consejo federal de inversiones

manual de metodologías estadísticas

parte tercera
estadística metodológica

curso de formación y entrenamiento
para funcionarios
de servicios estadisticos provinciales

buenos aires
1965

La Parte Tercera estuvo a cargo de los Profesores Dra. Martha G. de Cabezas (capítulos I, III, IV y V) y el Contador Juan B. Collado (capítulo II)

Impreso en Argentina - Printed in Argentine - Hecho el Depósito que previene la ley 11.723 (c) by "Consejo Federal de Inversiones"

Alsina 1407 - Buenos Aires

República Argentina

INDICE DE LA
PARTE TERCERA

ESTADISTICA METODOLOGICA

Capítulo I

LA DESCRIPCION ESTADISTICA

1. VOCABULARIO ESTADISTICO

Universo = Población = Colectivo

Es un conjunto o colección de individuos o elementos.

La población puede ser finita, cuando se compone de un número limitado de individuos. Ej.: los alumnos de un curso, las familias de un barrio o de una ciudad, etc.

La población o universo, es infinita cuando no se conoce exactamente el número de elementos que la componen. Ej.: la producción ininterrumpida de tornillos que sale de una máquina en funcionamiento, donde cada tornillo es un elemento de población.

El concepto de población no ha de referirse necesariamente a una colección de organismos vivientes. Así, no sólo puede hablarse de la población constituida por los habitantes de un país, o por los árboles de un bosque, sino también, de la población de establecimientos comerciales de una ciudad o de la constituida por las letras de un diccionario, o al conjunto de medidas de talla o de peso de un grupo de soldados, etc.

En resumen:

Una sola observación	=	individuo
Observaciones repetidas	=	población o universo

Atributo = Variable

Es una magnitud sujeta a variación. Se denominan también variables.

Constante = Parámetro = Cantidad Fija = Coeficiente

Es una cantidad fija, determinada.

Constante absoluta: tiene siempre un mismo valor, por ej.:

el número Pi = 3,1416...
el número e = 2,718 ...

Constante relativa: tiene un valor determinado dentro de un determinado problema o ecuación, pero adopta valores diferentes en diferentes problemas:

Ej. $Y = a\,x$, donde a, puede tomar distintos valores.

Ej. $Y_1 = 300\ \$\ x$

$Y_2 = 50\ \$\ x$

Atributo Cuantitativo:

La variable puede adoptar cualquier valor numérico. Existe para todos los individuos del universo, pero cada uno la posee en distinto grado. Ej. la talla, el peso, el ingreso, el gasto en alimentación, etc.

Atributo Cualitativo:

La magnitud se presenta por conceptos o cualidades no mensurables en forma gradual. Se caracteriza porque los individuos de la población poseen o no poseen, el atributo en cuestión. Ej. :

atributos dicotómicos	obrero sindicalizado	-obrero no sindicalizado
	persona ocupada	-persona desocupada
	varón	-mujer
	sano	-enfermo

atributo policotómico: soltero - casado - viudo - separado -divorciado.

Los atributos cuantitativos pueden ser:

VARIABLE DISCRETA: Se mide por números enteros. Ej. No. de hijos, No. de cuartos de la vivienda, grados de escolaridad, etc.

VARIABLE CONTINUA: Se mide por grados sucesivos de incidencia. Ej. Ingresos, Edad, Estatura, Superficie, etc.

DISTRIBUCION: Es la forma sintética de presentar las observaciones o frecuencias.

DISTRIBUCION UNIDIMENSIONAL: Las frecuencias se clasifican según una sola propiedad o variable. Ej.:

Cada grupo de valores se llama intervalo de clase. Frente a él aparece la frecuencia o cantidad de veces que aparecen los valores incluidos en cada intervalo.

Obreros ocupados según salario percibido

Salario $	Obreros
140 - 149	10
150 - 159	60
160 - 169	100
170 - 179	40
180 - 189	1
Total	220

DISTRIBUCION BIDIMENSIONAL (Tabla de doble entrada)

Las frecuencias se clasifican según la variación simultánea de dos atributos o variables.

Ej. Clasificación de familias según la edad del esposo y la esposa.

edad del esposo	edad de la esposa	de 20 a 25	de 25 a 30	de 30 a 35	de 35 a 40	de 40 a 45	de 45 a 50	de 50 y más	Total
55 y	más							2	2
50 -	55						3		3
50 =	45				6	15	5		26
45 -	40			5	13	4			22
40 -	35		10	15	17				42
35 -	30		8	20	9	1			38
30 -	25	7	2						9
25 -	20	8							8
TOTAL		15	20	40	45	20	8	2	150

DISTRIBUCION CUALITATIVA:

Las frecuencias de caracteres cualitativos pueden presentarse de la siguiente manera:

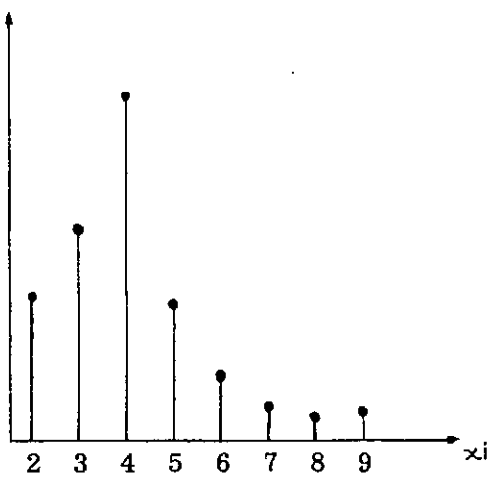
Ocupación del jefe de familia	frecuencia	%
Total	150	100
Profesional	6	4
Comerciante	15	10
Empleado	50	33
Agricultor	12	8
Obrero Industrial	67	45

El tratamiento de estas distribuciones, se reduce en general, al cálculo de porcentaje a frecuencias relativas que permita una más fácil comparación así como una rápida visión de las proporciones existentes en la población.

DISTRIBUCION CUANTITATIVA A VARIABLE DISCRETA:

Si el grupo de familias anteriormente considerado, lo observamos ahora, según el número de miembros de la familia, tendremos una distribución cuantitativa a variable discreta cuya presentación numérica y gráfica son las siguientes:

Número de miembros de la familia x_i	frecuencias y_i	%
2	21	14
3	36	24
4	60	40
5	20	13, 3
6	9	6
7	3	2
8	0	0
9	1	0, 6
	150	100, 0

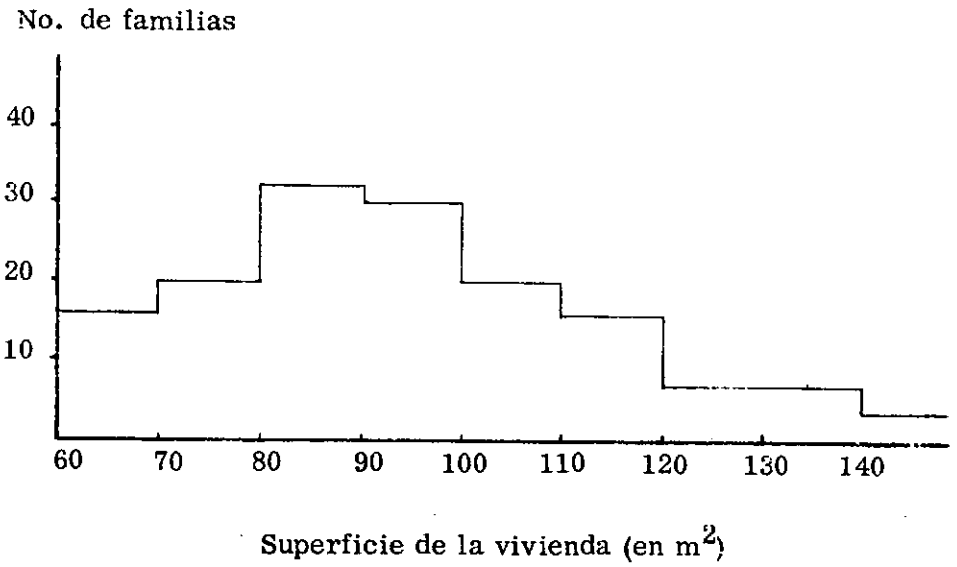


No. miembros

DISTRIBUCION CUANTITATIVA A VARIABLE CONTINUA (con intervalos iguales)

Si nos referimos a la superficie de la vivienda que ocupa la familia, estamos frente al caso de una variable "continua". Si agrupamos los valores de superficie en intervalos de clase convenientes, se observa una distribución de la siguiente forma: en la que todos los intervalos son iguales.

Superficie de la vivienda m2.	familias y_i
60 a 69, 9	16
70 79, 9	20
80 89, 9	32
90 99, 9	30
100 109, 9	20
110 119, 9	16
120 129, 9	6
130 139, 9	6
140 149, 9	4
Total	150



DISTRIBUCION CUANTITATIVA A VARIABLE CONTINUA(con intervalos diferentes)

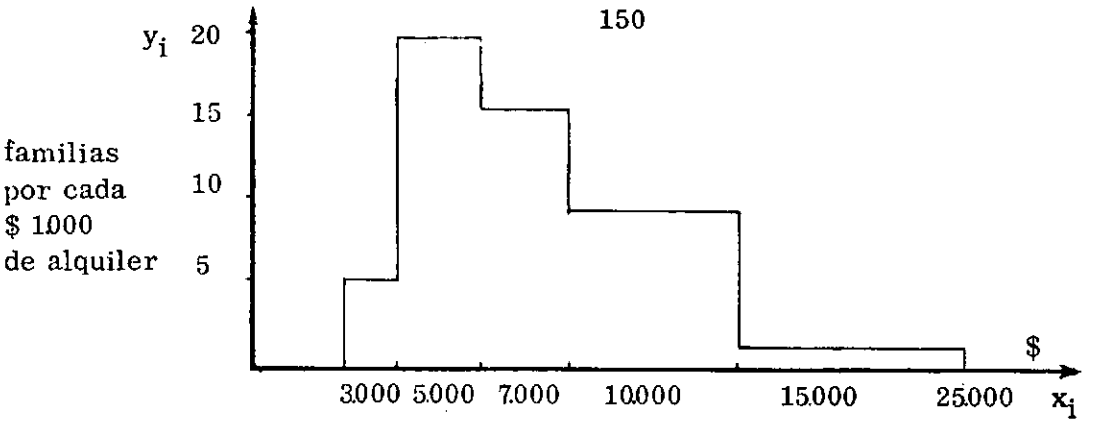
En algunos casos -muy frecuentes en observaciones de carácter económico-social- el atributo varía en intervalos excesivamente amplios, por lo cual no es aconsejable ordenar los datos en series de intervalos iguales entre sí. Por el contrario, las frecuencias se distribuyen entre pequeños intervalos al comienzo y, luego, otros gradualmente mayores. La representación gráfica en estos casos requiere un tratamiento especial. Para ello, previamente se debe calcular la proporción de frecuencias que le corresponde a cada unidad de intervalo, mediante el cociente del número de frecuencias de cada uno, por la respectiva amplitud de intervalo (w).

Si en nuestro ejemplo, agrupamos los datos del "ingreso familiar" en intervalos de clase, convenientes, se observa una distribución de la siguiente forma numérica y gráfica:

Intervalos de Ingreso x_i $	familias Y_i	Amplitud en miles $ w	Familias por cada mil $ y_i:	w	Fam.
menos de 2000	35	2	17, 5		
2000 4000	55	2	27, 5		
4000 6000	21	2	10, 5		
6000 8000	12	2	6		
8000 10000	9	2	4, 5		
10000 20000	10	10	1		
20000 30000	5	10	0, 5		
30000 40000	2	10	0, 2		
40000 50000	1	10	0, 1		
	150				

Es indispensable homogeneizar las frecuencias en cada intervalo a fin de que la superficie de cada rectángulo (base x altura) sea realmente proporcional a las frecuencias. Así, por ejemplo, en el 2o. y 4o. intervalos de la siguiente serie, se registra el mismo número de frecuencias. Sin embargo, al ser mayor la amplitud del 4o. intervalo, la base del rectángulo deberá ser también mayor en el gráfico, y en consecuencia habrá que dar menor altura al rectángulo para que la superficie sea efectivamente equivalente al que representa al 2o. intervalo.

Alquileres pagados por la vivienda	Amplitud en $ w	familias y_i	familias por cada $ 1.000 de alquiler (y_i : w) x 1.000 $
de 3.000 a 4.999	2.000	10	5
5.000 6.999	2.000	40	20
7.000 9.999	3.000	50	16,6
10.000 14.999	5.000	40	8
15.000 25.000	10.000	10	1
		150	



alquileres pagados por la vivienda.

FUENTE PRIMARIA: Entidad que organizó y compiló o proporcionó la información.

FUENTE SECUNDARIA: Se aprovecha de datos obtenidos por otras instituciones y ella los analiza.

FUENTES OFICIALES: Entidades públicas encargadas de una labor permanente u ocasional de relevamiento estadístico.

En nuestro país: Dirección Nacional de Estadísticas y Censos (D.N.E.C.).

Direcciones de Estadísticas de Provincias: Instituto de Investigaciones Económicas y Tecnológicas de Mendoza (I.I.E.T.).

FUENTES PRIVADAS: Entidades no públicas. Ej. Sindicatos, Empresas.

Estadística estática: Población o universo en un momento determinado; en general son censos porque averiguan los datos en un momento determinado y luego no se continuan. Por ej.: Datos de una elección. Es una fotografía de la estructura social o económica, en un momento determinada. Estadística dinámica: Es la que continuamente va acumulando datos. Ej. Registro Civil: nacimientos, defunciones, matrimonios. Otro ejemplo sería el comercio exterior, de exportación e importación. Se compara con una película.

2. ETAPAS DE UNA INVESTIGACION ESTADISTICA

Investigación es un proceso por el cual el investigador busca determinar los elementos:

- determinantes o causas
- concurrentes
- influyentes
- presentes

que intervienen

- en un problema
- en una situación social

para

- resolver el problema o
- mejorar la situación

Por ser proceso, es una serie de etapas. Veamos esas etapas o fases de una investigación (como especial referencias a investigaciones estadísticas por muestreo).

I - Fase: Indicación de los Objetivos de la Investigación o planeación

Deberá determinarse previamente :
- Qué objetivos se persiguen, y
- Cómo se espera utilizar los resultados.

Es necesario establecer los fines u objetivos muy concretamente y no limitarse a una enunciación vaga de los mismos. Así, se debe expresar claramente.

1o. La definición de la población o universo total, que se aspira conocer y la parte del mismo que se considerará en la muestra.

En el caso que se elimine alguna parte de la población total o ideal, hay que justificar los fundamentos que se tuvieron para prescindir de ellos. En el caso de practicar una "poda" o "desmoche" (out-off) debe recordarse que los resultados de la encuesta se refieren estrictamente a la población considerada de un modo efectivo. Así debe hacerse cuando se prescinde de un grupo humano por ser nómade (bracero, trabajadores "golondrinas"); o difícilmente accesible, o por presentar una parte muy poco importante dentro de la población total (en un censo industrial, no se incluyen los establecimientos que no tienen fuerza motriz instalada o que no tienen por lo menos un empleado a sueldo).

Como generalmente es bastante difícil definir la población, se deben extremar los cuidados para establecer criterios muy claros y adelantarse a cualquier problema que se le presente a los entrevistadores en su trabajo de campo. Veamos un ej.:

Se desea averiguar el número de personas entre 18 y 44 años inclusive, de una ciudad. Aunque hiciéramos un censo completo, -esperaríamos el número exacto? -Debemos definir con precisión lo que queremos. Ej.:

"Contar todas las personas de la ciudad entre 18 y 44 años inclusive, en un instante determinado, excluyendo a cualquiera que esté de paso e incluyendo a los que normalmente viven en esa ciudad pero están ausentes en ese momento".

Esta definición aunque sea clara, podría ser difícil de aplicar con exactitud en una entrevista de campo, sobre todo si es instantánea. Por ejemplo:

La expresión "que normalmente viven en esa ciudad", -se incluyen o no las personas que viven en esa área, mantienen allí su casa, pero viven 2 o más meses, quizás 6 meses fuera de ese lugar (vacaciones, trabajo, estudio, servicio militar, enfermedad, etc.)? . Supóngase que hay personas que tienen casa en esa ciudad, pero trabajan en otra área y vienen allí para los fines de semana, y también otras que tienen una casa en dos áreas diferentes y dividen su tiempo de estadía entre ambas, -en cuál de los dos lugares viven?

También hay personas que no tienen casa allí, sino que están en esa ciudad temporalmente, o simplemente están de paso en un remolque que es su único hogar.

Como se ve, hay muchos problemas para definir seguramente a una población aparentemente tan simple de contar.

Aunque se haya definido el verdadero valor, todavía quedan, seguramente, muchos problemas al aplicar la definición en la entrevista del campo.

Así, pueden surgir como errores:

1) omisiones de gente que debió incluirse
2) inclusión de gente que no debió censarse
3) duplicación en la enumeración de algunas personas.

Además, si estuviéramos de acuerdo sobre el tipo de personas a incluir o no, surgiría otro tipo de error: es el referido a la edad; debido a:

1) desconocimiento de la edad exacta
2) tendencia a responder en números redondos
3) intención de responder incorrectamente la edad.

Los errores pueden ser aun más serios si la pregunta estuviera referida a:

-empleo, tarea que realiza, jerarquía en la empresa
-ingresos
-opinión política, etc.

En el Censo Nacional del 30 de septiembre de 1960, de población, vivienda y agropecuario, se establecieron las siguientes normas básicas para el censista que visitaría a cada vivienda:

Quiénes deben ser incluidos en esta planilla?

Ud. debe incluir en esta planilla a todas las personas que pasaron la noche del 29 al 30 de septiembre en esta vivienda, y a las que por razones de trabajo nocturno o por estar viajando o por cualquier otra causa de fuerza mayor, no pasaron dicha noche en la misma, pero que viven en ella y se encontraban presente en la mañana del día del censo, siempre que no hubiesen sido censadas en otro lugar.

También debe incluir a los nacidos antes de la 0 hora del día 30 de septiembre y a los fallecidos después de esa hora.

Quiénes no deben ser incluidos en esta Planilla?

Ud. no debe incluir en esta planilla a las personas que, aunque vivan en la vivienda no pasaron la noche del 29 al 30 de septiembre de 1960 ni estaban presentes en la mañana del día del censo.

Tampoco debe incluirse a los nacidos después de las 0 horas del día 30 de septiembre ni a los fallecidos antes de esa hora. Tampoco debe incluirse a las personas que habiten en esta vivienda, pero que en el momento del censo se hallen internadas en algún establecimiento sanitario (hospital, sanatorio, etc.), o pupilas en algún establecimiento educacional (oficial o privado) o cumpliendo el servicio militar obligatorio (conscripto) o recluidos en algún establecimiento carcelario o similar (cárcel, alcaidía, etc.) aunque en el momento de su visita ocasionalmente en la vivienda.

2o. El motivo del muestreo; la forma en que se van a utilizar los resultados; y el modo en que habrán de influir en decisiones posteriores. Si la información se hace a título informativo o si será decisivo. Esto interesa para determinar luego, hasta qué punto puede llegar el error probable.
Ej.: Aceptar o rechazar un nuevo medicamento, adoptar decisiones políticas.

3o. Las especificaciones de la característica a estimar. Acá se hará un detalle de las características que se desean: Promedio, total, mediano, porciento.

4o. Las tabulaciones y grados de confianza y de precisión que se consideran adecuados.

Se preparan anticipadamente todos los esquemas de cuadros estadísticos que resultarán de la compilación final de los datos, poniendo gran cuidado en la distribución de los encabezamientos, columnas y filas.

