se puede llevar a cabo un recuento total.

En efecto, si una muestra es realizada cuidadosamente proveerá de materiales suficientes para hacer generalizaciones respecto de las características de la población total de la que se obtuvo la muestra. Por lo tanto, la eficiencia y la economía recomiendan las técnicas del muestreo al investigador en el campo de los negocios y de la economía.

Poblaciones finitas y poblaciones infinitas:

En general hay dos clases de poblaciones en las cuales se trabaja en estadística: poblaciones finitas y poblaciones infinitas.

Por ejemplo, los estudiantes sin títulos actualmente matriculados en la Universidad de Buenos Aires constituyen una población finita.

Los cazadores de la Provincia de Buenos Aires, con licencia forman una población finita.

La sucesión de números de puntos obtenida arrojando un par de dados sucesivamente, es una población infinitamente grande.

El número de arandelas que se fabrican con un torno puede considerarse que pertenece a una población indefinidamente grande, ya que la población se engendra por la producción de una arandela detrás de la otra, y es fácil imaginar que no hay una última arandela.

Debe observarse que esta población de arandelas cambia realmente según la maquinaria empleada, el tipo de materia prima utilizada, el cambio de operaciones, etc.

Para un grupo particular de operarios la población puede permanecer razonablemente constante y una muestra de las arandelas, durante la actuación del grupo, puede considerarse como una muestra de una población infinitamente grande de arandelas.

En el caso de los puntos obtenidos arrojando un

par de dados sucesivamente, la población depende de varios factores tales como los propios dados, que pueden estar ligeramente cargados, el método de lanzarlos y la superficie sobre la que caen.

Si los dados fuesen en su construcción "perfectamente" homogéneos y se los agitara "perfectamente" antes de cada prueba y fueran arrojados sobre una superficie "perfecta", podríamos imaginar que tenemos una población "ideal"

Podemos usar la teoría de las probabilidades para predecir las características de esta población ideal

Cuando un cierto procedimiento de muestreo se aplica a una población finita, debe cuidarse de que el procedimiento usado no llegue a engendrar una población diferente de la población finita que se está estudiando.

Por ejemplo, si tomásemos uno de cada 20 nombres de la guía telefónica y marcásemos el número para pedir una información sobre la casa, tal procedimiento de muestreo puede servir para obtener referencias aceptables acerca de la población formada por casas con teléfono. En realidad existiría un cierto número de casas en las que no se obtendrá respuesta. Debemos considerar que, si tomamos la muestra de residencias en las que se ha obtenido contestación, nuestro procedimiento de muestreo no se refiere a la población de casas con teléfono sino a la población de casas con teléfonos que contestaron al llamado. Estas dos poblaciones de casas son realmente diferentes. Por ejemplo, la segunda población debe estar formada por familias más numerosas y de gente anciana y otros tipos de personas que permanecen en casa.

Por otra parte, si repetimos bastantes veces las llamadas telefónicas a las casas que no contestaron originariamente, hasta obtener respuesta, haríamos recien el muestreo de la población primera.

Est. A

43a. Conferencia

PARAMETROS Y ESTADISTICAS

¿Cuál es nuestra finalidad al obtener muestras de observaciones?

La razón principal no está simplemente en acumular un conjunto de números, sino en obtener algunas características de este conjunto, tales como su promedio, como varían de uno a otro, su varianza, etc. con el propósito de obtener inferencias acerca de la población. Lo primero que se ha de aprender en estadística es condensar los datos de las muestras y presentarlos satisfactoriamente.

El punto fundamental consiste en saber que clase de inferencias o conclusiones definitivas se pueden hacer a partir de la muestra para la población de la que aquella procede y que grado de confianza merecen tales inferencias.

Lo más sencillo que puede hacerse para condensar y describir muestras de datos cuantitativos es formar distribuciones de frecuencias y describirlas calculando determinados tipos de promedios.

Tales cantidades calculadas a partir de los datos de las muestras que permiten describirlas se llaman estadígrafos. De modo similar, las poblaciones se describen mediante números llamados parámetros de la población.

Sólo en raras ocasiones es posible conocer precisamente los valores de los parámetros de la población, simplemente porque, en contados casos disponemos de los datos para la población entera.

Frecuentemente sólo se tiene una muestra de la población.

Por lo tanto el problema habitual es calcular estadígrafos de las distribuciones de frecuencias de las muestras y a partir de ellos obtener los valores más verosímiles de los parámetros de la población.

En el caso de muestras muy grandes los valores de sus estadígrafos serán muy próximos a los de los co-

rrespondientes parámetros de la población.

Por ejemplo, la media aritmética de una muestra muy grande de mediciones "tomadas al azar" de una población será muy próxima a la media aritmética de la población total.

En el caso de pequeñas muestras las discrepancias entre estadígrafos de varias de ellas llegan a hacerse mayores y el problema de deducir los valores de parámetros de la población a partir de muestras es más complicado y ha de resolverse mediante la teoría de las probabilidades.

Ejemplos:

1) Si se le recomendase seleccionar una muestra de 100 estudiantes de la Universidad de Buenos Aires, que son hijos de ex-alumnos de la misma Universidad ¿cómo seleccionaría la muestra? ¿Cuál es la población muestreada?

Si esos 100 estudiantes han recibido por correo un cuestionario y sólamente 60 lo contestan y devuelven ¿qué población habrá sido muestreada

2) Suponga que Ud. fuese encargado de emprender un estudio del consumo de corriente eléctrica en las casas particulares de Buenos Aires durante diciembre de 1955 sobre la base de una muestra de 300 vivendas. ¿Cómo procedería Ud. a la selección de la muestra de direcciones para tal propósito? ¿Cuál es la población muestreada?

Est. A

44a. Conferencia

SELECCION DE LA MUESTRA

Se usan dos técnicas en la selección de los casos que formarán la muestra:

1) Muestra simple al azar
2) Muestra estratificada al azar

Muestra simple al azar:

Cuando la selección de los casos se ha hecho en forma de dar a cada uno de ellos, dentro del universo, la misma probabilidad de ser seleccionados, este método se llama "muestreo al azar"

Si bajo estas condiciones, se selecciona un número suficientemente amplio de casos, la muestra será una sección transversal, en miniatura del total de donde se seleccionarían los casos.

El método de muestreo al azar, exige que la selección se haya hecho en forma tal que cada caso en el universo tenga las mismas probabilidades de ser incluído en la muestra.

Para asegurarse una verdadera muestra al azar para un estudio estadístico, se debe proceder con mucho cuidado en la selección de las muestras obtenidas del universo.

Una forma elemental de muestra al azar puede ser usado cuando el detalle correspondiente a cada caso puede ser escrito en una tarjeta y ubicar estas en un fichero.

Si las tarjetas están bien mezcladas se puede hacer una selección al azar tomando una tarjeta del universo, escribiendo su valor; se repone la tarjeta en el fichero se mezclan nuevamente y vuelve a extraerse otra tarjeta y así sucesivamente hasta obtener el número de casos fijados para la muestra.

Un medio más conveniente para obtener una muestra al azar es enumerar cada caso y luego usar una tabla de números al azar como la de Tippett llamada "Números de muestra al azar"

Se seleccionan los casos incluyendo en la muestra aquellos cuyo número aparece en la tabla de números

al azar.

El proceso para obtener una muestra al azar de, por ejemplo, datos económicos y comerciales, muy pocas veces es simple, aún cuando los diseños al azar puedan jugar un importante rol en la solución de un problema complejo de muestreo.

Consideremos el problema que se le presenta a un investigador que desea analizar el mercado de jabón en Buenos Aires. El dato debe ser solicitado a las amas de casa y seleccionado en forma tal que constituya una muestra representativa.

Si se usa el sistema de muestra al azar debe dividir primero la ciudad en secciones y numerar cada sección, luego dividir estas en distritos y numerarlos, dividir los distritos en manzanas y numerarlas.