TABULACION

Criterios

1o. A la vista de los objetivos propuestos, se analizarán cuáles son las variables independientes y cuáles las dependientes.

Las primeras constituirán la columna matriz del cuadro, y los otros llenarán los encabezamientos de las columnas auxiliares.

2o. Se harán tantos juegos de cuadros, cuantas variables independientes haya, por ejemplo, uno en que la columna matriz corresponda a la variable independiente edad y otro juego, a la variable lugar de residencia.

3o. Se debe preparar en principio, el siguiente material:

1o.) Planilla de vuelco de los cuestionarios:
En el caso de que intervengan varios encuestadores convendrá que cada uno prepare y presente la planilla con los resultados de su zona.

2o.) Planilla de resumen de vuelcos:
Reúne las planillas de cada encuestador.

3o.) Cuadros de Referencias:
Sintetizan en cifras absolutas los resultados finales encontrados para cada aspecto de la variable que se estudia.
Acompañan el estudio final en forma de apéndice.

4o.) Cuadros analíticos:
Aunque no difieren de los cuadros de referencia en su encabezamiento, se distinguen porqué sólo se usan para presentar los resultados de algún análisis efectuado con las cifras de aquellos.
Así se presentan tablas o cuadros analíticos para porcientos, promedios, índices, tasas o coeficientes y relaciones entre dos o más fenómenos.

Además, se establecen los grados de error admisibles según la naturaleza del trabajo.

II - Fase: Condiciones, Recursos y Limitaciones

Entre las condiciones y circunstancias que puedan influir en el plan de trabajo, debe tenerse presente:

1. Qué organismo o personas serán responsables o auspiciadores del trabajo.
2. Cuánta información existe ya sobre el tema para evitar duplicación de trabajo y aprovechar sus resultados (ficheros, padrones, cartografía, etc.)
3. Qué amplitud o limitaciones y qué seguridad tiene esa información (fuente secundaria).
4. Qué limitación existe para encarar la investigación en cuanto a:
   a) Presupuesto disponible para: trabajos de preparación, de campo, de supervisión, de elaboración, de análisis y de publicaciones.
   b) Tiempo a emplear y fecha de iniciación.
   c) Personal
   d) Disposiciones administrativas o legales.
   e) Equipo mecánico.

III - Fase: Programas de Operaciones.

Si se opta por el muestreo, hay que estructurar:

1. Diseño o esquema del muestreo, para captar los datos.
2. Los metodos para estimar la seguridad de los resultados que se obtienen.
3. Las medidas de precisión y confianza.
4. La redacción del cuestionario o escala sociométrica para medir opiniones o actitudes. Libre; grupal; controlada.
5. Los metodos de recolección (por correo, por entrevista, por consultas de fuentes preexistentes.
6. Los manuales de instrucción (definiciones)
7. La organización de trabajo de campo y de oficina.
8. La preparación y entrenamiento del personal y equipos.
9. Las disposiciones sobre:
   a) Codificación de datos.
   b) Supervisión
   c) Tratamiento del: "no consta" o "no hay respuesta"; o "seguimientos" o "visitas repetidas".
10. El calendario de operaciones.

IV - Fase: Ejecución del Programa y Recolección de Datos o Realización.

Si no se dispone de experiencia sobre trabajos análogos, es imprescindible realizar una o más muestras pilotos para:

1o. Poner a prueba el cuestionario proyectado.
2o. Ensayar las instrucciones.
3o. Pulsar las reacciones de los encuestados.
4o. Estimar el por-ciento de faltas de respuesta.
5o. Calcular el tiempo que insume cada encuesta.
6o. Obtener informaciones sobre la variabilidad de la población en estudio para evaluar el tamaño definitivo de la muestra, para hacer estimaciones cuya precisión y costo se aproximan a los establecidos en la I Fase.

Una vez obtenidos los datos, se procede a la compilación o recuento de datos mediante las siguientes operaciones:

1o. Operaciones críticas
   a) Prueba de integralidad (revisar si faltan respuestas).
   b) Prueba de consistencia (revisar si los datos son congruentes).

2o. Operaciones mecánicas
   a) Compilación manual (...
   b) Compilación mecánica

V - Fase : Análisis y Aprovechamiento de los Resultados.

Los cálculos con las correspondientes tabulaciones básicas y representaciones gráficas, deben ir acompañadas por comentarios e interpretaciones.

El análisis de las cifras comprende:

1) Cálculo de las estimaciones de características.
2) Cálculo de los errores de muestreo.
3) Análisis de las características mediante el uso simple o combinado de los métodos descriptivos típicos, de:
   a) Las series de frecuencias: media, mediana, modo, dispersión absoluta y relativa, asimetría, kurtosis, cuartiles y porcentuales, distribución y concentración, correlación, análisis factorial.
   b) Las series de tiempo: índices, tendencias, variaciones estacionales y cíclicas, tasas demográficas, índices de rendimiento y productividad.

VI - Fase : Presentación del Informe Final, Debe Contener :

Indice, Metodología empleada-Limitaciones-Comentarios-Gráficos-Conclusiones-Soluciones-Bibliografía General y Especial.

R E S U M E N

Es muy conveniente que una vez efectuada la investigación -sobre todo si se hizo por muestreo- se indique cuál ha sido el desarrollo efectivo de las faces antemencionadas, destacando detalladamente el capítulo de costos y haciendo una crítica general de las operaciones realizadas y de las diferencias entre lo que se trataba de hacer y lo que efectivamente se hizo.

Tiene especial interés la comparación entre el error máximo que se proyectaba, como admisible en la primera fase, y el error de muestreo calculado a posteriori. También hay que analizar las posibles fuentes de error "o vicios" de la muestra, como el "no consta", "no responde", etc.

En síntesis puede decirse que las distintas fases responden a las preguntas:

1a. fase	Qué queremos?
2a. fase	Con qué medios contamos?
	A qué vehículos debemos someternos?
3a. fase	Qué cosa debemos hacer?
4a. fase	Realización
5a. fase	Qué es lo que hicimos?
Es decir, nos planteamos:	Quién ? Qué? Dónde? Con qué? Para qué? Cómo? Cuándo?

Conviene tener muy en cuenta el cumplimiento de las distintas fases enunciadas ya que, de nada servirá un perfecto diseño matemático del muestreo, fruto de un estudio detenido y dificultoso si fallan los supuestos en que se basa.

Por ejemplo: el incumplimiento de las condiciones por parte de los entrevistadores, ya sea por incompetencia, irresponsabilidad o parcialidad, mala contestación de las preguntas del cuestionario, utilización defectuosa, etc.

F U E N T E:

Tomado de Azorin Poch, Francisco, Curso de Muestreo y Aplicaciones (Venezuela, 1961) pág. 27.

Hansen, Hansen, Hurwits and Madow, Sample Sur... Methods and Theory, Vol. I (New York, 1953).

Uribe Villegas, Técnicas Estadísticas para Investigadores Sociales (México).

3. REPRESENTACION ESTADISTICA

3.1 - Representación Tabular

Reglas Generales de Trabajo y Presentación Tabular

Norma de presentación tabular: 1o. Definiciones:

Las partes principales de una tabla son: Título - Cuerpo - Encabezamiento y Columna Indicadora - Fuente - Notas Aclaratorias.

2o. El Cuerpo : de la tabla comprende columnas y líneas que contienen, respectivamente, las series verticales y horizontales de informaciones. La intersección de una columna con una línea se llama casilla.

3o. El Encabezamiento : es la parte de la tabla en que se designa -con toda claridad-la naturaleza del contenido de cada columna.

4o. Columna Indicadora : es la parte de la tabla en la que se indica el contenido de cada línea, pudiendo una misma tabla tener más de una columna indicadora.

5o. Título : es la parte superior o inicial de la tabla, en la cual se indica con toda precisión y claridad:

a) Naturaleza del hecho que se estudia (cuyos datos aparecen en el cuerpo de la tabla).

b) Lugar al cuál se refieren los datos:

c) Epoca en que fue observado.

6o. Fuente : es la indicación al pie de la tabla de la entidad responsable que organizó y compiló el material estadístico, o bien la entidad que proporcionó los datos respectivos.

7o. Notas o Llamadas : son las informaciones en lenguaje conciso (pero no incompleto) colocadas al pie de la tabla, después de la fuente cuando la materia contenida en la tabla exige aclaración.

a) Se usa la Nota para aclaraciones de conceptos constantes y de carácter general en la tabla.

b) Se usa la llamada para aclarar ciertas minucias en la relación con la información de las casillas, las líneas o columnas.
Se enumeran con números arábigos entre paréntesis siguiendo la sucesión de los números naturales en su colocación de izquierda a derecha, de arriba hacia abajo.

Reglas Generales para la Presentación de la Tabla Estadística.

A) Generalidades:

1o. Cada tabla debe tener un significado propio de modo que, cuando se consulte aisladamente, pueda ser interpretada sin el auxilio de textos anexos (o información verbal del empleado que la confecciona o su anterior compilador).

2o. Ninguna casilla debe quedar en blanco. Siempre debe presentar un No. perfectamente legible o bien el signo convencional que corresponda. Los signos convencionales son: Un guión (-) para señalar que la magnitud es cero o no alcanza a la mitad del último dígito usado; Un punto (.) para indicar que no existe el concepto en el período correspondiente o no puede ser obtenido; Tres puntos (...) significa que el dato no ha sido compilado o elaborado hasta la fecha de publicación; Un asterisco (X) señala que la cifra es ahora provisional o estimada; Un parágrafo () indicará que la cifra que se publica a su lado, fue antes publicada pero sin la advertencia que se trataba de una cifra provisoria (cifras que rectifican informes anteriores).

3o. Como la principal finalidad de la tabla estadística es revelar la evidencia numérica de determinados fenómenos, se evitarán la presentación de tablas en las que en la mayor parte de las casillas indiquen la inexistencia del fenómeno.

4o. Ninguna tabla será dispuesta de manera que la lectura exija colocar el volumen fuera de su posición normal, es decir que deben evitarse en lo posible los cuadros apaisados.

5o. Las tablas deben cerrarse en la parte superior e inferior con líneas horizontales tipo gruesa.

a) Cuando la tabla, en sentido vertical debe continuar en la página siguiente, deberá repetirse el encabezamiento en dicha página.

6o. Las columnas muy extensas deben tener de cinco en cinco o de diez en diez líneas, un intervalo en blanco.

7o. Cuando en una tabla más de una columna se presenta bajo una misma especificación, sepárese a ese conjunto por una línea más gruesa.

8o. Los conjuntos tabulares deben ser precedidos de una indicación de las señales o unidades empleadas y al final, una relación completa de las fuentes con sus respectivas direcciones.

B) Enumeración de Títulos y Subtítulos:

1o. Las publicaciones en las que se publiquen varias tablas estadísticas, éstas deben tener un No. de órdenes.

2o. Se ha adoptado el siguiente orden de preferencia para las diferentes indicaciones de títulos y subtítulos de las tablas:

a) En primer lugar, números romanos seguidos de un trazo. Así: I - II - etc.

b) En segundo lugar, algoritmos arábigos seguidos de un punto. 1. 2. etc.

c) En tercer lugar, letras minúsculas seguidas de paréntesis, a), b), etc.

d) Habiendo necesidad de una cuarta enumeración, se continúa con letras de mayúsculas de imprenta. A), B), etc.

e) Tornándose imprescindible una quinta enumeración o subdivisión, se recurre a las letras griegas seguidas de paréntesis.

C) Modo de Presentar las Especificaciones de una Columna Indicadora.

Después de una especificación en la columna indicadora, se hará una línea punteada (.....) hasta encontrar el comienzo de la primera columna del cuerpo de la tabla.

En el caso de existir columnas indicadoras sucesivas a la principal, las líneas punteadas se indican donde termina cada tópico y se lleva hasta el comienzo de la primera columna.

Santa Rosa	Almendros..... Cerezos..... Damascos..... Durazneros..... Manzanos.....

D) Colocación y Denominación de los Totales.

Siempre que sea posible, la suma de los datos de una columna, será intitulada por ej.:

La Provincia .....
Total del País .....
Costo total por kilo.....

Cuando no fuera posible, se anotará la palabra total. En cualquier caso se evitará la palabra (Suma).

a) Los totales se escribirán en forma destacada.
Cuando se realice el trabajo en imprenta, estos totales deberán imprimirse en negritas, sin trazo horizontal que corte la columna.

b) Por lo general, los totales se escribirán después de los parciales. No obstante, si existen subtotales o un total está discriminado según diversas especificaciones parciales, el total figurará antes de los parciales.

E) Empleo de Denominaciones Especiales.

a) Se usará la denominación "otros" o "no especificado" cuando varias especificaciones se incluyan en un solo título. Estas deben ser aclaradas al pié de la tabla.

b) Se usará la denominación "No declarado" para significar el agrupamiento de casos que los declarantes de la encuesta omitieron informar.

c) No debe usarse la expresión "diversos".

F) Abreviaturas

Las abreviaturas se indican siempre en singular. Van seguidas de punto, sin perjuicio de la puntuación corriente del texto, Ej.:

m. Tm. m$n. Ha. Hl.

Cuando se desea expresar el nombre de los meses y no existe lugar para indicarlos in extenso, se indica con números romanos; así:

Enero I
Febrero II
etc.

G) Redondeo de Cifras:

1o. Siempre que fuera necesario redondear un dato estadístico, se adoptará la siguiente regla:

a) Cuando el primer número a ser despreciado fuera menor que 5 (0, 1, 2, 3, 4) se desprecia (redondeo por falta o por defecto) y se mantiene invariable la última cifra.

b) Cuando el primer número a ser despreciado fuera mayor que 5 (6, 7, 8, 9) se aumenta el número anterior en una unidad 9 redondeo por exceso).

c) Cuando el número a despreciar fuera justamente 5: si las cifras siguientes son cifras significativas (distintas de cero) aumentar la última cifra en 1.

Si todas las cifras siguientes son cero, aumentar la última cifra en 1, si ella es par. Se mantiene invariable si es impar.

Ejemplos :

23,753.....	23,75
23,758.....	23,76
23,75512.....	23,76
23,75500.....	23,75
23,74500.....	23,75

2o. En caso de suma, deben redondearse el parcial y todos los parciales.

3o. Si la suma de los parciales fuera superior al total redondeado, vuélvase a la serie original para dejar de redondear por exceso tantos parciales cuantas fueran las unidades excedentes, dentro de esos parciales escoger aquellos cuyas fracciones despreciadas fueren un número que más se aproxime a 5, 50, 500, etc. según el caso. Seguir procedimiento similar para el caso que los parciales sumados den un número inferior al total redondeado.

Serie Original	Redondeo por la Regla 3	Ajustamiento
22, 55	23	22
6, 00	6	6
18, 52	19	18
9, 71	10	10
12, 53	13	12
3, 57	4	4
10, 64	11	11
7, 63	8	8
3, 21	3	3
5, 62	6	6
99, 98	103 100	100

4o. El redondeo se efectúa de la siguiente forma en caso de subtotales: Los subtotales en base al total general y los parciales en base a los subtotales.

H) Tipo y Disposición de los Caracteres (títulos, subtítulos, encabezamientos, columna indicadora).

1o. Para distinguir los números deben emplearse los siguientes tipos de imprenta:

Para totales : tipo negrito grueso en texto manuscripto o dactilografiado, subrayar con línea doble.

Para los porcentajes: (números relativos) cuando se presentan conjuntamente con números absolutos, deben imprimirse en letras bastardilla o itálica. Si esto ocurre en textos manuscriptos o dactilografiados, subrayar con una sola línea.

2o. En los encabezamientos, las diferentes designaciones del contenido de las columnas, deberán escribirse en el mismo cuerpo. En el caso de subdivisión, el cuerpo deberá ser gradualmente menor e idéntico para las subdivisiones del mismo tipo.

I) Reglas Generales de Trabajo y Presentación

1o. Cada hoja de cálculo -o preparación de cuadro estadístico- debe quedar archivada ordenadamente, dactilografiado o escrita en tinta.

2o. Al iniciar cualquier trabajo, lo primero que debe hacerse es colocar el título claro y conciso, con indicación de tema, lugar y fecha.

Jamás deben surgir dudas respecto a la identificación de los datos contenidos en el cuadro.

3o. No deben aparecer en dichas hojas,, números borrosos o dudosos. Cualquier error debe ser salvado convenientemente.

4o. Todo gráfico debe indicar en el eje de las ordenadas y abscisas, las escalas y unidades utilizadas.

5o. Ningún trabajo se considera completo sin un breve sumario del método empleado, variantes observadas y proyecciones cuando sea posible.

PROCEDIMIENTO A SEGUIR PARA LA CONSTRUCCION DE UNA DISTRIBUCION DE FRECUENCIAS AGRUPADAS

Tabla No. 1

102	113	127	111	109
99	114	138	118	111
109	124	115	122	134
108	118	122	99	109
106	122	133	124	108
102	130	107	104	141
103	108	118	113	138
101	109	112	103	104
111	106	126	114	102
107	146	108	100	114

1. Preparar una hoja de trabajo con las siguientes columnas:

Tabla No. 2

Valores	Tabulación en damero									
	0	1	2	3	4	5	6	7	8	9	Fcia.
90										/	
100	/	/									

2. En la columna de Valores, listar en orden creciente todos los valores posibles incluidos entre los datos a ser tabulados, y en éscala de 10. (También puede listarse en orden correlativo -a escala de 1- en cuyo caso se trabaja con una sola columna de tabulación, donde se van anotando con tildes, el número de veces que aparece cada uno de los valores).

3. Inspeccionar la Tabla No. 1 de valores originales. Tildar sucesivamente cada valor, mientras simultáneamente se coloca una marca en el cuadro del damero correspondiente al número. En nuestro ejemplo: 102, 99, 103, etc.

4. Si en cada celda del damero se han ubicado muchas marcas que dificultan la lectura, se cruza con una línea, al anotar la quinta observación, cada grupo de cuatro marcas. Luego se cuentan los grupos y se multiplica por cinco para obtener el total de unidades.

Ejemplo : 𝍸 𝍸 // = 12

5. Contar el número de marcas de las celdas de cada línea y anotar este total en la misma línea de la columna de frecuencias. La suma de estas frecuencias debe ser igual al total de los datos de la tabla número uno.

6. Determinar el intervalo de variación o rango de la serie. Está formado por los valores más bajos y más altos de la serie. En nuestro ejemplo: 99..... 146; por lo tanto el intervalo de variación es igual a 47.

7. Determinar la amplitud del intervalo de clase.

Generalmente es preferible trabajar con series de intervalos iguales. Para determinar la amplitud del intervalo de clase w, no existen reglas fijas aplicables a todos los casos, pero se aconseja arreglar los datos de modo que la serie tenga entre 10 y 20 intervalos. A tal fin, el valor de w, se determinará así: Dividir el intervalo de variación por 10 y por 20; redondeando el cociente se obtendrá el valor de w.

Ejemplo: 47 : 10 = 4, 7 redondeando será w = 5
47 : 20 = 2, 3 redondeando será w= 2

Pero no siempre se pueden formar series con intervalos iguales entre sí.

Diferentes tipos de datos requieren diferentes métodos de trabajo; factores que son importantes en una situación, no son en otras importantes. Por lo tanto, el buen criterio del investigador, aplicado al conocimiento del problema, aconsejará en cada caso el método más oportuno a seguir.

Tratándose de problemas de morbilidad y mortalidad, se utilizan intervalos desiguales, ya que la incidencia de las enfermedades no es igual en todas las edades. No tendría sentido clasificar un grupo de pacientes de sarampión en intervalos de 10 años, ya que todos los casos caerían en el 1o. y 2o. intervalo (de 0-9 y de 10 a 19). Tampoco tendría sentido el hacer un estudio de morbilidad y mortalidad por cáncer, usar la mencionada clasificación en intervalos de 10 años, porque muy pocos caerían en los intervalos inferiores. Por todo ello, cuando se trabaja con problemas sanitarios, educacionales, laborales y demográficos se usan los siguientes intervalos:

Menos de 1 año	Lactante	población pasiva
de 1 a 4 años	preescolar	
de 5 a 14 años	escolar	
de 15-24 años	adolescente	población activa
de 25-44 años	adulto joven	
de 45-64 años	edad madura	población pasiva
de 65 y más años	edad mayor	

La clasificación de las rentas o ingresos de las personas también se efectúa en intervalos desiguales detallando en intervalos pequeños, las rentas de la clase baja y media; y agrupando en intervalos muy amplios las correspondientes a la clase adinerada. Ejemplo:

INGRESOS MENSUALES:

Hasta $ 7.000.--	
de 7.500 a 10.000.--	w = 2.500
de 10.001 a 12.500.--	
de 12.501 a 15.000.--	
de 15.001 a 20.000.--	w = 5.000
de 20.001 a 25.000.--	
de 25.001 a 30.000.--	
de 30.000 a 50.000.--	w = 20.000
de 50.001 a 75.000.--	w = 25.000
más de 75.001.--	

Si se trabaja con un problema para el cual no se ha adoptado ya intervalos convencionales, como en los ejemplos dados en el punto 7, habrá que proceder por "tanteo". Se preparan dos o más distribuciones de frecuencias con distintos intervalos de clase, llamados también módulos. Se hace la representación gráfica (histograma), y se elige la más suave y regular. Los distintos rectángulos del histograma deben escalonarse paulatinamente.

8. Construir la distribución de frecuencias.

En la primera columna de la tabla, se inscriben los límites de cada intervalo ($L_1 - L_2$)

Donde L_1 = límite inferior del intervalo de clase o módulo

L_2 = " superior " " " " " "

Los valores de la variable, deben satisfacer esta condición:

$$L_1 < X_i < L_2$$

Esto implica una condición importante: Los intervalos deben ser mutuamente excluyentes para que no se pueda colocar una misma observación ya sea en uno u otro intervalos sucesivos.

9. Confeccionar la tabla número tres.

W	X_i	f_x	f_r	f_a

En la primera columna se anotan los límites de los intervalos de clase, en orden creciente. Se comienza con el intervalo que contiene al valor más pequeño y se continúa con todos los intervalos hasta aquel que contiene el valor más grande.

La segunda columna se reserva para anotar el punto medio del intervalo.

$$X_i = \frac{L_1 + L_2}{2}$$

En la columna siguiente se anotan frente a cada intervalo, las frecuencias obtenidas en el damero para cada grupo de valores. Estas cifras o "repeticiones" se llaman también frecuencias absolutas u observadas.

En la columna cuatro podrán anotarse las frecuencias relativas que se obtienen por cociente entre la frecuencia de cada intervalo y el total de frecuencias. Para más comodidad, estos cocientes se multiplican por 100. Cada uno de estos porcentajes indican la "importancia" relativa de cada grupo. También puede confeccionarse la columna quinta destinada a registrar las frecuencias acumuladas, que consiste en anotar la suma con arrastre de las frecuencias de cada intervalo. El valor anotado para el último intervalo debe coincidir con el total de las frecuencias observadas.

Práctica: - Desarrollar el procedimiento expuesto con el ejemplo:

"Variación de precios relativos de 50 artículos de consumo.

10. Interpretación y lectura de la serie de frecuencias obtenidas.

Con las observaciones de la tabla número uno, ordenadas en la serie número tres, se puede saber:

1) Cuántos valores o casos observados son menores que un determinado valor de la variable.

2) Entre qué valores, mínimo y máximo, ha variado la variable.

3) En qué valores de la variable se registraron las frecuencias mínimas y máxima.

4) Qué valores de la variable no se presentaron (frecuencia cero).

5) Qué importancia relativa, dentro del conjunto tiene cada intervalo. Ej.: El 50% de los artículos aumentó sus precios alrededor de....%

6) Cuántas observaciones se hicieron para "valores menores que..." Ejemplo... artículos que sufrieron un aumento de menos del 30%.

7) Si los valores se distribuyen a lo largo de todo el intervalo de variación o si se concentran en algunos puntos más que en otros;

8) Qué magnitud tiene esa concentración alrededor de un valor central.

9) Cuál es la forma de la distribución simétrica o asimétrica?

DISTRIBUCIONES DE FRECUENCIAS CON INTERVALOS IRREGULARES Y/O CON EXTREMOS ABIERTOS.

En problemas como los ya mencionados de carácter demográfico, laboral, sanitario, educacional, económico y social en general, se presentan muy a menudo series con intervalos desiguales y/o con extremos abiertos.

En tales casos, los puntos medios de cada intervalo son diferentes entre sí, por lo que debe aplicarse una gran precaución en el tratamiento e interpretación de estas series.

Para tener un cuadro válido de la distribución, es necesario efectuar un reajuste de los datos. Ello se logra: "dividiendo el número de frecuencias de cada intervalo, por el módulo". La cifra resultante representa el número de frecuencias que corresponde a cada unidad del módulo.

1. - PACIENTES AFECTADOS POR CANCER

Edad	Pacientes	Número de pacientes por años de edad
22-30	18	2, 25
30-35	45	9, 00
35-40	79	15, 80
40-55	225	15, 00
55-60	63	12, 60
60-70	45	4, 50
70-90	13	0, 65

2. DISTRIBUCION DE FAMILIAS DE EE. UU.
POR NIVEL DE INGRESO EN 1950

Nivel de Ingreso	Familias (en miles)	Ingreso medio ponderado (en dólares)
menos de 1.000	3.704	501
de 1.000 a 1.999	7.328	1.524
2.000 a 2.999	8.044	2.504
3.000 a 3.999	8.463	3.494
4.000 a 4.999	6.980	4.472
5.000 a 7.499	8.484	6.035
7.500 a 9.999	2.860	8.468
10.000 y más	2.727	17.377
TOTAL	48.590	4.461

En este caso la tercera columna muestra el promedio -previamente calculado- y no el punto medio de cada intervalo. Las cifras de esta tercera columna son las verdaderamente representativas. Así deben presentarse las series con intervalos abiertos o desiguales.

3.2 - Representación Gráfica:

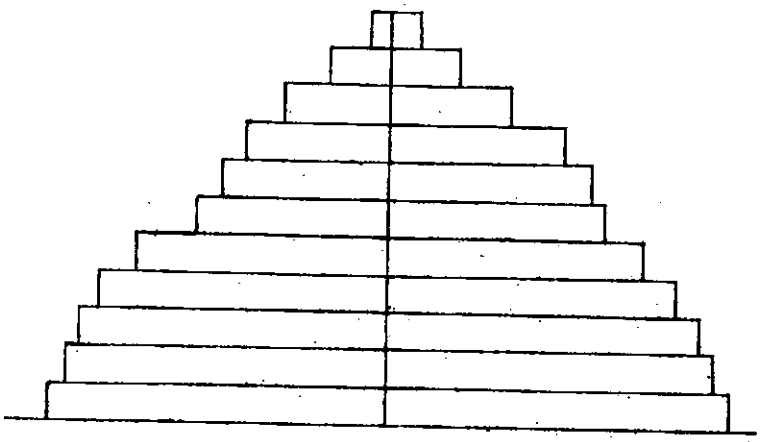
Concepto: Los gráficos tienen por objeto ilustrar mediante el uso de figuras, el desarrollo de un fenómeno estadístico.

Consideraciones: De acuerdo a lo antes enunciado, vemos que la representación gráfica es muy importante para la rápida comprensión del fenómeno que se debe estudiar, ya que se sintetiza una serie de datos numéricos complejos y de largas cifras.

Es menester señalar que el gráfico precede y luego sigue el análisis estadístico dado que, primeramente ofrece un aspecto general del problema, el que una vez estudiado y analizado es representado en sus conclusiones.

Para sintetizar más lo dicho, podemos dar un ejemplo:

1o.) Con los datos del último censo se construye una pirámide de población:

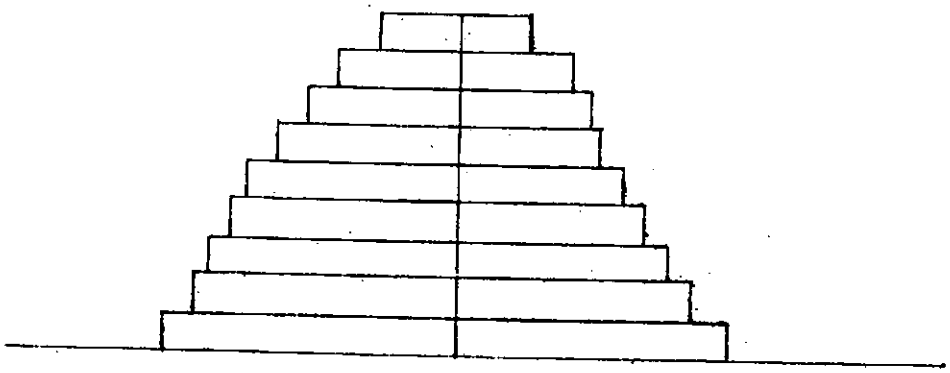


1960

Este gráfico permite tener una rápida visión de la población actual que, de ser presentada con las cifras solamente, sería difícil de interpretar.

2o.) Mediante métodos analíticos y sistemáticos se proyecta la estructura de la población para 1980.

3o.) Mediante un nuevo gráfico se representa la población actual y la calculada para 1980.



La representación gráfica tiene diferentes formas de presentación, mediante puntos líneas, curvas, barras, áreas, volúmenes, etc. Para la confección de los gráficos se debe tener en cuenta los elementos científicos y artísticos que conjuntamente deben obrar para la justa y rápida comprensión.

II)- Clasificación de Gráficos:

- Puntos
  - Mapas
  - Areas
- Líneas
  - Rectas
    - Barras
      - Simples
      - Compuestas
    - Poligonales
  - Curvas
    - Curvas de frecuencia
    - Ojivas
- Superficie
  - Rectángulos (Histograma)
  - Círculos
- Volumen
  - Dibujos convencionales
  - Esterograma
  - Figuras (Pictogramas)

III)- Escala:

Definición : Es una relación aritmética en la cual el denominador es la cantidad a representar (magnitud lineal), y el numerador la longitud del segmento que la representa.

En las escalas lineales, la unidad de medida del numerador y del denominador será la misma debiendo quedar en consecuencia indicada en la escala, solamente la relación de las mismas.

Si tenemos que representar un objeto de la realidad que tiene dimensiones que no es posible reproducir en un papel de tamaño oficio, por ejemplo, es menester recurrir a la escala. Por medio de ella se establece convenientemente de acuerdo a las necesidades, qué medida de papel equivale a la de la realidad.

Ejemplo:

Si un cm. del papel equivale a 100 cm. estamos trabajando con escala 1 : 100 o sea que hemos reducido cien veces la realidad.

Aplicaciones de la Escala: con relación a la escala hay que tener verdadero cuidado cuando se construyen gráficos y sobre todo usar el sentido común.

La escala vertical se construye mediante la situación de cantidades positivas del origen hacia arriba, y las nagativas del origen hacia abajo. La escala horizontal lleva las cantidades positivas del origen hacia la derecha, y las negativas viceversa, o sea del origen hacia la izquierda.

Observemos también que partiendo del origen los valores positivos se sitúan en orden ascendente, mientras que los negativos en orden descendente. Ya que la mayor parte del trabajo Estadístico se efectúa con valores positivos, el cuadrante que se usa casi siempre es el primero, lo cual ha engendrado la costumbre de indicar la escala vertical a la izquierda del gráfico y la horizontal en lo más bajo del mismo. Si el gráfico es de grandes dimensiones es conveniente exponer además, ambas escalas en la parte superior y a la derecha a la vez.

También debemos recordar que: distancias iguales en cada escala representan magnitudes iguales. Pero como las escalas verticales y horizontales son independientes entre sí, igual espacio entre una y otra, pueden representar valores diferentes.

Escala Logarítmica:

Deficiencias de la escala Aritmética:

Supongamos una cantidad de $ 10.000 colocada para rendir el 20% de interés anual. Podemos resumir las operaciones de los tres años como sigue:

1o.	$ 10.000	tasa anual de aumento	--------
2o.	$ 12.000	" " " "	20%
3o.	$ 14.400	" " " "	20%
4o.	$ 17.280	" " " "	20%

Si quisiéramos estudiar los capitales en valor absoluto, resultaría bien la escala aritmética, pero si quisiéramos observar la tasa de aumento o mejor las proporciones de variación , el gráfico nos daría una impresión errónea, pues para la misma tasa del 20%, el gráfico presenta trechos de curvas cada vez más inclinados. (Representar los datos en las dos escalas y observar las diferencias).

Por lo tanto es evidente que, la escala aritmética es muy eficaz para representaciones de valores absolutos, e ineficaz para la representación de valores relativos.

Imaginemos otro caso. En cierto país la producción de calzado aumenta cada año en un 50%. Un cierto productor aumenta también su propia producción en un 50% anualmente. Admitamos el siguiente cuadro con las producciones en millones de pares de zapatos.

	1er. año	2o. año
Producción nacional.....	100	 150
Productor X.....	60	 90

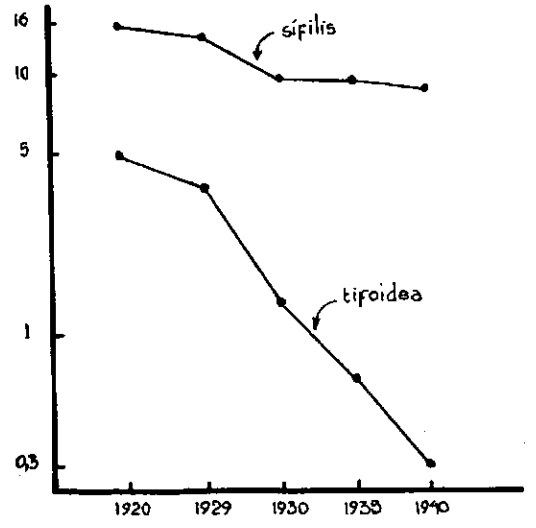
Si observamos los porcentajes vemos que el productor X acompaña perfectamente el ritmo de la producción nacional; ambos progresaron en 50%.

El gráfico de escala aritmética nos informa en cambio que la producción de X es menor en relación a la totalidad de la producción, o sea, que su fábrica no acompañó al ritmo de las demás.

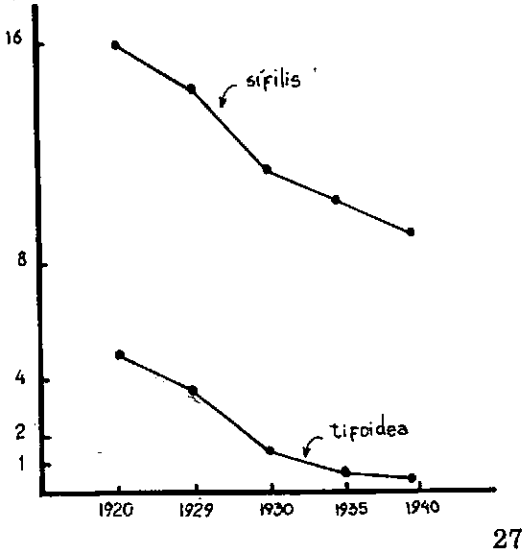
Otro ejemplo: El cuadro siguiente demuestra las tasas de muerte referidas a tifoidea y sífilis en los EE. UU. de raza blanca.

Año	Tifoidea ‰	Sífilis ‰
1920	4,9	16,1
1925	3,7	15,3
1930	1,5	12,1
1935	0,7	11,0
1940	0,3	10,3

a) Escala Logarítmica



b) Escala Aritmética



Análisis de los Gráficos:

Mientras la tifoidea ha experimentado una reducción del 94%, la sífilis se ha reducido en un 35% solamente. Vale decir que, mientras la primera se redujo en 16 veces la segunda sólo lo hizo un poco más de la tercera parte. Si realizamos las representaciones en papel aritmético y semilogarítmico respectivamente, tendremos que el gráfico aritmético demuestra los valores absolutos pero no las proporciones, y que el crecimiento proporcional es evidenciado por el gráfico en papel semilogarítmico que nos da la posibilidad de demostrar la proporción de crecimiento o decrecimiento de un fenómeno. Para este fin también podría usarse a falta de papel semilogarítmico, el aritmético, pero trasladando a él los logaritmos de los datos originales (tasas) obtendremos la misma evolución evidenciada en el papel semilogarítmico.

Construcción del Trazado Logarítmico

A diferencia del trazado común, el logarítmico debe tener muchas líneas horizontales, no sólo por la dificultad de guiarse a simple vista en una escala a la cual no estamos habituados, sino también porque el gráfico tiene una finalidad más técnica expuesta a lecturas más minuciosas que los otros tipos de diseños.

Pero, como en el trabajo estadístico el uso del papel logarítmico o doble logarítmico, que así también se lo llama, es raro y sólo se usa en rectificaciones de curvas, la moderna tendencia es llamar simplemente "logarítmicos" a los gráficos con escala horizontal aritmética y la vertical logarítmica y reservar el nombre de "doble logarítmico" para los diseños ocasionales hechos con dos escalas logarítmicas.

Construcción del Papel Logarítmico

Tenemos 10 cms. para ser graduados de 1 a 10 y hagamos:

log. 1	0,00	0,0 cm.
log. 2	0,30	3,0 "
log. 3	0,48	4,8 "
log. 4	0,60	6,0 "
log. 5	0,70	7,0 "
log. 6	0,78	7,8 "
log. 7	0,85	8,5 "
log. 8	0,90	9,0 "
log. 9	0,95	9,5 "
log. 10	1,00	10,0 "

Tenemos ahora los números 1; 10; 100; 1000 cada uno diez veces mayor que el precedente, pero en el mismo gráfico, la distancia entre 1 y 10 debe ser la misma que entre 10 y 100 o entre 100 y 1000.

De este modo, en tres veces la altura requerida para marcar de 1 a 10, tendremos la presentación de 1 a 1000, lo que es una ventaja inconfundible sobre el papel aritmético.

Para cada múltiplo de 10 que necesitemos, bastará con agregar una nueva sección de 10 cms.

Representación Gráfica Logarítmica:

Cuando representamos gráficamente a escala natural una serie de números, la curva que de ella resulta nos indica que una variación dada representa tantas unidades de incremento o disminución, en cualquier lugar de la escala en que éstos se produzcan. Es decir que representamos por una misma ordenada al incremento de 480 a 500, por ejemplo que al incremento de 20 a 40. Sin embargo, mientras en el primer caso el incremento es sólo del 4%, en el segundo es el 100%. Por ello es necesario usar papel logarítmico para que la curva refleje variaciones relativas y no absolutas.

La curva obtenida en papel logarítmico nos dará un verdadero reflejo de los cambios porcentuales, ya que la diferencia entre los logaritmos de los números será mayor que la diferencia de los logaritmos de otros dos, siempre que la razón entre los números del primer par sea mayor que la razón entre los números del segundo par.

Ventajas de la Representación Gráfica Logarítmica:

a) Representa razones: En un papel logarítmico, una diferencia vertical representa igual razón de crecimiento en cualquier parte del diagrama en lugar de iguales incrementos, como sucede en la escala natural.

b) Facilita la comparación de series cronológicas: La semejanza entre dos curvas logarítmicas significará su correspondencia en las variaciones de proporciones, mientras que la semejanza entre dos curvas representadas a escala natural, significaría correspondencia en las variaciones absolutas cuando estén representadas a igual escala.