Si las áreas son aproximadamente iguales en tamaño, la selección al azar de los números localizará las manzanas de la ciudad adonde deben enviarse a los visitadores.

Los censistas deben interrogar a las amas de las casas seleccionadas al azar de cada una de las manzanas enumeradas.

Mediante este método de selección de una muestra simple al azar podrán conocerse las preferencias respecto a las condiciones del mercado de jabón al determinarse si se lo prefiere perfumado o no, blanco o de color, etc.

Est. A

45a. Conferencia

2) Muestra estratificada al azar

Es la técnica que se usa para aumentar la seguridad de los resultados de la muestra/

Cuando se emplea la muestra estratificada se usan características conocidas del universo para guiar la selección.

Por ejemplo, es muy probable que las opiniones de los habitantes de la ciudad difieran de las de los agricultores en un determinado aspecto. Debemos entonces conocer la proporción de población urbana y población rural para que teniéndola en cuenta en la muestra ésta resulte más representativa. Así si la población urbana es el 70% y la rural el 30% la muestra será mejor diseñada si se toman para la muestra elementos de población en la proporción indicada que si sólo se la diseña teniendo en cuenta los centros metropolitanos de fácil acceso.

En general, serán más representativas las muestras cuando se tengan en cuenta las características más importantes de la muestra y que sean comunes con el universo.

Al plantear un diseño para medir el estado de la opinión pública, por ejemplo, debe tenerse en cuenta que la gente joven difiere de las opiniones de los mayores; la opinión de personas del sexo masculino difiere de aquellas del sexo femenino; la de los habitantes del norte no siempre opinan como los del litoral; el rico puede tener diferentes opiniones que las personas de menores recursos.

Debido a la diversidad de características es necesario estratificar la población de maneras diferentes para hacer una muestra representativa.

Para asegurarse una buena muestra, sobre la opinión en materia política, cada uno de los grupos citados debe estar representado en la muestra en la misma proporción en que están presentes en la población considerada como un todo.

Debemos usar controles en la muestra estratificada. Por ejemplo, para asegurar una estratificación apropiada, las instrucciones al censista pueden dirigirlo a interrogar a 20 personas que ganan menos de $ 1.000 mensua-

les; 30 personas que ganan entre $1.000 y $3.000 y a 10 personas que ganan más de $3.000; o puede habérselo instruído para obtener datos de 10 productores agropecuarios, 15 obreros industriales y 5 empleados administrativos. Con estas indicaciones el censista no interroga a cualquiera sino que, controla cuidadosamente su selección, de manera que la proporción de cada clase cuyas opiniones pueden ser significativamente diferentes de las opiniones de otras clases sobre la pregunta formulada, será la misma en la muestra en proporción a lo que es con respecto al todo.

Es obvio que la estratificación que se haga de la población dependerá del tipo y propósito de la investigación. Si por ejemplo, intentamos medir la importancia que tiene una bolsa de harina en el costo de venta, la estratificación según la afiliación política parecería casi innecesaria, ya que no hay razón para pensar que la población se dividiría en lineamientos políticos sobre este punto. En cambio, dado que los ingresos pueden tener alguna relación con la cantidad de harina usada en los hogares, tendría importancia una clasificación por ingreso como base para la estratificación.

Una vez que se ha determinado la estratificación y se ha decidido sobre el número de casos que han de ser seleccionados dentro de cada estrato, dichos casos deberán ser elegidos al azar en cada estrato. De allí la denominación de "muestra estratificada al azar"

Se puede resumir en cuatro frases la muestra estratificada al azar:

1) El universo se divide en estratos de acuerdo a una o más características importantes.

2) El número de casos que ha de seleccionarse en cada estrato está determinado por la exigencia de que la representación proporcional de cada estrato en la muestra debe ser la misma que en el universo.

3) La muestra se toma al azar en cada estrato, en igual proporción a la establecida en el punto anterior.

4) Los diseños de cada estrato se combinan luego para obtener la muestra estratificada.

Los casos antes indicados ilustran suficientemente respecto a la forma en que la estratificación debe ser usada para asegurar la veracidad de los resulta-

dos de la muestra. También permiten apreciar que las estratificaciones pueden determinarse únicamente después de un análisis cuidadoso del propósito de la investigación y del examen de la relación entre las características del universo considerado y las características por las cuales se efectúa la estratificación o estratificaciones.

Naturaleza de los resultados del muestreo

Hemos considerado antes el caso de una encuesta dirigida a los conductores de vehículos tendientes. a determinar, entre otras cosas, la cantidad media gastada en la adquisición de combustible.

Supongamos que dos censistas están encargados de interrogar a 1.000 conductores seleccionados al azar, en el mismo día y en el mismo lugar. Sin duda, algunos dirán que compran 5 o 10 lts, a la vez, en cambio otros compran hasta llenar el tanque y habrá un gran número que estarán entre estos dos extremos. En otras palabras, existe variación en el universo considerado.

Es evidente entonces, que habrá también variación entre los valores medios obtenidos por los dos censistas al tomar cada uno una muestra al azar diseñada bajo condiciones uniformes respecto de una misma población. Aún cuando estén a veces muy cerca un valor medio del otro en general nunca serán coincidentes.

Podemos concluír diciendo que aún cuando el universo que se estudia no cambie; la técnica de muestreo empleado sea al azar y las muestras sean todas del mismo tamaño, cada muestra será en cierto grado diferente de toda otra muestra.

Las diferencias son atribuíbles a las circunstancias casuales que gobiernan la selección de los casos individuales que componen las muestras.

Tales variaciones en los resultados de las muestras son conocidos como "errores al azar del muestreo"

La probabilidad que tiene cualquier valor de aparecer en una muestra elegida al azar, está determinada por la frecuencia relativa con la cual el valor o peso aparece en el universo.

Si el universo contenido en una urna es de 100 bolillas, 50 blancas y 50 negras, la probabilidad de obtener bolilla blanca es igual a 0,5 siempre que las bolillas sean extraídas una por vez de la urna y bajo las condiciones de muestreo simple al azar previamente estable-

cido.

La probabilidad de extrer, en las mismas condicio nesm una bolilla negra es también de 0,5

Si se extraen dos bolillas a la vez, existe la pro babilidad de obtener dos bolillas blancas o dos bolillas ne gras a la vez, aún cuando la probabilidad de tal hecho rea- zado al azar, es menor que la probabilidad de obtener una bolilla blanca y una negra.

En este caso los resultados, al extraer 100 veces dos bolillas a la vez, serán aproximadamente de 25 extraccio nes con dos bolillas negras; 50 con una blanca y otra negra; y 25 con dos bolillas blancas.

Por supuesto, los resultados del muestreo serían mucho más fáciles de interpretar, si cada extracción fuera una miniatura del universo, en este caso una bolilla blanca y una negra.

Pero debemos recordar que la selección al azar tam bién nos puede dar dos bolillas blancas o dos negras.

Est.A

46a. Conferencia

El problema de la inferencia estadística.-

Una vez llevada a cabo la muestra, la tarea analítica consiste en construir un panorama del universo sobre la base dada por la muestra.

En el caso citado antes la extracción simultánea de dos bolillas cuando se hacen 100 extracciones sabiendo que $N(p + q)^n$ nos dará

$$25\ p^2 + 50\ pq + 25\ q^2$$

cuando se usa $100(\frac{1}{2} + \frac{1}{2})^2$.

Por tanto, podemos inferir de la distribución de 25, 50 y 25 que hay un número igual de bolillas blancas y negras en la urna de donde se han hecho las extraciones.

De ahí que aún en esta situación más simple, los resultados de la muestra no son descripciones claras, precisas ni definitivas del universo.

El panorama dado por la muestra esta oscurecido por las variaciones que son atribuibles a la selección al azar de los casos que se separan por la muestra.