Las curvas así representadas pueden ser trasladadas verticalmente hasta obtener la máxima correspondencia entre ellas, después de lo cual podrán ser examinadas. Este artificio no es posible en la escala natural, porque en ella el origen o cualquiera otra parte de la escala, deben siempre coincidir para todas las curvas.

c) Aplicación de los Index Numbers: El uso de la escala logarítmica es importante cuando los números originarios representan razones.

Suponiendo que 100, 75, 50, son los números índices de los precios de 3 años en la escala natural, los decrecimientos estarán representados por iguales ordenadas. Sin embargo, la disminución en el primer caso es de 25%, mientras que en el segundo es de 33%. Estos porcentajes se reflejarán exactamente si representamos los números índices en logaritmos.

d) Se pueden representar valores grandes y pequeños: Se pueden representar fácilmente valores tan pequeños como 20 a 50 y valores tan grandes como 20.000 y 50.000, lo cual sería imposible en papel aritmético.

Limitaciones al uso de la escala logarítmica:

1) Esta escala se debe usar cuando las oscilaciones de la curva son muy grandes, pues poco se ganará con este método si el punto máximo de la serie no va más alla del 100% del mínimo, dado que a medida que las oscilaciones van disminuyendo a 60%, 30%, 15%, la diferencia entre ambas escalas se va haciendo tan poco notable que ya no se justifica su uso.

2) Este método se puede usar sólo para representar aquellas series cuyos valores no pasen de ser positivos a negativos o viceversa, puesto que cero es el límite absoluto de toda progresión geométrica y de todo lo que opere bajo las reglas del crecimiento orgánico, como lo es la población de un país que nunca puede ser cero, y nunca negativo.

Interpretación de los gráficos logarítmicos:

No tiene línea de cero, luego, la escala podemos empezarla en el punto que nos resulte más conveniente, al contrario de la escala común.

El gráfico logarítmico objetiva las variaciones en proporciones, pues como ya sabemos, distancias verticales iguales representan siempre la misma proporción, o sea la misma tasa de variación.

Significado de las diferencias de ordenada en escalas logarítmicas:

En la escala logarítmica la longitud de los segmentos a que dan origen los valores, son proporcionales a sus correspondientes logaritmos y por ello se encontrarán representados a la misma distancia aquellos valores que presenten la misma relación por cociente.

Como $\frac{a}{b}$ indica una relación.

Aplicando logaritmos: $\log. \left(\frac{a}{b} \right) = \log. a - \log. b$

En el gráfico tendremos los puntos A = log. a; B= log. b, en consecuencia la diferencia de ordenadas que hay entre A - B, es lo mismo que $\log. \left(\frac{a}{b}\right)$ o sea, el logaritmo de la proporción A respecto a B.

IV) - Elementos de un Gráfico. Condiciones:

1) Título:

Todo gráfico como todo cuadro, deberá tener un título que exprese clara y suscintamente lo que la gráfica se propone demostrar. El título de una gráfica impresa puede aparecer en la parte superior o en la inferior de la gráfica. El título debe expresar claramente la clase de datos representados, el período y área a que se refieren.

2) Fuente:

Además, como en el caso de un cuadro, cada gráfica deberá contener una referencia a la fuente, indicando el autor, título, volumen, página, editor y fecha de la publi

cación de donde se tomaron los datos, o el nombre de la institución que proporcionó la información.

3) Referencias o notas:

Cuando sea necesario hacer aclaraciones con respecto a los datos consignados (por ejemplo, señalar que un valor corresponde a una estimación y no algo verdaderamente observado, o que otro dato es poco confiable, que en determinado caso se ha utilizado un criterio de clasificación distinto, no incluyéndose a determinado grupo de edad.

4) Escalas:

Tanto para la línea cero como para los trazos que unen cada dos puntos consecutivos, deben emplearse líneas más gruesas que las que presentan las coordenadas de dichos puntos. Los números que acompañen a las escalas se escribirán al pie y a la izquierda del gráfico, pudiéndose repetir a la derecha de éste, los de la escala vertical para facilitar su lectura. La disposición de los números debe ser tal que puedan ser leídos desde el origen del gráfico hacia arriba, en el eje vertical y desde ésta, hacia la derecha en el otro eje.

5) Algunas reglas universales:

La comisión conjunta de Standards para la Representación Gráfica designada en los EE. UU. por ciertas instituciones científicas y reparticiones oficiales, formuló 17 sugestiones que exponemos a continuación:

1) "El ordenamiento general del diagrama debe ser de izquierda a derecha". Puede asegurarse que éste es el principio aceptado universalmente, aun en los casos en que el gráfico abarca más de un cuadrante, porque presumiblemente, la línea se ha iniciado con signo negativo.

2) "Dentro de lo posible, conviene representar las cantidades por medidas lineales y no por áreas o volúmenes, pues éstos son más suceptibles de interpretación errónea". Efectivamente, los gráficos de barras o líneas, dan una impresión inmediata más efectiva y exacta.

3) "Para gráficos de curva la escala vertical debe disponerse en lo posible, de modo que la línea no aparezca en el diagrama". Esta norma ayuda a interpretar mejor el gráfico y coincide con la base matemática de este tipo de diagrama.

4) "Cuando el cero no puede aparecer en el dibujo, conviene representarlo al pie; mostrando el gráfico como cortado en su parte inferior". Este principio es complementario del anterior y el caso se presenta muy a menudo cuando se quiere representar porcentajes.

5) "La línea cero debe distinguirse nítidamente de las restantes. También se relaciona con las dos normas precedentes. Puede suceder que el 0 corresponda a la línea inferior, pero sí es posible asimismo, que se halle a cierta altura por la posibilidad de que la línea tenga que representar magnitudes negativas, en este caso se

marcaría con más fuerza que el resto de las abscisas.

6) "Para las curvas que deben representar porcentajes es usual destacar la del 100%". Esta magnitud puede estar a la mitad de la escala o en otro lugar de ella, porque la curva sube o baja alternativamente de 100; al marcar la línea base con más nitidez se facilita la interpretación del gráfico.

7) "Cuando la escala se refiere a fechas y el período representado no es completo, es mejor no destacar la primera y última ordenada. Ya que el respectivo dibujo no representa ni el principio ni el fin del tiempo.

8) "Cuando las curvas se diseñen sobre ordenadas logarítmicas, las líneas que limitan el diagrama, deben corresponder a alguna potencia de 10 en la escala logarítmica".

9) "Conviene no mostrar más líneas coordenadas que las necesarias para guiar el ojo en la lectura del diagrama". Ciertamente, salvo cuando se trata de gráficas destinadas a destacar observaciones intermedias, es preferible dibujar estrictamente, las líneas que corresponden a las cifras que se quieren ilustrar; el exceso confunde la vista y dificulta la interpretación.

10) "La línea curva del gráfico debe distinguirse precisamente de las restantes". Esto es natural, ya que es esa línea la que da la información que el diagrama quiere representar.

11) "En las curvas que representan una serie de observaciones, es aconsejable siempre que se pueda, indicar claramente en el diagrama todos los puntos que corresponden a observaciones separadas".

12) "La escala horizontal en la gráfica de curvas, deberá usualmente leerse de izquierda a derecha y la vertical de abajo hacia arriba". Es una norma universalmente aceptada, que no hace sino seguir lo que es corriente con la escritura.

13) "Las cifras indicadas en la escala deben colocarse en la izquierda y debajo de los respectivos ejes". Esto es, fuera del gráfico propiamente dicho, lo que facilita su lectura.

14) "A menudo es útil incluir en el gráfico la información numérica o fórmula representada". Lo corriente es que la escala señale cifras límites: 1-20-30.....100 por ejemplo; pero no fracciones de la serie. Los datos numéricos en cambio, no corresponderán exactamente a aquellos límites que podrían ser en aquel caso -1.50-11.30, etc.; apareciendo en el gráfico por los puntos de coincidencia que señala la curva. Indudablemente, la inclusión en el dibujo en la parte superior o derecha de los importes exactos, completa perfectamente la ilustración aunque ello no es siempre posible.

Bibliografía: Mills, Cecil E., Métodos Estadísticos Aplicados a los Negocios, página 31.

VI) - Tipos de Gráficos

1) Pictogramas:

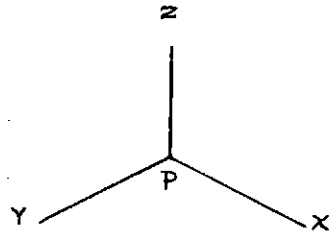
Un tipo de representación gráfica muy usado en las publicaciones de carácter divulgativo, es la de diagramas por medio de figuras. Estos están formados por figuras semejantes entre sí y que por su forma recuerdan el objeto al cuál se refieren (por ejemplo: figura humana, en el caso de que se quiera representar a los habitantes; figura de barril, cuando se quiera representar la producción vinícola; figura de barco, cuando se quiera representar el tráfico marítimo o el tonelaje de la flota mercante;etc.).

El pictograma consiste en un grupo de figuras repetidas, en las que cada figura corresponde a una unidad de medida del fenómeno (mil habitantes o cien automóviles), y se repite tantas veces como corresponda a la intensidad del fenómeno, dando lugar a un conjunto de figuras, todos del mismo tamaño, colocadas unas al lado de otras, cuya longitud o, respectivamente, altura, resulta por lo tanto proporcional a la intensidad del fenómeno y arreglado de tal modo que se forme una gráfica de barras.

2) Esterogramas:

A base de coordenadas cartesianas en tres dimensiones:

Cuando la serie contiene datos referidos a tres fenómenos, se hace preciso construir diagramas utilizando un sistema de coordenas cartesianas de tres variables. Una combinación de dos números reales da origen a dos valores: X e Y que permiten situar un punto en el plano. Un tercer valor determinará un segmento (Z) perpendicular al plano Y que establecerá la posición de un punto en el espacio. Las coordenadas del punto P son X e Y, Z.

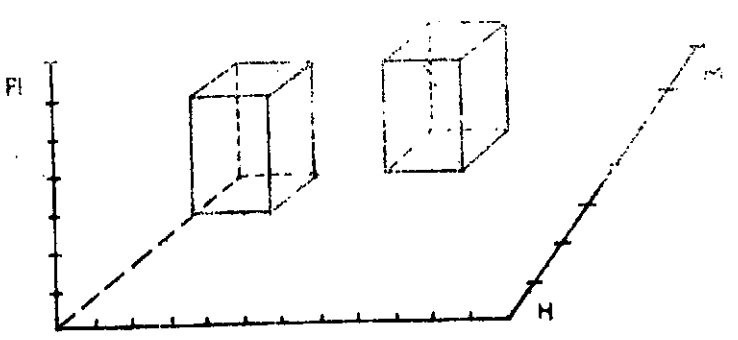


Puede suceder igual que en el caso de dos fenómenos que coincidan varios puntos, sobre todo si el número de observaciones es muy elevado, pues éste no es sino una variante del mismo tipo de diagrama.

Cuando se dispone de los datos de dos fenómenos expuestos en tablas de dos entradas, el diagrama corpóreo puede adoptar dos formas distintas que se corresponden en el histograma y el polígono de frecuencia. Se debe esta correspondencia, a la circunstancia de tratarse de distribuciones de frecuencias combinadas.

La primera clase de ellas se denomina esterogramas y en figura consistente en un conjunto de paralelepípedos. Se construye de la siguiente manera: en los ejes X e Y se sitúan los intervalos en que se ha dividido la variación total de los dos fenómenos, y so-

bre el cuadrado (o rectán[illegible]) correspondiente a cada [illegible] se levante un paralelepípedo de altura proporcional a la frecuencia [illegible] sión. Si los intervalos [illegible] serán todos iguales, la altura que [illegible] cociente resultante. Si dividimos la frecuencia por el producto de la amplitud de [illegible] intervalo correspondiente; en caso de que el volumen de [illegible] nueva homogeneidad.



Bibliografía: García Alvarez y Ayuso Orejana. Estadística.

3) Histogramas o diagrama de columnas:

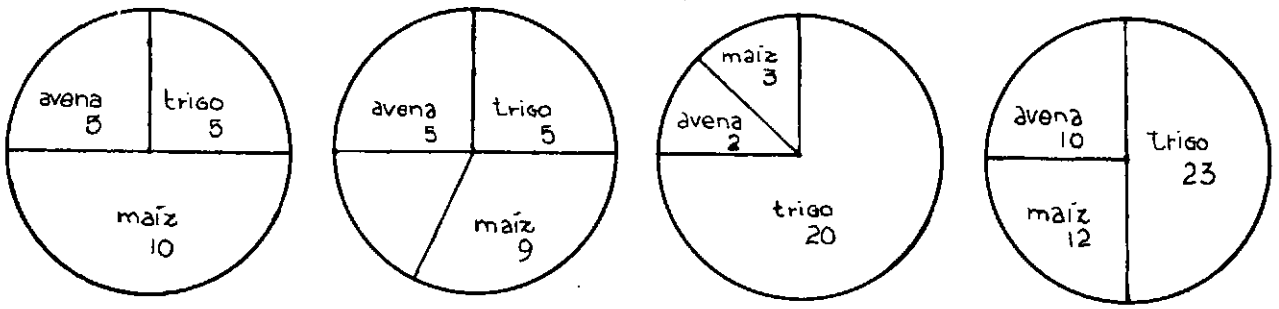
Si queremos representar por ejemplo, de variable contínua (ingreso, pesos, talla, etc.) el número de familias según el ingreso de que disponen, la intensidad correspondiente a cada categoría de ingreso, es la frecuencia que posee la clase misma. La figura se representa como una serie de rectángulos que tienen sus bases en una misma recta y adyacentes y se llama Histograma. La frecuencia es proporcional a la superficie de rectángulos respectivos y esta representación es la que se llama diagrama de columna o Histograma.

Si las bases del rectángulo son iguales, la frecuencia es entonces proporcional a la altura de cada rectángulo. Pero si las bases son desiguales, esto es si los intervalos de clase tienen diversa extensión, esto no sucede, y es necesario homogeneizar la frecuencia que corresponderá a cada unidad del intervalo. Si tenemos la frecuencia total para obtener la frecuencia media de cada punto, se dividirá la frecuencia total de cada intervalo por su respectiva amplitud. A este cociente lo denominaremos frecuencia homogeneizada que designamos con F'. Los valores D, F', darán la altura de cada rectángulo.

Bibliografía : Conrado Gini: Estadística, pág. 68.

Gráficas circulares:

Esta forma de gráfica se usa para variables que se han dividido en partes, como por ejemplo la cosecha de granos de diferentes Estados. Dibujamos círculos cuyas áreas son proporcionadas al total de las cosechas en los diferentes Estados. Esto implica que los radios se hagan proporcionales a las raíces cuadradas de las cosechas totales de cada Estado. La circunsferencia de cada uno de los círculos se divide en segmentos proporcionales a la cantidad de diversos granos que se cosecharon. Después, se trazan radios que unan los puntos marcados sobre la circunsferencia. Esto hace que el área incluida en cada sector circular sea proporcional a cada cosecha.



Esta gráfica circular no es útil para comparar las cosechas entre los diferentes Estados entre sí.

Uso de las Gráficas Circulares:

Estas son muy usadas y con éxito, cuando se pretenden exhibir datos. Tienen la ventaja de llamar la atención al público porque atraen la mente popular. El ejemplo vívido es el de la gráfica que indica cómo la gente gasta su dinero. Qué parte del peso ganado se emplea para comprar alimentos, qué parte para pagar Rentas, etc. Las monedas tienen forma circular, así también la gráfica.

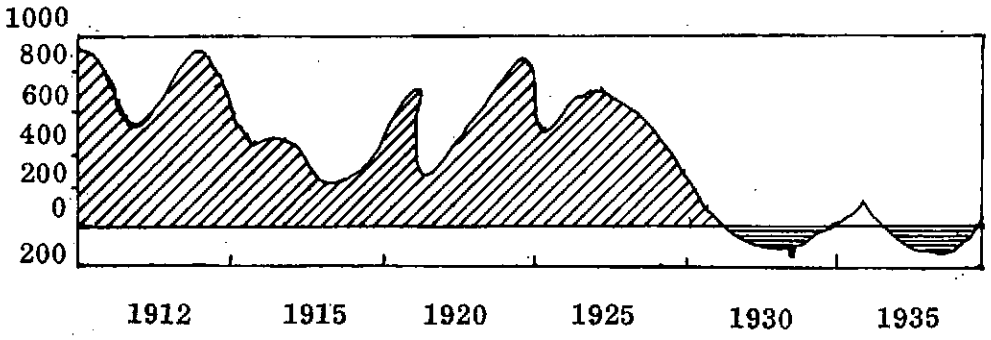
Pero no es un gráfico recomendable para estudios más rigurosos, pues si los sectores son de superficie más o menos similar, no es fácil apreciar exactamente las diferencias existentes entre ellos. El uso de estos gráficos se reserva para demostraciones populares.

Gráficos de Siluetas:

Son diagramas que representan saldos netos de dos series. Para cada uno de los años se sustrajo la emigración total de la emigración total, y el resultado se trazó como una cifra positiva o negativa. El saldo comercial (valor de las exportaciones menos valor de las importaciones) puede mostrarse de la misma manera, lo mismo que las ganancias y pérdidas. En la gráfica se ilustran un método alternativo para mostrar los datos de migración. En este caso serán las curvas de inmigración y de emigración y el exceso de aquellas sobre éstas, se indica por medio de la altura del área sombreada, en tanto que el exceso de emigración se muestra por la altura de la porción negra.

La gráfica ilustra no sólo la representación de las cantidades netas más bien que las cantidades brutas, sino también la práctica de sombrear el área entre las curvas a fin de hacerlas destacar. Se representan las fluctuaciones por encima y por debajo de la línea base. El resultado es una representación más llamativa de los "máximos" y "mínimos" de la curva. Las gráficas de este tipo son aun más eficaces cuando se pintan de negro las áreas "positivas" y de rojo las "negativas".

INMIGRACION NETA EN MILLARES



Bibliografía: Estadística Aplicada, Croxton y Cowen.

Gráfico de Barras:

El gráfico de barras horizontales es la forma más simple de comparar diferentes

ítems a fecha determinada. (Datos clasificados cuantitativamente o geográficamente). Es muy sencillo para su construcción y su comprensión por parte del público.

Barras Verticales:

Es muy sencillo para su construcción y comprensión por parte del público. Las barras se originan a la derecha de una línea base común y se miden por unas pocas marcas de escalas horizontalmente colocadas.

Espacios entre Barras:

Comúnmente, el espacio entre las Barras deben ser de la mitad del ancho de las barras. Este espacio no debe ser igual que la superficie de las barras.

Referencias de las Barras:

Los valores numéricos deben insertarse en el extremo a la derecha de cada barra; de este modo se puede omitir la escala vertical y se facilita la lectura del gráfico (sobro todo, si se utiliza en conferencias).

Si las barras no son lo suficientemente anchas para incluir en ellas los números, se indicará su referencia a la izquierda de la línea cero.

Rotura de Barras:

Como regla general, las barras no deben cortarse, de lo contrario se da una impresión falsa del gráfico. No obstante, si la longitud total de cualquier barra no es esencial al gráfico en general, puede cortarse cerca del extremo derecho, dejando un espacio suficiente para colocar en él al número que representará a la barra total.

Importancia Relativa:

Sombreando una porción del conjunto de divisiones, servirá para facilitar la lectura (ejemplo: se sombrea la parte correspondiente al 100% o al promedio que sirve de base para comparar las barras restantes).

Ordenamiento de las Barras:

Las barras se pueden ordenar de distintos modos, según el propósito: Orden numérico: cuando se desea mostrar un "Ranking" o posición relativa de los distintos aspectos, las barras se ordenan numéricamente de mayor a menor, pudiendo destacarse así, cuáles están por encima del valor típico (total, promedio, etc.) y cuáles por debajo.

Orden alfabético

En el arreglo alfabético, las barras aparecen irregularmente, pero cada unidad(geo

gráficas, actividad industrial, etc.) puede ser localizada fácilmente sobre todo cuando se incluyen muchas barras.

Orden progresiva:

Es el conocido gráfico de Gant que se usa cuando se quiere demostrar el proceso de un plan(de trabajo, gastos, etc.) a través del tiempo.

Orden Cronológico

Cuando se quiere mostrar datos de unos pocos períodos de tiempo, las barras se pueden arreglar cronológicamente. No se usará este gráfico si el período es largo, en ese caso es mejor el gráfico de líneas o de columnas.

Barras Cualitativas:

Un material estadístico clasificado cualitativamente, puede ser bien presentado en gráficos de barra, colocando las referencias a la izquierda de la línea cero (ejemplo, causa de muerte).

Barras Agrupadas:

Los gráficos de barras agrupadas compara un grupo de ítems en dos o tres aspectos. El número de barras debe limitarse a tres para evitar confusiones. Estos gráficos son útiles para ser comparaciones con un promedio o un total, o una categoría especial de datos. La secuencia de los grupos, así como el arreglo dentro del grupo, usualmente se determina por el valor numérico de mayor categoría arreglado en orden descendente. Una vez dispuesto esto, para el primer grupo se debe mantener el mismo ordenamiento para los grupos restantes. El sombreado de las barras debe ser muy distinto para distinguir por contraste cada característica. El espacio entre grupos debe ser por lo menos igual al ancho de una barra.

Barras Subdivididas:

Se usan para representar distintos componentes de un total. Mientras menor sea la cantidad de segmentos, más fácil será la lectura del gráfico. La única lectura verdadera que puede hacerse es respecto a los segmentos adyacentes a línea cero. Por lo tanto, la componente más importante debe graficarse primero. Si se usan barras agrupadas y subdivididas simultáneamente, debe limitarse el número de barras. Nc son gráficos útiles para todo público.

Barras de Porcientos:

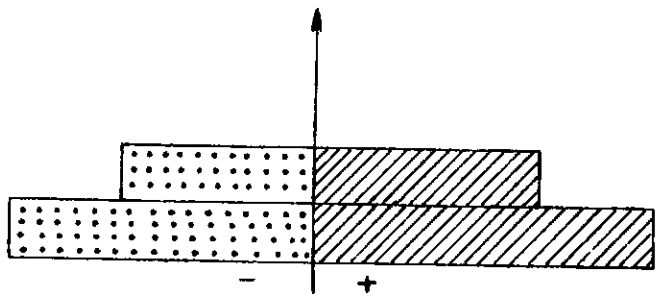
Estas barras tienen todas la misma longitud y son divididas en segmentos que representan la distribución en porciento en cada categoría. Son útiles cuando los porcientos parciales son pocos.

Barras Apareadas:

Se usan cuando es necesario comparar un numero de ítems en dos aspectos, aunque estén medidos en diferente escala. Los pares de barra se forman oponiendo una o otra a derecha e izquierda de la lista de característica (ejemplo, personal ocupado y salarios pagados por distintas actividades). También se utilizan para representar dos variables en una serie de tiempo.

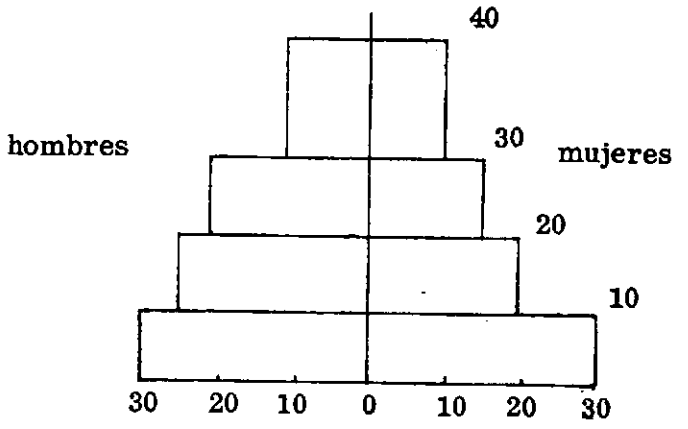
Barras de Desvíos:

Se usan cuando se desea representar saldo positivo y negativo, aumentos y disminuciones, pérdidas y ganancias o hechos similares. Las escalas se extienden a ambos lados a partir de una línea.



Triángulos de Población:

Diagramas de barras: Los sencillos diagramas de barras se adaptan para representar datos de clasificación combinadas, donde una clasificación es cuantitativa y la otra es no cuantitativa o el agrupamiento en intervalos de clases de una variable cuantitativa. Las longitudes de las barras son proporcionales a la magnitud de las características, mientras que las distintas barras representan las categorías o intervalos de clases de la otra característica. Hay varios grados de complejidad en diagramas de barras. Las barras son horizontales cuando las diferentes barras representan la categoría de una clasificación no cuantitativa. Las barras son verticales cuando cada barra representa un intervalo de clase de una característica cuantitativa. La anchura de las barras y el espacio entre las barras no tiene significado estadístico. Los espacios entre las barras no deberán ser del mismo ancho que las barras, para evitar la confusión si las barras no se sombrean. La escritura vertical se suprimirá para evitar el tener que dar vuelta el diagrama para leer.



Las pirámides de población son un ejemplo de diagramas de barras bilaterales con barras contiguas.

Bibliografía: Estadística para Sociólogos, Margaret J. H.A.G.O.O.D. Pág. 43.

Triángulo de Población:

Es una variante de los gráficos de barra yuxtapuesta que se utiliza para el estudio de la estructura de población. Las barras se ubican a ambos lados de un eje central donde puede leerse en escala, los grupos de edades de los habitantes. A la derecha de dicho eje, se dibujan las barras que representan en proporción la cantidad de mujeres correspondientes a cada grupo de edad. A la izquierda, las barras reflejan la cantidad de varones. Cuando se desea hacer comparaciones entre la estructura de dos o más poblaciones, conviene representar dichas cantidades en forma de porcientos para cada edad. Estos gráficos también se conocen con el nombre bastante generalizado de pirámide de población. Las características de una población joven se reflejan en un triángulo de base amplia que gradualmente va disminuyendo por efecto de la mortalidad, hasta llegar al vértice con muy escasos habitantes, en las edades muy avanzadas.

Análisis de algunos triángulos de población:

Gráfico número uno : Población autóctona de Tananarive (Madagascar), la población de Africa negra y Madagascar puede considerarse como población joven. Las tasas de mortalidad son muy elevadas alrededor del 20% (Mendoza, en cambio, tiene un 8‰), por ello se reduce bruscamente la población. La esperanza debida es por lo tanto, muy débil: Escasamente de treinta a treinta y cinco años (en Mendoza, es superior a los 64.

La tasa de natalidad es también muy alta 40‰, por lo cual, este triángulo presenta una base tan ancha (en Mendoza la tasa de natalidad es de sólo 26‰). La falta de fuentes estables de trabajo, provocan muy fuertes migraciones de todo tipo: interregionales, estacionales o definitivas. Se considera que alrededor de 1/3 de los hombres entre los 15 y 40 años están permanentemente afectados por la migración. Las barras salientes que se observan en este triángulo, reflejando mayor cantidad de personas de una edad, respecto a la edad anterior, es una representación típica de: Poblaciones que reciben o dan fuertes contingentes migratorios, o bien, representa una falla estadística ya supera

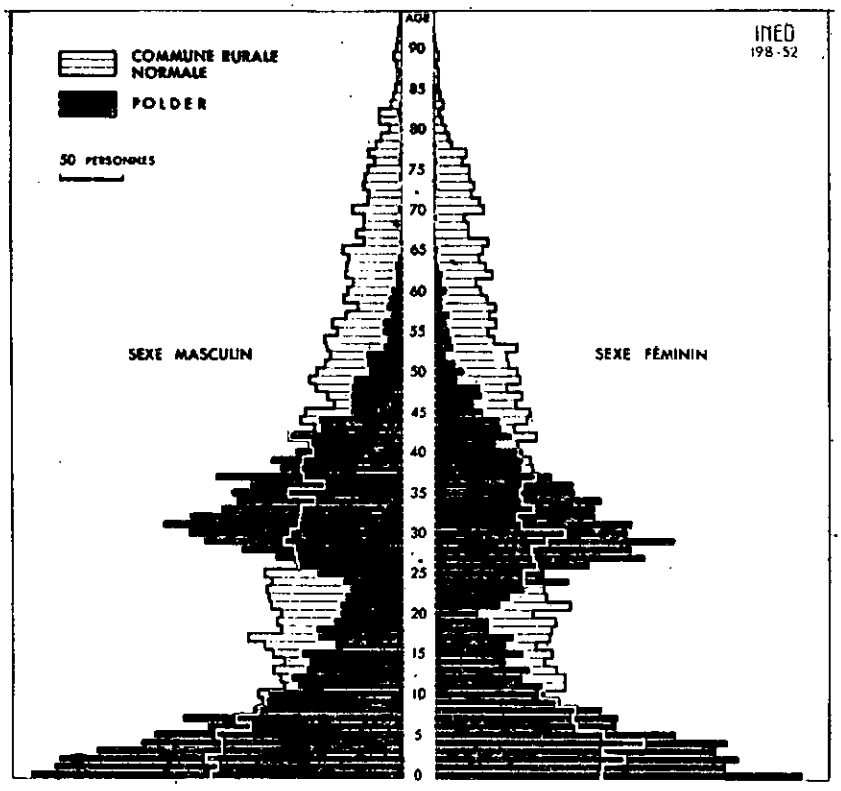


gráfico 4

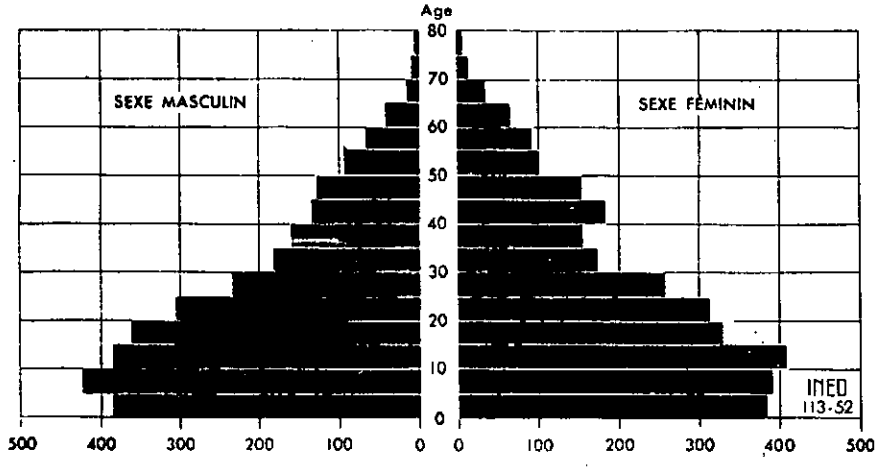


gráfico 5

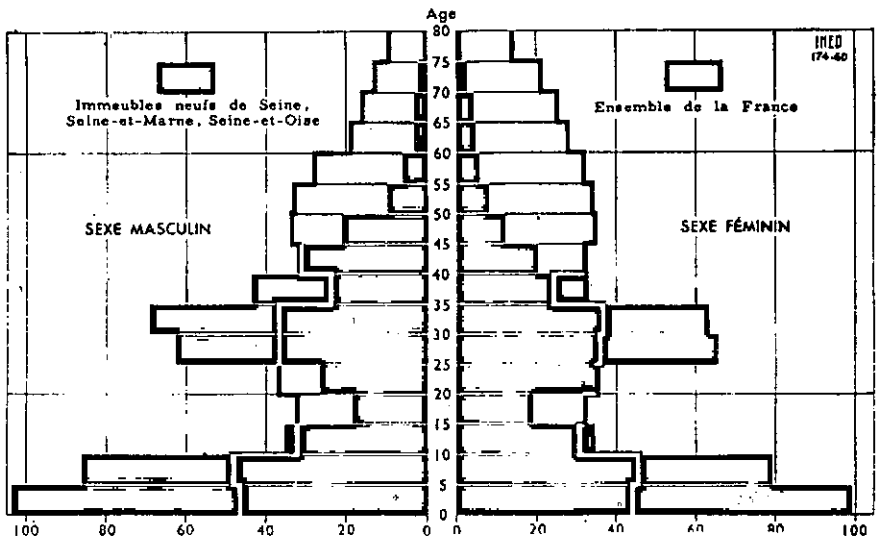


gráfico 6

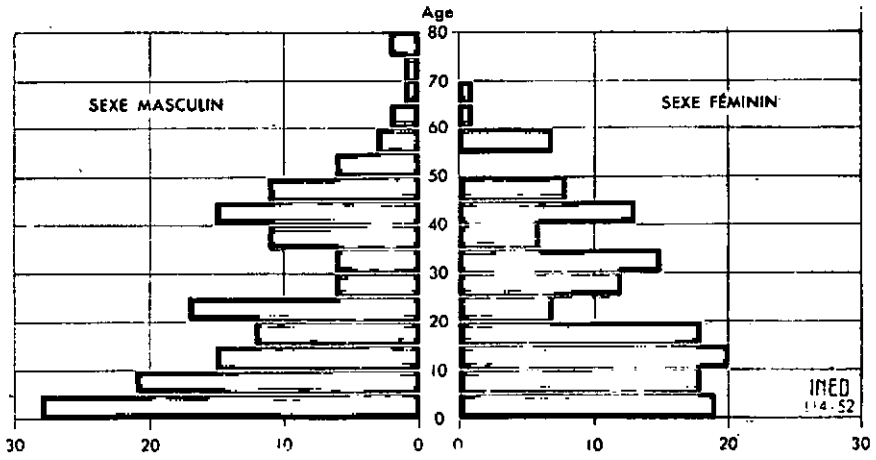


gráfico 7

do en los censos más modernos. Ocurre que si se pregunta: Qué edad tiene Usted?, muchas personas manifiestan una tendencia sistemática a redondear su edad, expresándose en números terminados en ceros o en cinco. De ahí que, las barras salientes se ubiquen frente a estos números.

Gráficos número dos y tres: Población de Francia y de la U. R. S. S. En el gráfico tres se ha representado en porciento la frecuencia relativa de cada grupo de edad. En la población de Francia y Rusia puede observarse para Francia una base menor que indica menor natalidad que en Rusia. Esta base estrecha, de dimensión muy próxima a los contingentes de adultos, hace tomar al triángulo una forma conocida como de "urna funeraria" que caracteriza a las "poblaciones viejas". Ambas representaciones presentan dos muescas correspondientes a las brechas por mayor mortalidad y menor natalidad como consecuencia de las guerras. En cuanto a la población mayor de los 50 años, se observa mayor proporción en Francia, lo que nuevamente confirma su característica de "población vieja"(escasa proporción de jóvenes, gran proporción de adultos y ancianos). El perfil demarcado por la mortalidad natural, después de los 50 años, es más agudo para el sexo masculino en los dos países, lo que confirma la mayor mortalidad masculina en esas edades.

Gráfico cuatro: Población del Polder del Nordeste y una comuna rural de estructura normal en Holanda (*). La población del Polder del nordeste en Holanda, es extremadamente joven debido al sistema de selección de los colonos que allí se radican, impuesto por el Gobierno. Los candidatos no deben ser mayores de 50 años, deben tener capacidad técnica, posibilidades financieras y aptitud de pionero. Al 1o. de octubre de 1950 la repartición de edades era;

Grupos de edades	Polder	Población total de los países bajos
0-15	43, 2	31, 0
16-20	4, 5	8, 4
21-24	3, 9	6, 7
25-39	36, 9	21, 8
40-49	8, 4	12, 1
50-64	2, 8	12, 9
65-y más	0, 3	7, 1
	100, 0	100, 0

El gráfico representa superpuestas, la población del Polder y la de una comuna rural considerada como normal y con el mismo número de habitantes.

(*) Populatión, oc. dic. 1952.

Gráficos números cinco y siete:

Esquimales de Groenlandia y de Thule. Pirámide obtusa de gran base, característica de población de gran natalidad y fuerte mortalidad. La sobre mortalidad femenina observada en Thulé no parece existir en los otros distritos de Groenlandia. La forma extremadamente dentada del triángulo de la población de Thulé se debe al hecho de que los datos fueron obtenidos por estimaciones. Debido a las condiciones del ambiente no se dispone allí de un curso completo.

Gráfico número seis: Población de ciudades nuevas (*):

Construir la Curva de Lorenz correspondiente a la capacidad de elaboración en la industria vitivinícola de Mendoza existente en 1957.

Intervalos	No. de Bodegas Insc.			Capacidad Inscrip. en miles de Hls.		
0 - 1.000	63	5,10	5,10	34,9	0,2	0,2
1.001 - 5.000	414	33,30	38,40	1.226,6	5,9	6,1
5.001 - 10.000	277	22,20	60,60	2.063,6	9,9	16,0
10.001 - 25.000	275	22,10	82,70	4.658,6	22,3	38,3
25.001 - 50.000	138	11,10	93,80	5.037,0	24,2	62,5
50.001 - 75.000	37	2,98	96,78	2.253,0	10,8	73,3
75.001 -100.000	21	1,69	98,47	1.810,8	8,7	82,0
100.001 -y +	19	1,53	100,00	3.756,8	18,0	100,00
					=	X

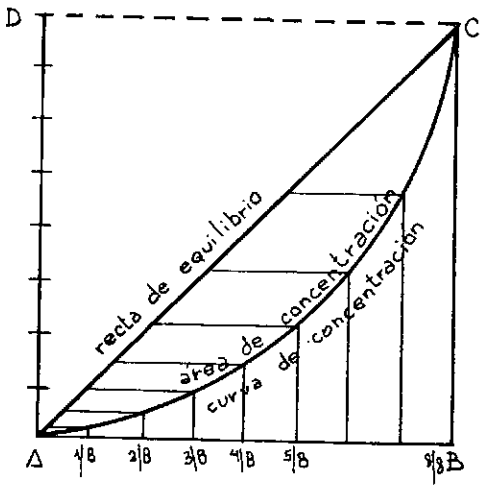
La composición por edades de conjuntos reducidos, rara vez está conforme a la normal. Así se presentan pirámides desplomadas, inclinadas o sobresalientes. Una estructura de edades anormal puede suscitar problemas si no se conserva una movilidad suficiente de los habitantes. La época en que el triángulo demográfico adopta la forma representada en este gráfico, hay muchos niños que necesitan clases elementales. En algunos años más necesitarán enseñanza primaria, secundaria y especial. Las escuelas necesarias a organizar serán muy grandes para sus sucesores. En las ciudades construidas por los empresarios para alojar a su personal habrá un golpe brusco-colapso- cuando los adultos actuales se jubilen. Dentro de un período más o menos corto, será necesario alojarlos en otra parte y encontrar nueva mano de obra que los reemplace.