Consideremos ahora un ejemplo de caracter económico.

Supongamos que hemos obtenido una muestra de 100 casos de un universo de 900 precios. El proposito de la investigación consistirá en mostrar el cambio típico en estos 900 precios según sus valores básicos anuales.

Imaginemos que estos valores están distribuidos en la siguiente forma:

50	casos	constituyen	el	85%	de	sus	precios básicos
140	"	"	"	95%	"	"	"
300	"	"	"	105%	"	"	"
200	"	"	"	115%	"	"	"
160	"	"	"	125%	"	"	"
50	"	"	"	135%	"	"	"
900							

Si se realiza una muestra al azar de 100 casos de aquel universo de 900, el resultado probable será:

6	casos	al	precio	de	85%	de base
16	"	"	"	"	95%	" "
33	"	"	"	"	105%	" "

22 casos al precio de 115% de base
17 " " " " 125% " "
6 " " " " 35% " "
100

Cada valor en esta distribución de la muestra de 100 casos, tenderá a tener la misma frecuencia relativa de aparición que tiene en el universo.

En los dos ejemplos considerados las distribuciones obtenidas constituyen el resultado más probable. Pero hemos visto que no se obtiene uniformemente este resultado más probable. que se define como el valor hacia el cual tienden todas las muestras al azar y, como veremos, es el valor al al que nos aproximamos a medida que la muestra aumenta de tamaño.

Al examinar los cambios entre 1926 y 1938 de 789 precios de productos de consumo, tomando dos muestras simples al azar de 100 casos cada una, se obtuvieron los siguientes resultados.

Base 1938 % de los precios de 1926	Número de casos con el porcentaje indicado de cambio desde 1926	
	Muestra I	Muestra II
25	5	4
45	9	10
65	29	23
85	33	36
105	20	23
125	4	4
	100	100

Como resultado de la selección al azar, aparecen en la la. muestra, 5 casos cuyos precios eran, en 1938, el 25% del nivel de 1926.

En la 2a. muestra se presentaron 4 de tales casos, La diferencia que aparece de grupo a grupo, es debida enteramente a los errores al azar del muestreo, puesto que la técnica del muestreo y el universo son constantes. Es casi improbable que cada una de estas muestras provea una miniatura precisa del universo, aún cuando da una aproximación bruta de él.

Veremos que existen métodos para limitar aquella determinación aproximada.

En toda muestra habrá siempre un elemento de duda debido a los errores al azar del muestreo, aún cuando este

área de incertidumbre sea muy pequeña cuanto más grande es la muestra.

Es de estas muestras particulares como las citadas que los estadígrafos deben,por inferencia,construir un panorama del universo del cual se han extraido las muestras.

Las situaciones causales conducen a una distribución normal.

Si en vez de las dos muestras consideradas antes hubiésemos extraido un mayor número de tales muestras,cada una de ellas nos daría una media aritmética diferente.

La distribución que se obtiene ordenando estos valores medios,presenta generalmente la forma de la curva "normal",debido a que las fuerzas que operan en la determinación de la selección de los casos individuales de cada muestra, son muy numerosos y de igual valor;en otras palabras,está afectada la selección por una multiplicidad de causas de igual importancia sin que domine el resultado una causa única o un grupo de causas.

En el caso de contar con dos o más muestras de igual dimensión,correspondientes a un universo común,el valor medio de las medias aritméticas se considera como el valor medio "verdadero" del universo.

Peligro del bias (vicio) en los resultados del muestreo.

Cuando solo se hallan presentes en una investigación,los errores debidos al azar,el valor medio de la muestra tiene la misma probabilidad de ser mayor que el valor verdadero,como de ser menor que él y las variaciones extremas son raras. Cuando se promedian tales valores medios,los errores pueden compensarse debido a las variaciones que los signos contrarios tienden a anular. Por esta razón la media de las medias es la mejor forma de estimar el verdadero valor.

Los errores debidos al bias que surgen de cualquier otra fuente,tiene na única dirección. No puede esperarse que se anulen,sino,que por el contrario determinará que la extimación del universo,varie más o menos según sea la dirección del bias.

Pueden admitirse errores al azar del muestreo,pero el bias,puede y debe evitarse en todo trabajo estadístico.

Por ejemplo,de un registro de estudiantes,se puede obtener una muestra,libre de bias,seleccionando un nombre de cada 10 que aparecen en el Registro. También se podría utilizar el nombre que aparece en la primera linea de cada página y se obtendrían resultados al azar similares.En cambio si se eligiera el nombre de estudiantes que empezará con una letra determinada,se estaría introduciendo un principio de selección ya que muchos nombres de estudiantes chinos empiezan con K.

Los nombres de una guia telefonica no constituyen una lista completa de todas las familias de la comunidad,ya que más de la mitad de las familias argentinas no tienen teléfono.

Aún cuando se lleve a cabo una enumeración completa en vez de una muestra el bias puede aparecer motivado por una investigación pobremente conducida. Por ejemplo,se sabe que existe una tendencia por parte de los individuos a aumentar sus ingresos exagerándolos,a mentir sobre la posesión de implementos eléctricos caseros,etc.

El bias es conocido también con el nombre de errores sistemáticos,provocados por defectos visuales del operador,instrumental falto de precisión,etc.

Equivocaciones.

La palabra "equivocaciones" puede usarse en su acepción general. Son equivocaciones por ejemplo los errores que se cometen al tabularse las cifras,al transcribirse los números,al totalizar una suma de casos.

La garantía de que no se han cometido equivocaciones,se realiza por medio del control y del cómputo. El tiempo que se pierde en el control,es un tiempo bien empleado, pues garantiza la seguridad del trabajo.

Por tanto debemos distinguir tres tiposde errores: los errores al azar del muestreo,el bias y las equivocaciones,que pueden aparecer en un trabajo extadístico.

Los dos últimos: el bias y las equivocaciones pueden ser evitados.Los errores al azar,por otro lado,son inherentes a la muestra;se suponen que han de producirse,debiendo ser considerados cuando se interpretan los resultados de las técnicas del muestreo.

Est.A

47a. Conferencia

Hemos dicho en clases anteriores que pueden estudiarse las fluctuaciones de las fluctuaciones de los estadígrafos (o estadísticas) de las muestras correspondientes a una población desde el punto de vista experimental y tambien desde el matemático, estando este último estudio fundado en la teoría de las probabilidades.

Dijimos tambien que podíamos contar con poblaciones finitas o infinitas.

Vamos pues a considerar ahora para cada clase de población la forma de aplicar el método experimental y el método matemático para el estudio de las fluctuaciones de sus estadísticas.

Muestreo de una población finita

Mediante un sencillo ejemplo aclararemos lo que se entiende por 1) muestreo experimental y 2) muestreo matemático de una población finita.

1) Muestreo experimental

Supongamos tener 6 bolillas numeradas del 1 al 6 en una urna; la composición de este grupo queda descripta en la siguiente tabla

x	y	$\frac{y}{N}$
1	1	1/6
2	1	1/6
3	1	1/6
4	1	1/6
5	1	1/6
6	1	1/6
	6	1

Extrayendo 3 bolillas simultáneamente o una despues de otra sin volver a reponerlas, hemos extraído experimentalmente una muestra de 3 bolillas de la población finita de 6.

Si introducimos de nuevo las 3 bolillas en la urna, podemos repetir el experimento cuantas veces nos plazca.

Cada vez que repetimos este proceso obtenemos una muestra experimental de 3 bolillas.

Supongamos ahora que estamos interesados en conocer la $\bar{X}$ de los números de las 3 bolillas de una muestra.