Curvas de concentración y equidistribución de Lorenz:

La concentración de la riqueza de una determinada población se puede representar sobre un sistema de ejes coordenados. Sobre uno de ellos (eje de las abscisas), representaremos las fracciones de la población y paralelamente al otro (eje de las ordenadas), las fracciones de la riqueza que globalmente posee dichas fracciones de población. Suponiendo

(*) Population, Agost. Set. 1960. Pág. 586

que los individuos están ordenados en razón al calor creciente de las riquezas que poseen, empezaremos por tomar en consideración al primer décimo de la población así ordenada, esto es, al décimo más pobre, que no poseerá una décima parte de la riqueza total, sino una fracción menor que esta cantidad, de manera semejante, los primeros dos décimos de la población tendrán también en total menos de dos décimas partes de la riqueza global, pero más del doble de la que poseía el primer décimo, los primeros tres décimos dispondrán de una riqueza total inferior a los tres décimos de la riqueza global, pero que representará más del triple de la que poseía el primer décimo, y así sucesivamente, hasta llegar a los 10 décimos de la población, los cuales poseerán necesariamente la totalidad o sea los 10 décimos de la riqueza. Si cada uno de los puntos que sobre el eje de abscisas representa el primero, segundo,... y último décimo de la población, levantaremos las ordenadas que representan las fracciones correspondientes de la riqueza total poseída, los extremos de estas ordenadas unidos por una línea continua darán lugar a lo que llamamos una curva de concentración. Esta se extiende de un extremo a otro de la diagonal de un cuadrado, cuyo lado es igual a la unidad, y permenece de manera constante por debajo de dicha diagonal, siendo convexa al eje de las abscisas: Si todos los individuos de la población considerada tuvieran la misma cantidad de riquezas, evidentemente al primero, al segundo, al tercero ... décimo de la riqueza total de forma que la curva de concentración se reduciría a la diagonal, que recibe el nombre de recta equidistribución. Si por el contrario, todas las riquezas se concentraran en un solo individuo todas las ordenadas de la curva de concentración serían nula, excepto la última, que tendría una longitud igual a la unidad, de forma que dicha curva se reduciría a los dos lados consecutivos del cuadrado citado o sea al ángulo A B C que revela en tal caso la situación de monopolio absoluto tanto la recta o línea de equidistribución A C, como el ángulo A B C, son dos situaciones extremas poco frecuentes. En la generalidad de los casos, se obtiene la curva de concentración. Cuanto mayor es el área comprendida entre la curva de concentración y la recta de equidistribución (área de concentración), tanto mayor es la desigualdad existente en la distribución de las riquezas o, como suele decirse la concentración de la riqueza. Si todos poseen la misma riqueza o sea la concentración es nula, el área de concentración equivaldría al área de un triángulo rectángulo cuyo catetos tienen una longitud igual a la unidad, y por lo tanto, es igual a la mitad del cuadrado, en los casos intermedios el área de concentración estará comprendida entre cero y medio (0 y $\frac{1}{2}$). La proporción entre el área de concentración y el máximo de la misma puede alcanzar, o sea $\frac{1}{2}$, podrá, por lo tanto, variar entre 0 y 1, dicha proporción se llama razón de la concentración de la distribución considerada (Gini).



El trabajo se ordena del siguiente modo:

(1o.) - La información básica debe contener:

a) - Cantidad de elementos clasificados según una escala progresiva del atributo a representar X.-

b) - Total del atributo X; que corresponde a los elementos de cada intervalo.

Antes de confeccionar el gráfico se requiere algunas sencillas operaciones previas que son:

1o.) - Calcular el porciento de elementos y el porciento de atributos que corresponde a cada intervalo.

2o.) - Calcular el porciento acumulado de ambas series.

Con estos datos se confecciona el gráfico. Para ello se dibuja un cuadrado, cuyos lados serán equivalentes al 100 % de cada serie.

En el lado correspondiente a la abscisa se indicarán los porcientos acumulados de la serie de elementos y en la ordenada los porcientos acumulados del atributo.

Con cada par de valores acumulados se determinarán puntos dentro del cuadrado uniendo con línea continua dichos puntos se obtiene la curva de concentración.

Punto 5:

Gráfico de líneas: Polígonos y Curvas de Frecuencias - Ojivas.

Tipos de gráficos (Tomado del libro: Spear, Marg, Charting Statistíes (N. Y., 1952).

Los más comunes son:

- de líneas
- de superficie
- de columnas
- circulares
- cartogramas
- pictogramas

Gráficos de líneas:

Es el más ampliamente usado por su construcción simple y claridad de interpretación.

Los puntos representados según las coordenadas cartesianas, se unen con líneas continuas o de puntos.

Las fluctuaciones muestran la variación de las tendencias, la distancia de la abscisas las cantidades.

Oportunidad de su empleo:

Los usos principales son:

1) - Cuando se desea poner énfasis en el movimiento más que en la cantidad.
2) - Cuando se comparan distintas series sobre la misma gráfica.
3) - Cuando los datos cubran un período muy largo de tiempo.
4) - Cuando una distribución de frecuencia se presenta con dos o más curvas componentes.
5) - Cuando se utilizan dos escalas en un mismo gráfico.
6) - Cuando se muestran estimaciones, predicciones, interpelaciones o extrapolaciones.

Ubicación de los puntos:

Los puntos que representan cantidades, se colocan sobre las líneas o entrelíneas?

En general los datos de primero a fin de mes, o totales mensuales o anuales se ubican sobre la línea.

Si los datos corresponden al promedio de un intervalo de tiempo o de frecuencia los puntos se colocarán entre las líneas verticales. Pero en las siguientes excepciones los datos de período se ubican sobre la línea.

Cuando los datos cubren un período largo (5 ó más años). En este caso, no tiene mayor importancia si los datos se ubican "entre" o "sobre" las líneas verticales.

Cuando se dibujan 2 ó más curvas con intervalos diferentes (mes, trimestre, año).

Cuando los datos son acumulativos.

Cuando se dibuja listogramas.

Gráficos de índice:

Se representan como gráficos de líneas. Se usa principalmente: Para comparar dos o más series que difieren grandemente en sus magnitudes (ejm. precios por toneladas y precios por kilos).

Cuando se quiere mostrar la relación de dos o más series expresadas en distintas

unidades (ejm. toneladas y precios). Las relaciones de cambio entre fenómenos de distintas escalas de unidades (miles de toneladas y millones de precios), se muestra más claramente convirtiendo los valores a índices.

Curvas múltiples:

Si se dibujan varias curvas en un gráfico, se pierde claridad y las líneas no se pueden identificar. Deberá confeccionarse dos o más gráficos en estos casos. Estos gráficos estarán relacionados entre sí, por ello, para posibilitar su comparación, las medidas de escala, deben ser las mismas. Esto se facilita dibujando los índices en vez de las cifras absolutas.

Polígonos de frecuencia:

Para trazar un polígono de frecuencias mediante el uso de las coordenadas rectangulares:

1) - Fórmase con los datos de una distribución de frecuencias o una distribución y frecuencias.

2) - En caso de tratarse de una distribución de frecuencia tómase como abscisas (o sea, como distancias al origen medidas sobre el eje horizontal), los valores de los datos; en caso de tratarse de una distribución de clases y frecuencia, tómese como abscisas los puntos medios o marcas de clase. En tales puntos levántese las perpendiculares correspondientes.

3) - Ya sea que se trate de una simple distribución de frecuencia o de una serie de clases y frecuencias, tómese como ordenadas (o sea como distancia al origen medidas desde el eje vertical), las frecuencias correspondientes. Levántese las perpendiculares al eje vertical, y

4) - Encuéntrese la intersección de cada una de esas perpendiculares con la levantada en el punto correspondiente al valor del dato del cual es la frecuencia.

5) - Unase los puntos así encontrados por medio de líneas rectas, a fin de obtener una línea quebrada que representa la distribución de las frecuencias y se conoce como polígono de frecuencias.

Lectura de un polígono de frecuencia:

1) - Tómese un punto del mismo.

2) - Mídase sobre la perpendicular (bajada de punto al eje), su distancia al eje horizontal y léase como primer elemento de la oración "hay tantos casos" (aquí la distancia medida)" "para las cuales la magnitud del fenómeno es de"...

3) - Léase sobre el eje horizontal la distancia entre el pie de la perpendicular y el origen y complétese la oración anterior "tanto" (aquí distancias recién medidas).

De esta forma, si se cuenta con un polígono de frecuencias como única fuente de observación y se desea saber cuántos casos de la distribución tiene una medida X dada:

a) - Se busca el valor de X sobre la horizontal (al eje horizontal).

b) - En ese punto se levantará una perpendicular, y

c) - Se medirá la distancia entre el pie de la perpendicular y la intersección de la misma con el polígono de frecuencia; dicha medida será la frecuencia correspondiente o sea, el número de casos que en el conjunto estudiado corresponden a esa magnitud del fenómeno.

Como es fácil comprender, el trazo de un polígono de frecuencias no es sólo útil para retratar la forma de distribución e incluso para tener una ideal del cuál pudiera ser su distribución teórica en caso de que se pudiera eliminar errores accidentales etc., sino que también permite tener una idea del cuál podrá ser la frecuencia con que se presentarán magnitudes intermedias del fenómeno que la agrupación de una serie de clases y frecuencias subsume bajo una misma denominación de magnitud (la magnitud del punto medio).

Los polígonos de frecuencias sirven también para comparar dos o más distribuciones, especialmente si se trata de polígonos para los cuales las frecuencias sean frecuencias relativas. Una frecuencia relativa se encuentra dividiendo la frecuencia absoluta correspondiente entre la suma de todas las frecuencias.

$$f = \frac{f}{\Sigma fi}$$

La representación de un polígono de frecuencias relativas sigue la misma rutina que la representación de un polígono de frecuencias absolutas.

Cuando la frecuencia relativa se la multiplica por 100, se obtiene una frecuencia porcentual. "La lectura será del siguiente tipo: "de cada 100 casos hay tanto (aquí frecuencia porcentual u ordenada), en los cuales la magnitud del fenómeno es de tanto (aquí el valor de la abscisa)".

En la representación del polígono relativas o porcentuales a más de los ejes coordenados que siempre deben representarse con trazos más gruesos, la paralela al eje horizontal que corresponde a la unidad (en los casos de los relativos) o al 100% (en los casos de los tantos por ciento).

Bibliografía: Técnica Estadística: Uribe Villegas.

Ojivas

Cuadro Nº 27

Distribución de frecuencias de las calificaciones de la clase que se graduó en la Academia Naval de EE. UU. en 1937.

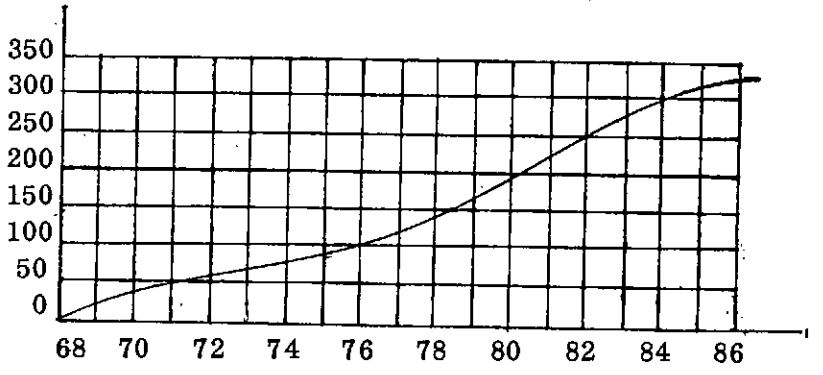
Calificaciones	Guardamarinas
68 - 69,9	4
70 - 71,9	17
72 - 73,9	39
74 - 75,9	62
.....	

En el cuadro 27 muestra la forma usual (no acumulativa) de la distribución de frecuencias y nos presenta el número de guardiamarinas en cada clase y en el cuadro 28 vemos cuántos alumnos o qué proporción de ellos tienen menos de determinadas clasificaciones estipuladas.

Cuadro Nº 28

Calificaciones	Nº Guardiamarinas	% Total
Menos de 70	4	1, 2
" " 72	21	6, 4
" " 74	60	18, 3
" " 76	122	37, 3

Distribución acumulativa de las mismas calificaciones. Mostrando el número de guardiamarinos recibieron "menos de":



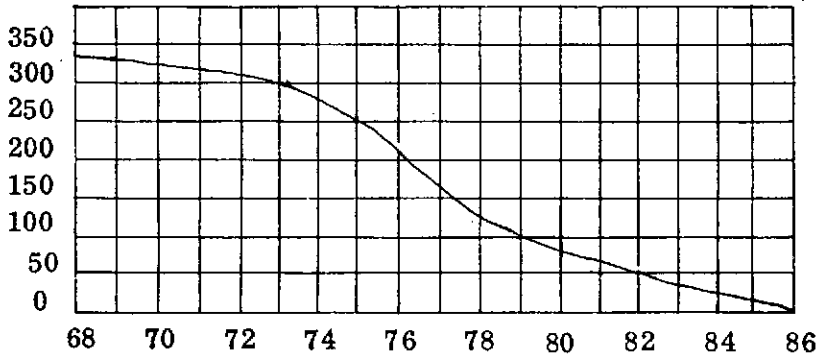
Ejemplo: 21 guardiamarinas o sea el 6.4% tuvieron 72 puntos o menos. Si se exponen las cifras de una distribución acumulativa de frecuencias por medio de una curva, esta será una "ojiva".

Pueden transportarse las cifras absolutas o las relativas. Cuando se trazó una curva de la serie no acumulativa, se transportaron las frecuencias de cada clase en relación con el valor medio de la clase. Sin embargo, las frecuencias acumulativas de cada clase se refieren a un solo valor, y si 4 guardiamarinos sacaron "menos de" 70 (el primer punto de la curva se traza en 4 sobre el eje de las Y, y 70 sobre X, de igual manera las demás.

En vez de buscar las calificaciones inferiores podemos buscar las dadas "o más".

Calificaciones	Nº de Guardiamarinos	Porciento
68 o más	327	100
70 o más	323	98, 8
72 o más	306	93, 6
.....		

Ejemplo:



Ejemplo: 323 o sea 98, 8 % tienen 70.... De nuevo se transportan las frecuencias elevando sobre el eje de las X los valores indicados (68 sobre X y 327 sobre Y).....

Los cuadros de frecuencias acumulativas y las ojivas se usan a menudo para:

1) - Datos de jornales y horas de trabajo. Respecto a los jornales nos permite averiguar cuántos (o qué proporción), recibe lo indispensable para vivir ("menos qué"), lo bastante o lo normal para vivir cómodamente. Averiguar si recibe lo indispensable "o más".....

Averiguar el jornal mínimo (o máximo), si pagó a 10, 25, 50 u otro % de obreros, respecto a horas de trabajo, el número (o proporción) de los que trabajan demasiado o pocas horas.

Comparación de Ojivas:

Si dos o más distribuciones de frecuencia acumulativas se basan aproximadamente en un mismo número de partidas, sus ojivas pueden trazarse y compararse en términos absolutos.

Sin embargo, si las dos series se basan en totales diferentes, es esencial que las comparaciones se basen en las frecuencias de los %.

Relación entre la Ojiva y Curva de Frecuencias:

La curva de frecuencia y la ojiva son simplemente dos disposiciones distintas de datos idénticos, presentando cada una de ellas sus ventajas características que aparecerán si se estudia la relación que entre ambas existe.

La curva que une los rectángulos es suave, mientras las frecuencias adquieren valores considerables y finalmente culmina cuando las frecuencias decrecen en la proximidad al límite de distribución.

Esta es la curva aumentativa de frecuencia de ojiva, es decir, que el área de cada rectángulo es proporcional al número de casos comprendido en el grupo correspondiente, ya las curvas de frecuencias corrientes no pueden ser comparadas entre sí, a menos que las agrupaciones sean idénticas, pero las curvas acumulativas no se hallan atadas a tal restricción.

Además de la desigualdad de los intervalos no produce en la ojiva la distorsión que produce en las curvas de frecuencia.

Curvas de Frecuencias:

La curva de frecuencia se construye tomando sobre el eje de abscisas las intensidades del carácter, y paralelamente al eje de las ordenadas, las respectivas frecuencias.

Entre las curvas de frecuencia encontramos a menudo, aunque no sea en su forma rigurosa, por lo menos en una aproximación suficiente para los fines de orden práctico la curva binomial determinada, por los términos del desarrollo del binomio $(p + q)^n$, que constituyen sus ordenadas.

Si $p = q$, la curva es simétrica, y si hacemos tender n hacia el infinito, tendremos al límite la curva de Gauss o de los errores accidentales que es asintótica por ambos lados al eje de las abscisas, o sea, que se acerca indefinidamente a este eje para aquellas abscisas que crezcan indefinidamente en valor absoluto.

En la práctica, realmente, no se encuentra nunca el caso de errores superiores a un límite determinado, pero esta divergencia entre la curva real y la curva de Gauss es despreciable ya que en esta última la probabilidad de cometer un error muy grande es muy pequeña.

Muchas curvas simétricas se alejan de la curva de Gauss, porque son más achatadas

y menos extendidas, o porque son más agudas y más extendidas; las primeras que se llaman hipobinomiales o hiponormales, y las segundas, hiperbinomiales o hipernormales.

Las curvas observadas hasta aquí que tienen un solo máximo, se llaman unimodales. Otras curvas presentan varios máximos y se llaman plurimodales.

Bibliografía: Estadística de Gini.

Gráficos de Frecuencias Acumuladas (Ojiva).

La acumulación de datos tiende a suavizar la curva. Se usa cuando se quiere representar datos "menos que"; "más que", determinado valor de la variable.

Ojivas:

Las ojivas son representaciones esiformes, sigmoides o en forma de s, correspondientes a distribuciones acumulativas. Para este tipo de distribuciones:

I) - Fórmese una distribución acumulativa:

A) - Sumando las frecuencias en forma progresiva, partir de la correspondiente a los valores inferiores de la variable (o a las clases inferiores) y continuando hacia la correspondiente a los valores superiores de la variable (o a las clases superiores), si se trata de la ojiva llamada "menos de".

B) - Sumando las frecuencias de esa misma forma progresiva pero en sentido inverso: de los valores superiores (o clases superiores) hacia los valores inferiores (o clases inferiores) si se trata de la ojiva llamada "más de".

II) - Tómese los valores de la variable más o menos un medio (en series de frecuencias), o los límites de clases (en series de clases o de frecuencias) sobre el eje horizontal:

A) - Tomando los valores de la variable más un medio (en series de frecuencias) o los límites superiores (series de clases y frecuencias) en el caso de la ojiva "menos de".

D) Tomando los valores de la variable menos un medio o los límites inferiores (en los casos correspondientes) en el caso de la ojiva "más de".

III) - Tómense las frecuencias acumulativas correspondientes sobre el eje vertical, trácense perpendiculares y en el cruce obténgase los puntos de intersección.

IV - Unanse los puntos de intersección para obtener la ojiva.

Lectura de la Ojiva:

Si se trata de una ojiva "menos de":

I) - Tómese un punto sobre la ojiva.
II) - Mídase sobre la perpendicular bajada al eje horizontal, la distancia del punto al eje, leyendo "hay tantos casos" (aquí la distancia medida), "para las cuales la magnitud del fenómeno es de menos de....".
III) - Mídase la distancia del pie de la perpendicular al origen sobre el eje horizontal, y complétese la creación:"...tanto" (aquí el valor de la distancia medida).

En esta forma análoga si se trata de una ojiva "más de", se leerá: "hay tantos casos para los cuales la magnitud del fenómeno es más de tanto".

Como en el polígono de frecuencia, en el caso de las ojivas es también posible utilizar relativos o tantos por ciento como frecuencia; en tales casos las lecturas serán como sigue:

"Hay tantos casos para los cuales la magnitud del fenómeno es menos de: (más de), tanto, si se considera el conjunto de la distribución como unitario".

"De cada 100 casos, haya tantos para los cuales la magnitud del fenómeno es menos de (más de) tanto".

Bibliografía: Técnicas Estadísticas de Uribe Villegas.

Punto 6 : Gráficos especiales -cartogramas- gráficos a doble escala- Gráfico a 2 variables en damero (gráfico de correlación) -Diagrama progresivo de Gannt.

Son mapas geográficos, corográficos, o topográficos, en los que la intensidad de un fenómeno cuantitativa, en las diversas zonas, regiones o circunscripciones (geográficas políticas, administrativas) se representan con diferentes colores o diferentes rayados).

En los casos en que el fenómeno sea por naturaleza contínuo (como por ej. la natalidad, la estatura, etc.), será necesario dividir en parte su intervalo de variación y hacer corresponder a cada una de estas una diferente intensidad o cualidad de coloración o rayado, indicando al márgen de la figura la escala de colores y rayados utilizados.

A veces conviene determinar la intensidad media del fenómeno cuantitativo y dividir los intervalos de variación superior e inferior a la media de un número igual de intervalos parciales, utilizando para las intensidades inferiores y superiores a la media, distintas gradaciones de un mismo color que serán cada vez más intensas, según nos alejemos de dicha media.

Bibliografía Estadística Gini.

Gráficos a Doble Escala:

Gráficos de escala múltiple. No son apreciados para representaciones popu lares. En vez de superponer 2 o más líneas con diferentes escalas, es preferible convertir los datos de ambas a índice y trabajar con una sola escala, o bien, utilizar escala logarítmica para dar cabida a las cifras extremas de ambas series.

No obstante, si el problema requiere ineludiblemente el uso de 2 escalas, deberá identificarse cada curva con su correspondiente unidad de escala. Además, las unidades de cada escala deberán ser opuestas en la misma magnitud. Ej. si en la escala de la izquierda se representa 59 millones y en la derecha 2,4 millones,

la 1ra. debe variar de 10 a 60
la 2da. " " "), 5 a 3

Gráficos de Escalones:

Se usan:

1) - Cuando se muestran fluctuaciones abruptas en los datos (ej. escalas de personal ocupado, sueldos pagados).

2) - Cuando se presentan en un mismo gráfico períodos irregulares de tiempo. (Ej. datos trimestrales y detalle mensual de un trimestre).

3) - Cuando se representa series de frecuencias -de variable contínua- agrupadas en intervalos. Si se acumulan los valores extremos de un solo intervalo, la superficie correspondiente al mismo debe sombrear se para llamar la atención sobre esa acumulación.

4) - Cuando se representan poblaciones (conocidas como pirámide de población).

Diagrama de Gannt:

La finalidad de este gráfico consiste comparar la cantidad realizada de una cosa, con la que se había proyectado. Consiste en una serie de columnas de igual anchura que en su parte superior llevan inscriptas la cantidad proyectada para el espacio de tiempo que representa y la acumulación a través de las unidades temporales de esas cantidades. Mediante 1 línea horizontal que ocupará todo o parte del ancho de cada columna, se indica si se cumplió o nó en su totalidad lo proyecta do, o en caso de que las cifras de lo realizado supere lo proyectado, se traza otra debajo de la primera, que abarcará todo el ancho de la columna hasta representar la cifra realizada.

Finalmente, una línea horizontal más gruesa que la anterior indicará la totalidad de lo realizado en todo el tiempo que abarca el gráfico; que si es menos que lo proyectado en total, no le cubrirá en toda su extensión y si es mayor ocupará

además, un sector del gráfico igual que se hizo en las unidades de tiempo.

Gráfico de Gannt;

Este gráfico constituye una útil herramienta de análisis para estudios de tiempo, movimientos y registros individuales de trabajo.

Se puede usar para registro de trabajos de años, semanas o meses. Usado visiblemente en los lugares de trabajo, incentiva la producción y ayuda a visualizar los problemas de la producción o del personal dentro del departamento o sector productivo.

Construcción:

El cuerpo del gráfico es una hoja rayada en sentido vertical a distancias uniformes. Horizontalmente, en cada columna se ubican los períodos de tiempo (días, semanas, meses).

En la primera columna (columna matriz) se indica nombre o número de empleado, producto, material, máquina, sección, etc. que será objeto de inspección.

Cada uno de los intervalos de tiempo representado por espacio de la misma amplitud, incluye 4 o 5 partes iguales que representa el porciento de cada registro frente al total para cada período. Igualmente se podrán representar cantidades absolutas en lugar de porcientos.

a) Porcientos:

Con una línea firme se llena la cantidad de subdivisiones necesarias para representar las cifras registradas en cada período.

Con una línea general o barra se lleva paralelamente los registros acumulados en cada período. Cuando se presentan causas especiales de paralización del trabajo, se anotan las iniciales correspondientes a cada causa de acuerdo a un código preestablecido.

b) Cifras Absolutas:

Si se trabaja con cifras absolutas, en lugar de porcientos se indica en el ángulo superior de cada espacio la cantidad proyectada, y en el de la derecha, la proyectada acumulada hasta el fin del mismo.

La línea fina representa la producción efectiva de cada mes. La línea gruesa indica la producción real acumulada durante dicho período. Cuando la producción proyectada iguala a la efectiva, la línea fina llena todo el espacio que corresponde a ese mes. Si la excede, se indica el exceso con otra línea fina por encima de la anterior.

En este caso, si bien los espacios en que está dividido cada mes (período) representan intervalos iguales de tiempo; las cantidades varían de acuerdo con la producción efectiva. Ej.: si bien el primer mes se ha proyectado una producción de 10.000 unidades, cada quinta parte del espacio correspondiente al ler. mes, representará 2.000 unidades. En cambio, si para el segundo mes se calculó una producción de 8.000, cada quinta parte representa 1600 U. Por lo tanto, las magnitudes absolutas del diagrama de-

ben leerse referidas a las cantidades que fijaron para cada mes.

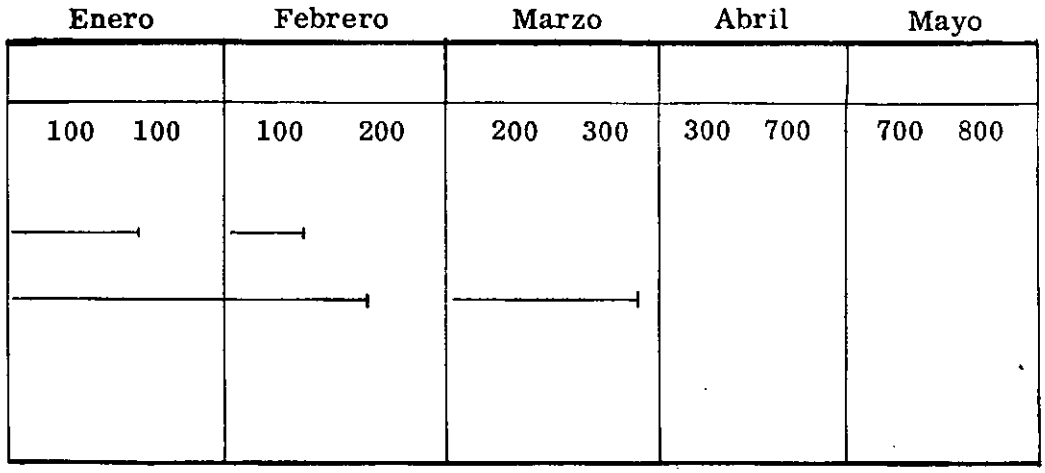
Ej.: Marcha del presupuesto para alimentación del hospital:

Presupuesto Anual	Cálculo de consumo mensual	Consumo efectivo	Consumo efectivo acumulado
Enero $			

La ventaja principal de este gráfico, consiste en que permite a medida en que va transcurriendo el tiempo, ir alargando las líneas de acuerdo con lo que se va realizando y constantemente facilita el conocimiento de la marcha del trabajo en proyecto. Ejemplo.: Representa gráficamente las cifras siguientes:

	Proyectado	Realizado
Enero	100	150
Febrero	100	50
Marzo	200	175
Abril	300	300
Mayo	100	25
Total	800	700

Gráfico



U. N. C.
Escuela Superior de Estudios
Políticos y Sociales

CATEDRA : Introducción a la Estadística (Bases Matemáticas)

1962

TRABAJO PRACTICO Nº 4

Representación Gráfica:

1) - Representar en papel aritmético y en semilogarítmico, los datos siguientes:

a) - Indices de costo de nivel de vida en Capital Federal y de Poder Adquisitivo de la moneda.

AÑO	Indices 1950 = 100	
Año	Costo de nivel de vida	Poder adquisitivo de la moneda
1950	100, 0	100, 0
1951	136, 7	73, 2
1952	189, 4	53, 0
1953	197, 0	50, 7
1954	204, 2	49, 0
1955	230, 0	43, 5
1956	260, 0	38, 4
1957	325, 0	30, 8
1958	350, 3	28, 5
1959	910, 3	11, 0
1960	1.163, 5	8, 6
11 primeros meses de 1961	1.308, 3	7, 6

b) - Indicar cuál es el gráfico más correcto, calculando el número de veces que aumenta o disminuye cada valor anualmente.

2) - Representar a cada escala logarítmica y natural los siguientes datos:

POBLACION SEGUN CENSOS NACIONALES

Territorio	1869	1895	1914	1947	1960
	en miles de habitantes				
Total del país	1.737,1	3.954,9	7.885,2	15.879,1	20.008,9
Buenos Aires	307,9	921,8	2.066,9	4.273,9	6.734,5
Mendoza	65,4	116,1	277,5	588,2	825,5
Sgo. del Estero	132,9	161,5	261,7	479,5	477,2

Indicar porqué conviene usar la escala logarítmica o natural que se haya elegido.

3) Representar en papel semilogarítmico y natural lo siguiente:

Evolución del monto de 100.-- pesos colocados a interés compuesto del 12% anual durante 6 años.

4) Representar en papel semilogarítmico y milimetrado, los siguientes datos de mortalidad, colocando en la escala de ordenada las cifras de fallecidos por cada 100.000 habitantes (tasas).

Tendencias de las tasas de muerte por tifoidea o sífilis, en personas blancas en EE. UU.

TASAS DE MORTALIDAD POR:

Año	Tifoidea	Sífilis
	‰ habitantes	
1920	4,9	16,1
1925	3,7	15,3
1930	1,5	12,1
1935	0,7	11,0
1940	0,3	10,3

Cuál el porcentaje de disminución en cada causa de muerte.

4. FUENTES ESTADISTICAS

El análisis estadístico, ya sea en problemas sociales, políticos, económicos, etc., no se puede realizar hasta haber:

1o. recopilado las estadísticas necesarias;
2o. comprobado su exactitud.

Fuentes Primarias:

El primer problema que el investigador debe considerar, es la recogida y reunión de datos.

En varios campos y especialmente en el social y económico, el investigador no puede, en todos los casos, preparar él mismo los datos, sino que ha de conformarse con los que pueda obtener en fuentes de información ya existentes y que por lo general trabajan con fines distintos de los que él persigue (censos, estadísticas, aduana, etc.). En otros casos, como en psicología, biología, medicina, meteorología, el investigador puede producir él mismo los datos necesarios o bien usar los de otras personas ocupadas en cuestiones análogas.

Así entonces, a veces actúa como productor de cifras, y otras, simplemente como usuario de las ya existentes. En el primer caso, diremos que las cifras provienen de una fuente primaria y en el otro, de fuente secundaria.

Fuentes primarias :	es la institución que originalmente planificó la investigación y recopiló los datos.
Fuentes secundarias :	es la institución o publicación que divulga o utiliza datos que originalmente fueron ya recopilados por otra organización.

Comprobación de Datos:

Muchas publicaciones son al mismo tiempo, fuentes primarias para unos datos, y secundarias para otros.

La confianza que merezcan los datos, debe ser examinada antes de basar en ellos ninguna conclusión. Será un despilfarro de tiempo y podrán sacarse conclusiones peligrosas o falsas, si se aplican métodos estadísticos refinados a datos que son desde un principio, sospechosos. Los datos estadísticos no deberán usarse "a ciegas". Sólo un cuidadoso estudio de todas sus limitaciones, llevará a un uso inteligente de cifras publicadas.

En general, es preferible utilizar los datos de las fuentes primarias que de las secundarias, porque:

1o.) Los datos en las fuentes primarias tienden a ser más completos que en los secundarios. En las primarias se puede encontrar distintas clasificaciones y aperturas en detalles que las fuentes secundarias suelen omitir.

2o.) En la fuente primarias se dispone de información complementaria sobre los objetivos, definiciones y limitaciones utilizadas en la investigación original; detalles que no siempre aparecen en la secundaria. Esta información complementaria a veces decide la utilización o el desprecio de la información. Ayuda mucho a valorizar e interpretar los datos.

3o.) Siempre existe la posibilidad de que en la fuente secundaria se deslicen errores inecistentes en la fuente primaria. Cuando se usan datos publicados se debe poner especial cuidado en conocer sus limitaciones. Estas pueden incluir:

1. Errores causados por el uso de técnicas imperfectas o inapropiadas en la recopilación.
2. Errores por la redacción incorrecta de preguntas.
3. Errores por prejuicios de los entrevistadores.
4. Equivocaciones personales de los informantes.
5. Equivocaciones del personal que recopila la información.
6. Errores tipográficos de transcripción en el procesamiento o en la publicación de los datos.
7. Cambios de conceptos, definiciones, métodos de recopilación de datos, etc. sobre un período de tiempo. Si esto ha ocurrido, es importante comprobar el efecto de estos cambios sobre los datos.

Sin embargo, las fuentes secundarias se usan muy a menudo y ello se debe a dos ventajas:

1o.) Acumulan datos dispersos en numerosas fuentes primarias (gubernamentales o privadas). Así ocurre con los Anuarios y Boletines Estadísticos de la Dirección Nacional de Estadística y Censos.

2o.) Las fuentes secundarias amplias, pueden utilizarse para localizar rápidamente a las fuentes primarias.

Fuentes de Datos en Argentina:

Consultar la publicación :

Instituto Torcuarto Di Tella, Centro de Investigaciones Económicas, Catálogo de Estadísticas Publicadas en la República Argentina.

Capítulo II

MEDIDAS ESTADISTICAS

Ya se ha dicho en este curso que la ordenación de los datos en determinada forma, nos sirven fundamentalmente para condensar la información que nos sería útil extractar para que con un solo número o expresión pudiésemos sacar una conclusión que nos defina el evento que estudiamos.

Se dijo que entre esas expresiones podríamos contar con los porcientos y razones por un lado, y también con los promedios, primero veremos a aquellos y luego estos últimos por razones de metodología.

1. RAZONES Y PORCIENTOS

Para expresar la razón, relación o proporción que existe entre dos números dados, tenemos que dividir dichos números entre sí (el mayor en el menor) como sería por ejemplo este caso: Qué relación, proporción o razón existe entre 1.200 y 400? Luego, dividimos 1.200 en 400 y obtenemos como resultado 3. Entonces concluimos diciendo: que 1.200 es a 400 como 3 es a 1; esto expresado con relación a la unidad o sea tanto por uno, pero si al resultado de la división lo multiplicáramos por cien o por mil, la relación quedaría expresada en tanto por ciento o tanto por mil respectivamente.

Todas estas razones y proporciones, las calculamos con finalidad de poder comparar eventos presentados en serie, ya sea desde el punto de vista dinámico o estático. Desde el punto de vista dinámico, sería por ejemplo comparar las cifras arrojadas en el cálculo del producto bruto al mes de diciembre de 1964 de Mendoza o San Juan a lo que arrojaron al mismo mes de diciembre de 1963. En cambio, desde el punto de vista estático sería comparar las cifras de ese mismo producto bruto para diciembre de 1964 entre Mendoza y San Juan.

Las variaciones relativas de dos o más eventos, pueden compararse o apreciarse más concretamente cuando efectuamos esas relaciones como lo vamos a mostrar ahora:

Cuando comparamos nosotros, uno o más números con respecto de otro, a este último lo llamamos "BASE". Encontramos la proporción dividiendo entre la base los números que queremos comparar con ella. Ejemplo:

Evolución porcentual de la producción de uva para San Juan y Mendoza tomando como base 1951 = 100. -

(Por razones didácticas hasta 1955)
(En Quintales)

Años	Producción San Juan	Evolución Porcentual	Producción Mendoza	Evolución Porcentual
1951	3.488.254	100,00	10.939.921	100,00
1952	3.906.330	111,99	10.159.253	92,86
1953	4.536.319	130,05	12.540.471	114,63
1954	4.468.432	128,10	9.373.634	85,68
1955	5.700.860	163,43	16.993.470	155,33
....	.....	.....	.....	.....
1963(*)	6.288.104	180,27	18.071.118	165,19

(*) Informativo.

También en este mismo cuadro podríamos nosotros determinar los porcientos de aumento de la producción con respecto del año inmediatamente anterior, lo que significaría que en cada cálculo nuestra base sería distinta para todos los años.

Cuando se trata de una serie de suma importancia y de componentes muy grandes, ambos casos se han presentado en nuestro ejemplo, conviene sacar los porcientos con decimales para dar una idea bien concreta de la comparación que queremos llevar a cabo. En otros casos en que no se presentan las dos premisas anteriormente expuestas, no es necesario calcular dichas razones con decimales, ya que vienen a ensombrecer un poquito nuestra visualización.

Algunas proporciones que se usan con frecuencia son las siguientes:

Números Indices:

La mayor parte de los números índices se representan con porcientos. Tomemos un ejemplo: Indice de evolución de los precios mayoristas en EE. UU. que toman como precio base el de 1926 y el resto lo expresan con relación a éste. Debe tenerse especial cuidado en la elección del año o número base, ya que éste debe ser el año o período en que

hayan ocurrido las cosas o problemas a que se refiere el evento en forma absolutamente normal.

Proporción de los Sexos:

La proporción de hombres con respecto de las mujeres se expresa normalmente en porcientos o por miles, así tenemos por ejemplo: Que en 1960 en Valle Fértil, Departamento de la Provincia de San Juan, había una población total de 3.865 personas de las cuales eran 1965 varones y 1900 mujeres, lo que equivale a establecer una proporción de varones sobre mujeres de 103,42.

Densidad de Población:

Para establecer la densidad de población de una Provincia o Nación determinada, ya conocemos el procedimiento y sabemos que no es nada más que un promedio, ya que no podíamos nunca decir que en 1963 la densidad de Valle Fértil es de 0,27 de persona y que, efectivamente, se encuentra distribuida así, una vez más nos encontramos frente a la utilización de un promedio.

Proporciones Per Cápita:

Lo mismo que en el caso anterior cuando decimos que a cada persona en San Juan, le corresponde una determinada parte del Producto Bruto Interno, no queremos decir que en la realidad deba o corresponda que cada uno tome una idéntica porción de ese Producto, sino que simplemente estamos frente a un simple promedio que nos da una idea más o menos precisa de cómo evoluciona cuando se lo compara cronológicamente el ritmo o nivel de vida de la Provincia o Nación que estudiamos.

Coeficientes de Natalidad y Mortalidad:

Se calcula dividiendo el número de defunciones y/o el número de nacimientos, el de población de esa zona a la mitad del año que se considera. También estamos en presencia de un promedio, en éste y otros casos que más adelante se nos han de ir presentando.