Recordando que al estar numeradas las bolillas pueden presentarse al extraer 3 bolillas valores que oscilan entre 6 y 15, si repetimos 100 veces el experimento, tendremos 100 medias aritméticas cuyas frecuencias fueron:

S(X)	$\bar{X}$	Y	$f(\bar{x})$	$\bar{X}$ Y	$\bar{X} - \bar{X} = x$	x^2	x^2Y
6	2	6	0,06	12,	[illegible]6	2,13	12,78
7	2,33	6	0,06	13,98	[illegible]3	1,28	7,68
8	2,67	8	0,08	21,36	- 0,79	0,62	4,96
9	3	17	0,17	51,	-0,46	0,21	3,57
10	3,33	18	0,18	59,94	- 0,13	0,02	0,36
11	3,67	12	0,12	44,04	0,21	0,04	0,48
12	4	14	0,14	56.	0,54	0,29	4,06
13	4,33	8	0,08	34,64	0,87	0,75	6,00
14	4,67	5	0,05	23,35	1,21	1,46	7,30
15	5	6	0,06	30.	1,54	2,37	14,22
		100		346,31		9,17	61,41

cuyos valores están comprendidos entre 2 y 5.

En efecto si en una extracción obtuviésemos como suma de las 3 bolillas 6 sería:

x	y	xy
1	1	1
2	1	2
3	1	3
4	0	0
5	0	0
6	0	0
	3	6

$$\bar{X}_6 = \frac{6}{3} = 2$$

Si la suma fuese 7 podría provenir del siguiente caso:

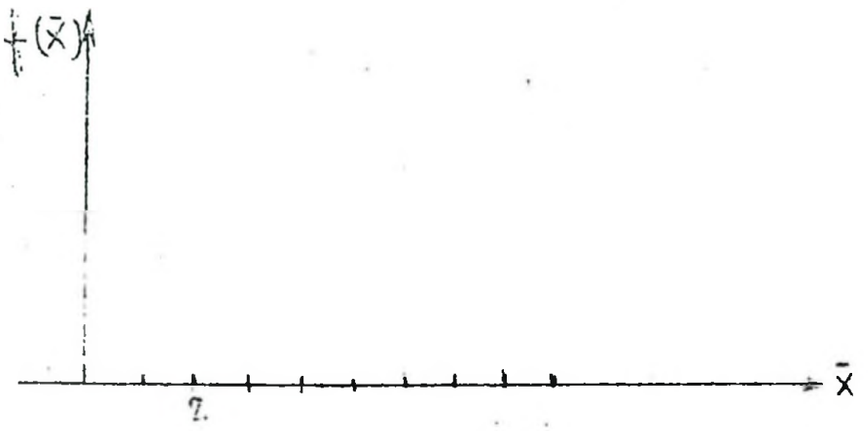
x	y	xy
1	1	1
2	1	2
3	0	0
4	1	4
5	0	0
6	0	0
	3	7

$$\bar{X}_7 = \frac{7}{3} = 2,33$$

y así sucesivamente, habiendo en este ejemplo una relación constante entre X y $\bar{X}$ dada por

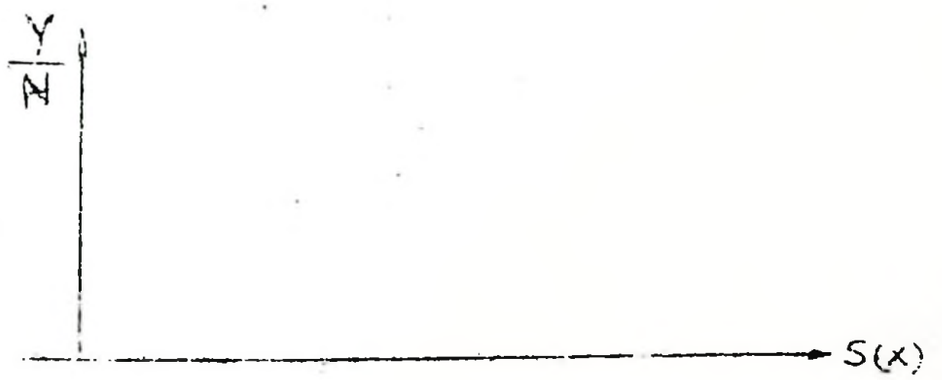
$$3\bar{X} = X$$

Las columnas 2a. y 4a. del cuadro ponen de manifiesto la distribución experimental de muestreo de las $\bar{X}$ de una serie de 100 muestras experimentales de 3 bolillas en una población de 6.



Análogamente las columnas 1a. y 4a. constituyen la distribución experimental de las sumas que pueden presentar las 3 bolillas al extraerse.

Su representación gráfica es:



Podríamos considerar otras estadísticas distintas de la $\bar{X}$ o $S(x)$ calculados de las muestras sucesivas, como la desviación típica, el rango, el valor máximo, la mediana, etc. Todas estas estadísticas tienen sus correspondientes distribuciones experimentales para las 100 muestras.

Si fueran extraídas otras 100 muestras obtendríamos una distribución de frecuencias no muy diferente para los valores de $\bar{X}$ y de $S(x)$.

Si fuesen extraídas 1000 muestras encontraríamos que las frecuencias relativas serían más estables de una serie de 1000 muestras a otra, que de una serie de 100 muestras a otra.

Pero la distribución de las $\bar{X}$ en 100 muestras da ya una buena idea de como varían las $\bar{X}$ en el ejemplo considerado.

Simbolizaremos a la media aritmética de las medias aritméticas por $\bar{\bar{X}}$ y a la desviación típica $S_{\bar{X}}$.

Calculando los valores hallamos

$$\bar{\bar{X}} = \frac{346,31}{100} = 3,46 \qquad S_{\bar{X}} = \sqrt{\frac{61.41}{99}} = \sqrt{0,6203} = 0,79$$

La idea del muestreo experimental de una población finita, tal como la obtenida puede ser extendida a muestras de "n" elementos de una población finita de "N" elementos.

Est.A

48 a. Conferencia

Muestreo teórica de una población finita.

Volvamos a nuestra población finita de 6 bolillas marcadas de 1 a 6.

El número de muestras de 3 bolillas que es posible extraer de esta población de 6 bolillas es simplemente el número de combinaciones de 6 objetos tomados de 3 en 3 es decir:

$$C_6^3 = \frac{6!}{3!\,(6-3)!} = \frac{6.5.4.3.2}{1.2.3.1.2.3} = 20$$

Esas 20 muestras son las siguientes:

1.2.3	2.3.4	3.4.5	4.5.6
1.2.4	2.3.5	3.4.6	
1.2.5	2.3.6	3.5.6	
1.2.6	2.4.5		
1.3.4	2.4.6		
1.3.5	2.5.6		
1.3.6			
1.4.5			
1.4.6			
1.5.6			

La distribución de frecuencias de la suma S(X) y de las $\bar{X}$ en esta serie de muestras posibles está dada en la siguiente tabla:

S(X)	$\bar{X}$	Y	y/N
6	2	1	0.05
7	2.33	1	0.05
8	2.67	2	0.10
9	3	3	0.15
10	3.33	3	0.15
11	3.67	3	0.15
12	4	3	0.15
13	4.33	2	0.10
14	4.67	1	0.05
15	5	1	0.05
		20	1

Las columnas 2a. y 4a. ponen de manifiesto la distribución teórica de muestreo de la $\bar{X}$ de las muestras de 3 bolillas en una población finita de 6 bolillas.

Análogamente las columnas 1a. y 4a. constituyen la distribución teórica en el muestreo de la suma S(X).

La distribución muestral teorica dada en la tabla anterior no es una que una distribución de probabilidad.

Es una predicción de la distribución que se obtendría en un muestreo experimental, extrayendo un número muy grande de muestras de 3 bolillas.

La precisión de la predicción para 100 muestras experimentales puede verse por compración de los valores de las columnas cuartas de las dos tablas obtenidas. Si las muestras experimentales hubieran sido 1000 en vez de 100, los valores hubieran sido más cercanos.

Ahora podemos obtener la media $\mu_{\bar{x}}$ y la desviación tipica $\sigma_{\bar{x}}$ de la distribución muestral teónica de $\bar{X}$ de la segunda tabla.

Sus valores son:

$$\mu_{\bar{x}} = 3.5 \qquad \sigma_{\bar{x}} = \sqrt{\frac{7}{12}} = 0.7638$$

Estos valores teoricos calculados predicen con bastante aproximación los valores experimentales antes hallados.

$$\bar{X} = 3.46 \qquad S_{\bar{x}} = 0.79$$

También podemos deducir en el caso teorico la media $\mu_{s(x)}$ y la desviación tipica $\sigma_{s(x)}$ de las sumas muestrales en las 20 muestras posibles de la población, a saber:

$$\mu_{s(x)} = 10,5 \qquad \sigma_{s(x)} = \sqrt{\frac{63}{12}}$$

que pueden calcularse, naturalmente directamente de $\mu_{\bar{x}}$ y $\sigma_{\bar{x}}$ pues al ser : $s(X) = 3\bar{X}$

resulta $3\mu_{\bar{x}} = \mu_{S(X)}$

$3\sigma_{\bar{x}} = \sigma_{S(X)}$

Si consideramos muestras de una bolilla en una población de 6 encontraríamos 6 muestras posibles.

La distribución muestral teorica, en este caso viene dada por las columnas la. y 3a. del cuadro de la página 208; es igual a la distribución de valores de X en la misma población.

Esta distribución tiene su propia media μ y su desviación tipica σ cuyos valores son:

$$\mu = 3,5 \qquad \sigma = \sqrt{\frac{35}{12}}$$

En otra palabra estos resultados nos dan los valores de la media y de la desviación tipica de la población de bolillas.

Mientras que los valores:

$$\mu_{\bar{x}} = 3.5 \qquad \sigma_{\bar{x}} = \sqrt{\frac{7}{12}} = 0.7638$$

son los de la media y de la desviación tipica de la distribución de las medias de todas las muestras posibles de 3 bolillas de esta misma población.

49a.Conferencia

Media y desviación típica de las medias de todas las muestras posibles de una población finita.

Nos preguntamos cuales son las relaciones entre:

$\mu_{\bar{x}}$ y μ

y entre $\sigma_{\bar{x}}$ y σ

para el muestreo teórico general de poblaciones finitas

Supongamos tener una población finita de N elementos numerados de modo que se tenga la distribución de esos números y que le corresponde por media el valor μ y por desviación típica σ .

Consideremos ahora todas las muestras posibles de n elementos para esta población.

Hay C_N^n de estas muestras.

Por otra parte sean X_1 , X_2 , X_3 , ... X_N Los números correspondientes a los N elementos de la población y

$X_1\ X_2 \cdots X_{n-1}\ X_n$

$X_1\ X_2 \cdots X_{n-1}\ X_{n+1}$

.....

$X_{N-n+1}\ X_{N-n+2} \cdots X_{N-1}\ X_N$

las C_N^n muestras.

La media de las medias de todas las muestras anteriores la simbolizamos por $\mu_{\bar{x}}$.

La probabilidad para cada muestra es $\frac{1}{C_N^n}$, resulta que siendo $\mu_{\bar{x}}$ la media de todas las medias muestrales

$$\sigma_N^n \mu_{\bar{x}}$$

será la suma de todas las medias muestrales, o sea

$$C_N^n \mu_{\bar{x}} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_N\right) + \frac{1}{n}\left(X_1 + X_2 \cdots + X_{N+1} + X_{N+2}\right) \cdots$$

$$+ \frac{1}{n}\left(X_{N-n+1} + X_{N-n+2} + \cdots + X_N\right) \qquad (1)$$

Ahora bien cada una de las X de la población aparecerá en C_{N-1}^{n-1} de las muestras, pues si se elige una X par-

ticular,una muestra de n elementos que contenga x, ha de estar formada eligiendo n-1 de las restantes x.

Como el número de formas en que pueden hacerse estas selecciones es

$$C_{N-1}^{n-1}$$

puede escribirse (1) según se elija X_1 , X_2 , X_3 ... del siguiente modo:

$$C_N^n \mu_{\bar{x}} = \frac{1}{n} C_{N-1}^{n-1} X_1 + \frac{1}{n} C_{N-1}^{n-1} X_2 + \ldots + \frac{1}{n} C_{N-1}^{n-1} X_N \qquad (2)$$

dividiendo ambos miembrso por C_N^n y teniendo en cuenta que:

$$\frac{C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N}$$

resulta en (2)

$$\mu_{\bar{x}} = \frac{1}{N}\left[x_1 + x_2 + \ldots + N\right] \qquad (3)$$

$$\boxed{\mu_{\bar{x}} = \mu} \qquad (4)$$

Es decir que la media de las medias de todas las muestras posibles de n elementos,es igual a la media de la población.

Para obtener el valor de la varianza $\sigma_{\bar{x}}^2$ partimos de la definición

$$\sigma_{\bar{x}_i}^2 = \frac{\sum_{i=1}^{N} \left[\bar{x}_i - \mu_{\bar{x}}\right]^2 y_i}{\sum_{i\ 1}^{N} y_i}$$

que desarrollando da

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{N} \bar{x}_i^2 y_i}{\sum_{i=1}^{N} Y_i} - \frac{2\mu_{\bar{x}} \sum_{i=1}^{N} \bar{X}_i Y_i}{\sum_{i=1}^{N} Y_i} + \frac{\mu^2_{\bar{x}} \sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} Y_i}$$

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{N} \bar{X}_i^2 Y_i}{\sum_{i=1}^{N} y_i} - \mu^2_{\bar{x}}$$

expresión que podemos escribir

$$\sigma^2_{\bar{x}} = \left[\frac{1}{n}(X_1+X_2+\dots X_N)\right]^2 \frac{1}{C_N^n} + \left[\frac{1}{n}(X_1+X_2+\dots+X_{N-1}+X_N\ 1)\right]^2 \frac{1}{C_N^n} +$$

$$+\dots+\left[\frac{1}{n}(X_{N-n+1}+X_{N-n+2}+\dots+X_N\right]^2 \frac{1}{C_N^n} - \mu^2_{\bar{x}}$$

Al desarrollar los cuadrados y teniendo en cuenta, segun dijimos que el número de veces que aparece cada cuadrado de x es C_{N-1}^{n-1} y el número de veces que aparece cada producto binario de las x es C_{N-2}^{n-2}, se puede escribir

$$\sigma^2_{\bar{x}} = \frac{1}{n^2}\frac{1}{C_N^n} C_{N-1}^{n-1}\left[X_1^2+X_2^2+\dots+X_N^2\right] +$$

$$+\frac{2}{n^2}\frac{1}{C_N^n} C_{N-2}^{n-2}\left[X_1X_2+X_1X_3+\dots+X_{N-1}\ X_N\right] - \mu^2_{\bar{x}} \quad (5)$$

Elevando al cuadrado ambos miembros de (3) se tiene:

$$\mu^2_{\bar{x}} = \frac{1}{N^2}\left[X_1+X_2+\dots+X_N\right]^2$$

$$= \frac{1}{N^2}\left[X_1^2+X_2^2+\dots+X_N^2\right]+\frac{2}{N^2}\left[X_1X_2+X_1X_3+\dots+X_{N-1}\ X_N\right]$$

sustituyendo este valor en (5) y recordando que:

$$\frac{1}{N^2}\frac{1}{C_N^n}\cdot C_{N-1}^{n-1} = \frac{\frac{(N-1)!}{(n-1)!\ (N-n)!}}{n^2\frac{N!}{n!\ (N-n)!}} = \frac{1}{nN}$$

y que

$$\frac{2}{n^2}\frac{1}{C_N^n}\cdot C_{N-2}^{n-2} = \frac{2\frac{(N-2)!}{(n-2)!\ (N-n)!}}{n^2\frac{N!}{n!\ (N-n)!}} = \frac{2(n-1)}{nN(N-1)}$$

se tiene que (5) resulta:

$$\sigma^2_{\bar{x}} = \left(\frac{1}{nN}-\frac{1}{N^2}\right)\left[X_1\ X_2+\dots+X_N\right]^2$$

$$+2\left[\frac{(n-1)}{nN(N-1)}-\frac{1}{N^2}\right]\left[X_1\ X_2+X_1X_3+\dots+X_{N-1}X_N\right]$$

$$\sigma^2_{\bar{x}} = \left\{\left[\frac{1}{N}-\frac{1}{N^2}\right]\left[X_1^2+X_2^2+\dots+X_N^2\right]\cdot\right.$$

$$\left.-\frac{2}{N^2}\left[X_1X_2+X_1X_3+\dots+X_{N-1}X_N\right]\right\}\frac{N-n}{n(N-1)} \quad (6)$$

La eepresión entre llaves puede escribirse:

$$\frac{1}{N}(X_1^2 + X_2^2 + \ldots + X_N^2) - \left[\frac{X_1 + X_2 + \ldots + X_N}{N}\right]^2$$

que es precisamente el valor de σ^2 varianza de una población finita.

Recordemos que al definir la varianza de una población finita dividimos por N y no por N-1; en vez en el caso de la varianza de una muestra de n casos, de una población, recuérdese que dividimos por n-1 no por n.

Finalmente reemplazando en (6) el valor indicado resulta:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad (7)$$

de donde extrayendo la raíz cuadrada se tiene el valor de $\sigma_{\bar{x}}$ buscado:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (8)$$

debemos observar que si el número N es muy grande, comparado con el número de elementos en la muestra, entonces el factor $\sqrt{\frac{N-n}{N-1}}$ es aproximadamente igual a 1 y el valor de $\sigma_{\bar{x}}$ se reduce a:

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}} \qquad (9)$$

Sin demostración comprobamos que la distribución muestral teórica de las sumas muestrales S(X); tiene las siguientes media y desviación típica:

$$\mu_{s(x)} = n\mu \qquad (10) \qquad \text{y por (4)} \quad \mu_{s(x)} = n\mu_{\bar{x}}$$

$$\sigma_{s(x)} = \sigma\sqrt{n}\sqrt{\frac{N-n}{N-1}} \qquad (11)$$

si N fuese muy grande en comparación con n se tiene que:

$$\sigma_{s(x)} = \sigma\sqrt{n} \qquad (12)$$

resultando al cuadrar ambos miembros

$$\sigma_{s(x)}^2 = n\sigma^2 \qquad (13)$$

que la varianza de las sumas muestrales S(X) es igual a n veces la varianza de la población.

Ejemplo:

Supongamos que se barajan las cartas de un mazo de 40. De ellas 4 están marcadas con 1, otras 4 con 2 y así sucesivamente hasta las últimas 4 con el número 10. ¿Cual es la media y la desviación típica de la distribución muestral teórica de las medias de todas las muestras posibles de 10 cartas para esta población de 40?.

X	Y	f(x)
1	4	0,1
2	4	0,1
3	4	0,1
4	4	0,1
5	4	0,1
6	4	0,1
7	4	0,1
8	4	0,1
9	4	0,1
10	4	0,1
	40	1

Resulta operando que:

$\mu = 5,5 \qquad \sigma = \sqrt{33}$

N= 40
n= 10

Por tanto:

$\mu_{\bar{x}} = \mu = 5,5$

$\sigma_{\bar{x}} = \frac{\sqrt{33}}{\sqrt{10}} \sqrt{\frac{40-10}{40-1}} = \sqrt{\frac{33}{13}}$

y

$\mu_{s(x)} = n \, , \, \bar{x} = 10 \times 5,5 = 55$

$\sigma_{s(x)} = \sqrt{10} \sqrt{\frac{33}{13}}$

Aproximación de la distribución de medias muestrales a la distribución normal

Hemos dicho que hay C_N^n muestras posibles de n elementos en una población de N elementos.

Comenzando con una población bastante numerosa y aun considerando muestras de extensión moderada, sería muy laborioso calcular una tabla semejante a la de la pág. 208 que nos mostrara las probabilidades en la distribución de las medias muestrales.

Pero podemos calcular esas probabilidades aproximadamente, ya que con alguna condición previa, esta distribución se ajusta a una distribución normal.

Veamos ahora como una distribución normal puede ser ajustada a una distribución de medias de muestras.

Para llevar a cabo este proceso de ajuste, siendo la función teórica:

$$F_N(\bar{X}) = \frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}} \int_{-\infty}^{\bar{x}} e^{-\frac{2(x-\mu_{\bar{x}})^2}{2\sigma_{\bar{x}}}} dx$$

necesitaríamos sólo los valores de:

$$\mu_{\bar{x}} \qquad \text{y} \qquad \sigma_{\bar{x}}$$

y la tabla que contiene los valores de $F_N(X)$ para los sucesivos valores de Z.

Ejemplo:

El peso de 1000 estudiantes, tiene por media 148,2 libras y una desviación típica de 5,4 libras. Si se toma una muestra al azar, de 100 estudiantes ¿cuál es la probabilidad de que el peso medio de esos 100 exceda de 149 libras.

$N = 1000$
$n = 100$
$\mu = 148{,}2$
$\sigma = 5{,}4$

Entonces:

$$\mu_{\bar{x}} = 148{,}2 \qquad \sigma_{\bar{x}} = \frac{5{,}4}{\sqrt{100}}\sqrt{\frac{900}{999}} = 0{,}513$$

La relación entre Z y $\bar{X}$ es en este caso

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X} - 148,2}{0,513}$$

La cuestión queda reducida a saber cuál es la probabilidad de que $\bar{X}$ exceda a 149 y esto sucedería si Z excede a:

$$\frac{149 - 148,2}{0,513} = 1,56$$

Podemos entonces escribir:

$$P(\bar{X} > 149) = P\left(\frac{X - 148,2}{0,513} > \frac{149 - 148,2}{0,513}\right) =$$

$$= P(Z > 1,56)$$

Pero

$$P(Z > 1,56) = 1 - P(Z < 1,56)$$
$$= 1 - 0,9406 = 0,059$$

Esto significa que aproximadamente el 5,9% de las muestras de 100 individuos que se pueden extraer de una población de 1000 estudiantes, tienen una media superior a 149.

Ejemplo 2

Hallar en el mismo ejemplo las probabilidades de todas las muestras posibles en la distribución de las sumas S(X).

En este caso la relación entre Z y S(X) es:

$$Z = \frac{S(X) - \mu_{s(x)}}{\sigma_{s(x)}}$$

y los valores de $\mu_{s(x)}$ y $\sigma_{s(x)}$ se obtienen en función de:

$$N,\ n,\ \mu\ \text{y}\ \sigma$$

conocidas.

$$\mu_{s(x)} = 14820$$

$$\sigma_{s(x)} = 5,4\sqrt{100}\ \sqrt{\frac{900}{999}} = 51,3$$

y la relación buscada entre Z y S(X) sería:

$$Z = \frac{S(X) - 14820}{51,3}$$

y con ella podría resolverse el problema de determinar la probabilidad aproximada de que el peso total de 100 estudiantes exceda de 14900 libras.

Es el valor de Z para $S(X) = 14900$

$$\frac{14900 - 14820}{51,3} = \frac{80}{51,3} = 1,56 = P(\bar{X} > 149)$$

luego:

$$P(S(X) > 14900) = P(Z > 1,56) \pm 0,059$$

En otras palabras, la probabilidad de que el peso medio de 100 estudiantes exceda de 149 libras, es exactamente la misma que la probabilidad de que el peso total de los 100 estudiantes supere a 14.900 libras.

Est. A

51a. Conferencia

Muestreo de una población infinita

Como en el caso de una población finita, puede estudiarse el problema del muestreo en una población infinita experimental o matemáticamente.

Muestreo experimental

Para una población infinita no difiere esencialmente del método seguido cuando la población es finita. Consiste en llevar a cabo las operaciones experimentales.

Por ejemplo, el lanzamiento de un dado 10 veces sucesivas puede ser considerado como una muestra de 10 elementos de una población infinitamente grande de las posibles tiradas del dado.

Podemos repetir la operación tantas veces como se desee.

Si nos detuviéramos en 100 muestras, por ejemplo, podríamos formar una distribución de frecuencias de las $\bar{X}$ y de las $S(x)$.

No haremos aquí el cálculo, pero ha de servirnos de base en el estudio del muestreo teórico de una población infinita.

Conceptualmente podemos considerar que tal muestreo es equivalente a la extracción de 10 bolillas de una urna que las contiene en número infinito y cada una de las cuales está marcada con alguno de los números 1, 2, 3, 4, 5, 6.

Si la urna habría de corresponder a un dado correcto, la proporción de bolillas sería 1/6.

Podemos tambien pensar que colocamos 6 bolillas marcadas 1, 2, 3, 4, 5, 6 en la urna, y que extraemos una bolilla reponiendola y repitiendo este proceso 10 veces.

Así, pues 10 extracciones con reposición, equivalen a extraer una muestra de 10 bolillas de una población infinita.

Media y desviación típica de las distribuciones teóricas de medias y sumas de muestras de una población infinita.

Los resultados obtenidos en el caso de una población finita para la media y la desviación típica estaban dadas por las fórmulas

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

y se observó que para n grande y N mucho mayor se aproximaba la distribución de las medias muestrales a la distribución normal.

Para N infinitamente grande resulta que:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (14)$$

en las que μ y σ son la media y la desviación típica de una población infinita, de la cual fueron extraídas las muestras.

Análogamente cuando N es infinitamente grande las fórmulas (10) y (11) resultan:

$$\mu_{s(x)} = n\mu \qquad (15)$$

$$\sigma_{s(x)} = \sigma\sqrt{n} \qquad (16)$$

Recordemos que una distribución infinitamente grande era simplemente una distribución de probabilidad.

Al extraer los sucesivos elementos de una muestra de una población finita, las probabilidades cambian, a medida que se van sacando más elementos de la población, por eso consideramos el muestreo de una población finita como un problema de combinatoria.

En una población infinita, al extraer sucesivamente los elementos de una muestra, la probabilidad de que una variable aleatoria x tenga un valor particular en una extracción, no viene afectada por el valor que haya podido tener x en otra extracción.

En otras palabras, los resultados de las sucesivas extracciones en una población infinitamente grande, son independientes uno de otro.

Las fórmulas (14) (15) (16) las obtuvimos haciendo N infinitamente grande en expresiones obtenidas para poblaciones finitas, pero tambien podemos deducirlas directamente.

Vamos a ilustrarlo para el caso de muestras de dos elementos en una población infinita a la que corresponde la variable aleatoria discreta.

x	f(x)
x_1	$f(x_1)$
x_2	$f(x_2)$
x_3	$f(x_3)$
°	°
°	°
°	°
x_k	$f(x_k)$

La probabilidad de que la variable aleatoria X tenga un valor x_a en la primera extracción es $f(x_a)$ y de que tenga el valor x_b en la segunda extracción es $f(x_b)$.

Siendo las extracciones independientes, si se extrae una muestra de una población infinita, la probabilidad de que X tenga los valores x_a y x_b en las dos extracciones sucesivas, es por la ley de las probabilidades compuestas $f(x_a)\ f(x_b)$.

La media de esta muestra de dos elementos es:

$$\frac{x_a + x_b}{2} \qquad (17)$$

El valor medio de la media muestral se obtiene multiplicando la fórmula (17) por la probabilidad $f(x_a)\ f(x_b)$ y sumando para todos los valores posibles de a y b se tiene

$$\mu_{\bar{x}} = \sum_{a=1}^{k} \sum_{b=1}^{k} \frac{x_a + x_b}{2} f(x_a)\ f(x_b)$$

la cual puede ser descompuesta en las dos sumas:

$$\mu_{\bar{x}} = \frac{1}{2} \sum_{a=1}^{k} \sum_{b=1}^{k} x_a\ f(x_a)\ f(x_b)$$

$$+ \frac{1}{2} \sum_{a=1}^{k} \sum_{b=1}^{k} x_b\ f(x_a)\ f(x_b)$$

$$= \frac{1}{2} \left[\sum_{a=1}^{k} x_a\ f(x_a) \right] \left[\sum_{b=1}^{k} f(x_b) \right]$$

$$+ \frac{1}{2} \left[\sum_{a=1}^{k} x_b\ f(x_b) \right] \left[\sum_{a=1}^{k} f(x_a) \right]$$

siendo

$$\sum_{b=1}^{k} f(x_b) = \sum_{a=1}^{k} f(x_a) = 1$$

y

$$\sum_{b=1}^{k} x_a\ f(x_a) = \sum_{b=1}^{k} x_b\ f(x_b) = \sum_{i=1}^{k} x_i\ f(x_i) = \mu$$

resulta que:

$$\mu_{\bar{x}} = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

$$\boxed{\mu_{\bar{x}} = \mu} \qquad (18)$$

Siguiendo igual procedimiento hallaríamos para muestras de n elementos

$$\mu_{\bar{x}} = \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu \quad (n \text{ términos})$$

o sea

$$\mu_{\bar{x}} = \mu$$

Es decir: la media de la distribución de medias de un número infinitamente grande de muestras de una población infinita es igual a la media aritmética de la población.

La varianza de $\bar{X}$ para muestras de dos elementos es por definición

$$\sigma^2_{\bar{x}} = \sum_{b=1}^{k}\sum_{a=1}^{k}\left(\frac{x_a + x_b}{2}\right)^2 f(x_a)\, f(x_b) - \mu^2_{\bar{x}}$$

Desarrollando el cuadrado y sumando los términos semejantes

$$\sigma^2_{\bar{x}} = \frac{1}{4}\sum_{a=1}^{k} x_a^2\, f(x_a) + \frac{2}{4}\left[\sum_{a=1}^{k} x_a\, f(x_a)\right]\left[\sum_{b=1}^{k} x_b\, f(x_b)\right]$$

$$+ \frac{1}{4}\sum_{b=1}^{k} x_a^2\, f(x_b) - \mu^2_{\bar{x}}$$

pero vimos que

$$\sum_{a=1}^{k} x_a^2\, f(x_a) = \sum_{b=1}^{k} x_b^2\, f(x_b) = \sum_{i=1}^{k} x_i^2\, f(x_i) = \sigma^2 + \mu^2$$

entonces

$$\sigma^2_{\bar{x}} = \frac{\sigma^2 + \mu^2}{4} + \frac{2\mu^2}{4} + \frac{\sigma^2 + \mu^2}{4} - \mu^2_{\bar{x}}$$

pero siendo por (18)

$$\mu_{\bar{x}}^2 = \mu^2$$

se tiene

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{2}$$

para muestras de dos elementos.

Por un razonamiento análogo se llegaría, para muestras de n elementos a:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Los valores de la media y la desviación típica de la distribución de una suma de muestras $S(\bar{x})$ en infinitas muestras de una población infinita son:

$$\mu_{s(x)} = n\mu \qquad \sigma_{s(x)}^2 = n\,\sigma^2$$

Normalidad aproximada de una distribución de medias muestrales en grandes muestras de una población infinita.

Como en el caso de la población finita, la distribución de medias de las muestras de una población infinita puede ser aproximada por una distribución normal bajo ciertas condiciones.

Para valores grandes de n la distribución de las medias de las muestras de n elementos, de una población infinita, que tiene media μ y desviación típica σ, es aproximadamente normal, con media $\mu_{\bar{x}} (=\mu)$ y desviación típica $\sigma_{\bar{x}} (=\frac{\sigma}{\sqrt{n}})$; la precisión de esta aproximación es tan perfecta como queramos, cuando n crece infinitamente.

La precisión de la aproximación depende del valor de n, del grado de asimetría de la distribución y de otras causas.

El procedimiento para lograr el ajuste es semejante al empleado en los dos ejemplos de ajuste anteriores y un caso práctico nos lo aclarará:

Ejemplo:

Un dado perfecto se ha arrojado 25 veces ¿Cual es aproximadamente la probabilidad de que el promedio de puntos obtenidos en las 25 tiradas sea menor que 4? (Equivale a preguntar ¿Cual es la probabilidad de que el total de puntos obtenidos sea menor que 100?)

x	f(x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$\mu = 3,5$ $\quad \sigma = \sqrt{\frac{35}{12}}$ $\quad n = 25$

$\mu_{\bar{x}} = 3,5$ $\quad \sigma_{\bar{x}} = \frac{1}{\sqrt{25}} \sqrt{\frac{35}{12}} = \sqrt{\frac{7}{60}} = 0,342$

La relación [illegible] es

$$Z = \frac{\bar{X} - 3,5}{0,342}$$

Para $\bar{X}=4$ es $Z = \frac{4 - 3,5}{0,342} = 1,46$

de donde $P(\bar{X} < 4) = P(Z < 1{,}46)$

acudiendo a la tabla 1 resulta

$$P(Z < 1{,}46) = 0{,}928$$

Observaciones sobre una distribución binomial, como una distribución muestral teórica.

La distribución binominal

$$P_X = C_n^x \, p^x \, q^{n-x}$$

es fundamentalmente una distribución muestral teórica, es la distribución muestral de la suma S (x) en muestras de n elementos, de una población infinita, en la cual la variable aleatoria X tiene la siguiente distribución:

x	f (x)
o	q
1	p

La aparición del suceso favorable corresponde a X=1 y el contrario a X=0

La media μ de esta población es:

$$o\ q + 1\ p = p$$

luego $\boxed{\mu = p}$

y la varianza de la distribución

$$\sigma^2 = \sum_{x=0} (x-\mu)^2 \, f(x) = (o-p)^2 \, q + (1-p)^2 \, p =$$

$$= p^2 q + q^2 p = p\,q\,(q+p)$$

$$\boxed{\sigma^2 = p\,q}$$

Si se obtienen muestras de n elementos do esta población, entonces S (x), suma de las x en la muestra, es simplemente el número total de 1 de la muestra

Siendo según vimos

$$\mu_{S(x)} = n\mu$$

$$\sigma_{S(x)} = \sigma\sqrt{n}$$

sustituyendo los valores de μ y σ hallados para el caso de muestras de la población binomial resulta que

$$\boxed{\mu_{S(x)} = n\ p}$$

$$\sigma^2_{S(x)} = n\ p\ q$$

$$\sigma_{S(x)} = \sqrt{n\ p\ q}$$

estos valores los hallamos directamente al considerar la distribución binomial.

Si lo que interesa es la media $\bar{X}$ (promedio de pruebas en las que el suceso esperado ocurre) de la distribución binomial, se encuentra que la distribución tiene las siguiemtes estadísticas.

$$\mu_{\bar{x}} = p \qquad \sigma_{\bar{x}} = \sqrt{\frac{p\ q}{n}}$$

Distribución teórica de muestreo de sumas y diferencias de medias de las muestras

Se presentan con frecuencia problemas estadísticos donde hay que operar con la diferencia entre las medias de dos muestras o la suma o promedio (ponderado o no) de dos o más medias de las muestras.

Diferencia de medias de las muestras

Supongamos que $\bar{X}$ es la media de una muestra de n elementos, extraídos de una población que tiene μ de media y σ^2 de varianza y $\bar{X}'$ la media de otra muestra de n' elementos de otra población con μ' de media y $(\sigma')^2$ de varianza, la media de la distribución de la diferencia de las medias $\bar{X} - \bar{X}'$ está dado por:

$$\mu_{\bar{x}} - \bar{x}' \quad \mu_{\bar{x}} - \mu_{\bar{x}'} = \mu - \mu'$$

esta fórmula sirve para poblaciones finitas o infinitas

La varianza de la distribución de las medias está dada por

$$\sigma^2_{\bar{x}-\bar{x}'} = \sigma^2_{\bar{x}} + \sigma^2_{\bar{x}'}$$

o también:

$$\sigma^2_{\bar{x}-\bar{x}'} = \frac{\sigma^2}{n}\,\frac{N-n}{N-1} + \frac{(\sigma')^2}{n'}\,\frac{N'-n'}{N'-1}$$

en el caso de poblaciones finitas en las que N y N' son los números de elementos de las dos poblaciones respectivamente.

Para el caso de poblaciones infinitas

$$\sigma^2_{\bar{x}\,\bar{x}'} = \frac{\sigma^2}{n} + \frac{(\sigma')^2}{n'}$$

Ejemplo

Supongamos que el 80% de los votantes de una ciudad (cuya población se supone infinitamente grande) se inclina a favor de cierta propuesta. Dos muestras de 100 votantes cada una, son registradas.

¿ Cuál es aproximadamente la probabilidad de que la diferencia entre los porcentajes favorables a la propuesta exceda del 10% ?

Las muestras se han extraido de la misma población y a la variable aleatoria X se le averigua el valor 1 por el voto a favor de la propuesta y o en otro caso

$$\mu_{\bar{x}} = \mu_{\bar{x}'} = p = \frac{80}{100} = 0{,}8$$

$$\sigma_{\bar{x}} = \sigma_{\bar{x}'} = \sqrt{\frac{0{,}8 \cdot 0{,}2}{100}} = 0{,}04$$

Resulta

$$\mu_{\bar{x}-\bar{x}'} = \mu_{\bar{x}} - \mu_{\bar{x}'} = 0{,}8 - 0{,}8 = 0$$

$$\sigma^2_{\bar{x}-\bar{x}'} = \sigma^2_{\bar{x}} + \sigma^2_{\bar{x}'} = 2\,(0{,}04)^2$$

$$\sigma_{\bar{x}-\bar{x}'} = 0{,}04\sqrt{2} = 0{,}057$$

El problema puede ser transformado en este otro ¿Cuál es el valor de

$$P\,(|\bar{X} - \bar{X}'| > 0{,}1) \;?$$

Resulta

$$P\left(\left|\bar{X} - \bar{X}'\right| > 0{,}1\right) = 1 - P\left(-0{,}1 < \bar{X} - \bar{X}' < 0{,}1\right)$$

Para el caso tenemos :

$$Z = \frac{\left(\bar{X} - \bar{X}'\right) - \mu_{\bar{x} - \bar{x}'}}{\sigma_{\bar{x} - \bar{x}'}} = \frac{\bar{X} - \bar{X}'}{0{,}057}$$

donde para:

$\bar{X} - \bar{X}'$	Z
- 0,1	$\frac{-0{,}1}{0{,}057} = -1{,}76$
0,1	$\frac{0{,}1}{0{,}057} = 1{,}76$

Acudiendo a la tabla I

$$P\left(-0{,}1 < \bar{X} - \bar{X}' < 0{,}1\right) = P\left(-1{,}76 < Z < 1{,}76\right) = 0{,}9212$$

La probabilidad entonces de que la diferencia entre los porcentajes de votantes favorables a la propuesta en las dos muestras supere al 10% es

$$1 - 0{,}9212 = 0{,}0788$$

I N D I C E