2. VALORES CENTRALES

Ya se ha visto cómo se construye una distribución de frecuencia o lo que es más frecuente escuchar, una serie de datos agrupados en determinada forma. De esa cantidad de datos agrupados, generalmente se presentan como los más frecuentes, los datos del centro de la serie, que al ser representados gráficamente asumen la forma de una campana; entonces nosotros al decir que los valores más característicos de esas series, se centran o sea que se presentan al centro de la distribución que estudiamos cuando buscamos un valor que sea significativo para que nos dé una idea total en una sola expresión decimos que estamos tratando de hallar las "medidas de tendencia central" de esa serie. Este tipo de medidas es lo que nosotros entraremos a analizar ahora, y sobre todo ya que se trata de un curso de adiestramiento y capacitación, intensificaremos los aspectos prácticos de este tipo de medidas. Más adelante veremos las medidas de "dispersión" que se refieren a la diseminación de una distribución o serie, y

luego veremos las medidas de "asimetría" que miden la dirección y cantidad de asimetría y por último estudiaremos la "kurtosis" que nos indicará en qué grado una serie tiene "picos".

Promedios: "Para los estadísticos es un término general aplicado a toda clase de medida de tendencia central derivada de un grupo de datos". Es equivalente al término "Intensidad media".

Condiciones de los promedios

a) Deben ser definidos y así expresados en forma precisa (expresan) y con ello se evitan ambigüedades.

b) Generales, entendiéndose por tal de que la población que se presenta contenga, sin excepción, todos los datos.

c) Elementales, y serán así fáciles de entender y sencillos de obtener.

2.1. - La Media Aritmética:

La media aritmética de datos no agrupados: es tan común entre todos nosotros usar el término "media aritmética", que siempre que nos referimos a ella posiblemente ni la mencionamos como tal, sino que la enunciamos diciendo simplemente "la media" o el "promedio"; en cambio cuando nos referimos a la "media geométrica" o la "media armónica" o a cualesquier otro tipo de media que normalmente pueden llegar a utilizarse, las denominamos como corresponde o sea por su asignación o denominación técnica.

La "media aritmética" de una serie, la obtenemos por un procedimiento sumamente sencillo o sea: Sumando las cantidades del evento presentado y lo dividimos por el número de veces que esa serie de eventos (variable) se ha presentado (frecuencia).

Ejemplo: Determinar cuál es la "media aritmética" del precio del pan que se vendió esta mañana en la panadería sabiendo que se vendieron: Uno de $ 27; Uno de $ 29; Uno de $ 34; Uno de $ 33; Uno de $ 36 y por último uno de $ 31 ?"

$$\frac{27 + 29 + 34 + 33 + 36 + 31}{6} = \$\ 31,66 = \bar{X}$$

Representándolo como en Estadística se acostumbra, y utilizando la simbología correspondiente, tendremos:

$$\frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{n} = \bar{X}$$

O también para reducir la expresión a su máxima condensación, podremos decir:

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

En el caso que damos anteriormente no contemplamos la posibilidad de que podamos haber vendido cantidades diferentes del pan de un mismo precio; por lo tanto ahora lo presentamos:

1.000 Kgs. a $ 27: 2.000 Kgs. a $ 29; 3.000 Kgs. a $ 34; 1.000 Kgs. a $ 33: 4.000 Kgs. a $ 36 y por último 2.000 Kgs. a $ 31.

Es evidente que la primera de las medias calculadas no podemos de dejarle de llamar "ponderada" ya que, se ponderó de acuerdo al número de veces que se presentó; en el segundo ejemplo le podemos decir que, además de considerársela ponderada por la unidad se la pondera por el número de veces que el evento se presenta.

Entonces, en consecuencia, toda "media" que nosotros calculemos la vamos a considerar ponderada (ya sea ponderada por la unidad) o (doblemente ponderada) y por lo tanto la llamaremos en todos los casos simplemente "media". La segunda serie presentada, la calculamos teniendo en cuenta las frecuencias o número de veces que se ha repetido cada característica que asume la variable, O sea:

$$\text{(a)} \quad \frac{X_1 f_1 + X_2 f_2 + X_3 f_3 + X_4 f_4 + X_5 f_5 + X_6 f_6}{N} = \bar{X}$$

Expresándolo numéricamente:

$$\frac{(27 \times 1.000)+(2.000 \times 29)+(3.000 \times 34)+(1.000 \times 33)+(4.000 \times 36)+31 \times 2.000}{13.000} = 32,77 = \bar{X}$$

La expresión (a) para condensarla podemos expresarla más cómodamente de la siguiente manera:

$$\frac{\sum_{x=1}^{n} X_i f_i}{N} = \bar{X}$$

Cuando es n (minúscula) estamos en presencia que las frecuencias en cada caso no son nada más que de la unidad, en cambio colocamos N (mayúscula) cuando cada posición que asume la variable se repite más de una vez.

Debe tenerse especialmente en cuenta que ninguna pieza de pan se ha vendido al precio que nos acaba de proporcionar el cálculo de la media, pero es en realidad el promedio, entonces concluimos diciendo que la media aritmética de una serie no es valor real, sino simplemente un valor calculado.

El signo que estamos usando en la expresión condensada de la media, o sea: $\sum_{x=1}^{n}$ es lo que se denomina "Sumatoria" de los valores de la variable desde que X = 1 hasta n; n en el primer ejemplo nuestro vale 6, y en el segundo es igual a 13.000, que es el número de veces que aparece repetida o no la variable en la distribución.

Inconveniente: El principal inconveniente para la media aritmética es la presencia de valores extremos de la serie, es decir, valores que salen de lo normal y que influyen demasiado en el promedio. Por ello, solo puede ser aplicado (la media aritmética) en serie homogénea, respecto al tiempo, espacio o al asunto que se trate, los datos deben estar dados en la misma unidad y los elementos de la serie deben tener la misma importancia. No se obtendrá ningún valor suficientemente representativo si en la serie entremezclan datos de importancia muy diversa.

Por ejemplo: Si hablamos de la producción de uva, puede darse el caso de que se mantenga más o menos uniforme durante un cierto número de años, y dar valores muy altos o muy bajos en algunos años considerados como extraordinarios o extremos. Esos extremos alterarían por completo los resultados de la media. Otro ejemplo: No podrá calcularse un sueldo promedio para individuos que perciben por ejemplo: $ 14.000; $ 15.000 y $ 70.000.

2.2. Media Geométrica

Se obtiene extrayendo la raíz de índice igual al número de factores que intervienen en la operación. Para dos valores se calcula: mg : $\sqrt{X^1 . X^2}$ o sea raíz de índice 2 del producto de esos valores.

Para N valores,

En serie simple se calcula:

$$M_g = \sqrt[n]{X_1 \quad X_2 \ldots . X_n}$$

Aplicando logaritmo a lo anterior, nos queda convertida en una media aritmética.

$$\text{Log } M_g = \frac{\log X_1 + \log X_2 + .. + \text{Log} X_n}{N}$$

O sea, logaritmo de la media geométrica (Mg) es igual a la media aritmética (M) de logaritmos de los valores dados, Ejemplo:

Media geométrica de los números 2, 4, 8, es:

$$M_g = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = 4$$

Su cálculo arroja siempre cifras menores a la media aritmética y sólo en casos excepcionales pueden darse valores iguales. Está menos influenciada por valores extremos, lo que le da cierta superioridad teórica que se hace notar en algunas series. Se utiliza

generalmente cuando los valores a promediar son muy desiguales por pertenecer a distribuciones asimétricas, porque da poca importancia a los valores extremos.

b) Es el promedio más adecuado en series de números que presentan una relación constante entre 2 contiguos como: 2, 4, 8, 16, etc. Se emplea para calcular promedios de series cuyas variaciones se expresan por grados más que por diferencias absolutas.

Tiene también la ventaja de ser una expresión algebraica, pero presenta la desventaja de ser más laborioso el cálculo que el de la Media Aritmética.

Podemos mencionar entre sus desventajas, la de no ser aplicable cuando hay valores negativos o cero. No es muy usada por esas razones, pero debería serlo aunque más no sea para controlar la Media aritmética. Si los valores de M y Mg guesen próximos probará que están bien hallados, si difieren mucho es preferible optar por la Mg.

Procedimiento para el cálculo de la Mg. : Dos procedimientos:

a) Para el caso de series simples: En este caso se utiliza el procedimiento general dado en la Definición de Media Geométrica.

$$Mg = \sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdots x_N}$$

b) Para el caso de Series de Frecuencias: Entra en estos casos, otro elemento: El de frecuencia fi, para ponderar a cada valor de la variable, según el número de veces que se haya presentado.

$$Mg = \sqrt[\Sigma f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdots x_N^{f_n}}$$ aplicando logaritmos

$$\log Mg = \frac{\log [x_1^{f_1} \cdot x_2^{f_2} \cdots x_N^{f_N}]}{\sum f_i} = \frac{1}{\sum f_i} \left[f_1 \log x_1 + f_2 \log x_2 + \cdots + n \log x_n \right]$$

$$\log Mg = \frac{1}{\sum f_i} \cdot \sum f_i \log x_i \quad \therefore$$ por lo tanto

$$Mg = \text{antilog} \left[\frac{\sum f_i \log x_i}{\sum f_i} \right]$$

CONCEPTO DE DESVIO O DESVIACIONES

Desvío o Desviación estadística es la diferencia (positiva o negativa) de un valor determinado con respecto a un valor constante o central tomado como término de comparación.

La desviación es un carácter de cada uno de los datos de una distribución. La dispersión es una característica propia de la distribución. La dispersión de una distribución indica la mayor o menor concentración de los datos en torno de un promedio central (principalmente en torno de la media) y por consiguiente, indica el grado de representatividad del promedio.

La dispersión de una distribución se mide en base al cálculo de promedio de desviaciones al cuadrado.

2.3.- Concepto de Media Provisoria o "de Trabajo". Cálculo de Media Aritmética de Series de Frecuencias Usando el "Método Corto".

1. Se forman los intervalos de clases o grupos de valores con amplitud "w".
2. Se forma una columna con los puntos medios de dichos intervalos (X_i)
3. A continuación se agrega una columna con las frecuencias con que se presentó cada grupo o intervalo (fi).
4. Las sumas o totales de las frecuencias las indicamos con N.

Si deseamos calcular la media aritmética, aplicamos la fórmula:

$$\overline{X} = \frac{\sum x_i f_i}{\sum f_i}$$ para lo cual debemos calcular

los productos de la segunda y tercera columna, es decir Xi . fi

Pero este procedimiento se usa para series cortas de pocos valores o de números pequeños, es el llamado Método Largo. Si se trata de estudiar series largas o de números grandes se aplica el Método Corto o Reducido que consiste en lo siguiente:

Se cambia el origen de la variable a un valor central que consideramos desde entonces, como origen y al que llamaremos Media Provisoria:

1 Intervalos W=5	2 Puntos Medios Xi	3 Frecuencias fi	4 Desvíos xi	5 Desvíos ponderados xi . fi
40-44, 9	42, 5	2	-3	-6
45-49, 9	47, 5	5	-2	-10
50-54, 9	52, 5	7	-1	-7
55-59, 9	57, 5	16	0	-23
60-64, 9	62, 5	12	1	12
65-69, 9	67, 5	8	2	16
70-74, 9	72, 5	4	3	12
		$\sum f_i = 54$		$\sum x_i \cdot f_i = -6$

Si tomamos el valor 57, 5 como media provisoria y calculamos en cuánto se desvía de él cada uno de los demás, mediante la fórmula: $xi = \frac{Xi - Mp}{W}$ obtendremos la nueva columna (4) que reemplaza a la (2) con la gran ventaja de manejar números pequeños. -Qué significan los números de esta columna? -Son los intervalos o distancias que hay entre cada valor y la Media Provisoria. Desde ahora en el cálculo, el origen es la Media Provisoria a quien le asignamos el valor cero; ya que siendo:

$$x_i = \frac{X_i - M_o}{w}$$

resulta $\frac{0 = 57,5 - 57,5}{5} = \frac{0}{5} = 0.$

La unidad en la que se miden los valores y en la que se desarrollará el cálculo es el intervalo de clase: W.

Entre 52, 5 y 57, 5, -qué distancia hay? : Un intervalo. Aplicando la fórmula:

$$x\,i = \frac{X\,i - M\,p}{W}$$

tenemos: $\frac{52,5 - 57,5}{5} = -1$

análogamente: $\frac{47,5 - 57,5}{5} = -2$

y en el otro sentido: $\frac{62,5 - 57,5}{5} = 1$

$$\frac{62,5 - 57,5}{5} = 2$$

De esta manera surge la columna No. (4)

Media Aritmética Verdadera

Es necesario repetir que el origen se cambia al centro de la serie, y ya no se considera desde el origen de la variable primitiva. Esto trae aparejado un cambio en la fórmula:

$M = \frac{\sum x_i f_i}{\sum f_i}$ se transforma ahora en $M = \frac{\sum (x_i - Mp + Mp) f_i}{\sum f_i}$

o sea:

$$M = \frac{\sum (x_i - Mp) f_i}{\sum f_i} + \frac{Mp \cdot \sum f_i}{\sum f_i}$$ (Mp sale afuera de la por ser una constante

por lo tanto:

$M = \frac{\sum x_i f_i}{\sum f_i} w + Mp$ siguiendo el lenguaje de Charlier

$\boxed{M = b + Mp}$ ya que $b = \frac{\sum x_i f_i}{\sum f_i} w$

Resumiendo: para encontrar la Media Aritmética verdadera por el método corto, se debe : 1 - Formar la columna xi . fi
2 - Encontrar su suma algebraica
3 - Multiplicar por w (valor de intervalo de clase
4 - Dividir (3) por la suma de las frecuencias.
5 - Sumar (4) al valor de la media provisoria.

Interpretación de b : El término b es la corrección necesaria de la Media provisoria Mp, para obtener la Media Aritmética verdadera. Vale decir que es la diferencia entre la Media Aritmética verdadera y la Media Aritmética provisoria.
b = M - Mp.

Es además un buen auxiliar que permite comprobar los cálculos.

2.4. - Mediana y Cuartilos:

Definición: "La mediana de una sucesión de valores ordenados en forma creciente, es aquel valor de la sucesión de que se trate, que divide a la misma en dos partes que tienen el mismo número de elementos cada una".

Para entrar a tratar en forma creciente el problema de la mediana, vamos a suponer una serie sencilla de valores que todos tienen una frecuencia de uno, o sea que se encuentran repetidos únicamente una sola vez, por ejemplo veamos:

Determinar la mediana de la siguiente distribución:

2, 6, 10 y 14 (para todos la frecuencia es = 1)

Es evidentemente sencillo darnos cuenta dónde está ubicada la mediana y decimos así: se ubica entre 10 y 6 por lo tanto sumamos ambos y los dividimos en dos y encontraremos la mediana con una simple media aritmética donde la variable x = a los valores centrales de la serie y la frecuencia a los efectos de ese cálculo es = 2; luego:

$$= \frac{6 \quad 4 \quad 10}{} = 8$$: : decimos que la mediana de ésta se encuentra ubicada entre los valores 2º y 3º de la serie, y es = 8.

Esto que hemos expuesto, es cuando se trata de una serie que posee valores pares, pero, cómo resolvemos el problema cuando se nos presentan distintas frecuencias y la suma de éstas es impar? :

Sea la siguiente distribución de la variable x discontínua :

x	f	fi acumulada
5	11	11
7	9	20
14	15	35
17	8	43
20	12	55
	N = 55	

Para determinar la mediana de la variable X, ordenamos los valores en forma creciente y obtendremos lo siguiente:

5, 5, 5, 5 (...11 veces.....)...5, 7, 7, 7, .(..9 veces.) 7, 14, 14, 14, 14, 14, (..15 veces) 17, 17, 17, 17, (.....8 veces.....) 17, 20, 20, 20, 20, 20 (12 veces.....20.

Una vez ubicado en esta forma, debemos buscar el número que es la mediana de la sucesión y si recordamos la definición de mediana, diciendo qué es aquel valor de la serie que divide a ésta en dos partes iguales, tendremos lo siguiente:

$$Ma = \frac{N + 1}{= 2} = \frac{55 + 1}{2} = 28$$

Por lo tanto, el valor que nosotros buscamos está ubicado en el lugar No. 28 de la serie y si acumulamos las frecuencias tendremos ubicado el valor de la variable que corresponde.

En consecuencia, decimos que se ubica dentro de la fase acumulada = 35 y por lo tanto es 14 el valor de la variable cuando es la Ma.

28 = Ma
14 = Ma

- Ma = orden de la mediana
  Ma = mediana o sea valor de la variable que es mediana.

Cuartilos: Existen el 1o, 2o, 3o. El 1o. y 3o. son los más importantes- ya que el 2o. no es nada más que la Ma.

1o. Cuartilo : El 1o. O de una sucesión de valores ordenados en forma creciente es aquel valor de la distribución que deja una cuarta parte (25%) de los N valores inferiores a ese valor determinado como 1o. Q.

3er. Cuartilo: ídem al anterior pero deja 75% o sea los 3/4 parte de los valores inferiores al valor del 3o.

Luego podemos decir que, si la mediana divide en 2, los cuartilos los dividimos en 4 y multiplicamos por el orden que le corresponde al Q.

Ejemplo:

$$^{\circ}Q\,1 = \frac{N + 1}{4} \quad 1$$

$$^{\circ}Q\,3 = \frac{N + 1}{4} \cdot 3$$

$$^{\circ}Q\,2 = \frac{N + 1}{4} \cdot 2 = Ma$$

3. DESVIO O DESVIACIONES

Generalidades

Desvío o Desviación estadística es la diferencia (positiva o negativa) de un valor determinado de la variable con respecto de un valor constante o central tomado como término de comparación.

La desviación es un carácter de cada uno de los datos de la distribución de frecuencias.

La dispersión es una característica propia de la distribución. La dispersión de una distribución indica la mayor o menor concentración de los datos de la distribución a que se refiere en torno a un promedio o valor central (casi el 90% de los casos la media = $\bar{X}$) y por consiguiente indica el grado de representatividad del promedio.

La dispersión de una distribución se mide en base al cálculo del promedio de las desviaciones con respecto de la media elevadas al cuadrado.

3.1.- Momentos :

Se denomina momento a la suma de los productos de la variable elevada a una cierta potencia por la frecuencia o repeticiones correspondientes, siendo el exponente de la potencia, el número que caracteriza el orden del momento.

- **Momentos**
  - **Potenciales**
    - Absolutos $M_s = \sum_{i=1}^{n} x_i \cdot r_i$
    - Relativos $m_s = \sum_{i=1}^{n} x_i \, y_i$ siendo $y_i = \frac{r_i}{N}$
      - NATURALES
      - CENTRADOS
      - REDUCIDOS
  - **Factoriales o Binomiales**
    - Sirven para ciertos estudios de tipo teórico.

De aquí en adelante nos referiremos al decir simplemente momentos a los momentos potenciales relativos, ya sean naturales, centrados o reducidos, esto último según se especifique.

Variable de los momentos naturales: La variable natural (X_i)

Variable de los momentos centrados: La variable centrada ($X_i - \bar{X}$)

Variable de los momentos reducidos: Variable reducida (t_i) $= \left(\frac{X_i - \bar{X}}{V}\right)$

Los momentos los vamos a simbolizar así:

Los naturales = m_s ; los centrados = $\mu_o =$;y por último los reducidos $= q_s$.

a) Momentos Naturales

$$m_o = \frac{1}{N} \sum_{i=1}^{n} x_i^o \, r_i = 1$$

$$m_1 = \frac{1}{N} \sum_{i=1}^{n} x_i^1 \, r_i = \bar{X}$$

$$m_2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 \, r_i :$$ momento de orden 2

$$m_3 = \frac{1}{N} \sum_{i=1}^{n} x_i^3 \, r_i =$$ momento de orden 3

$$m_4 = \frac{1}{N} \sum_{i=1}^{n} x_i^4 \, r_i =$$ momento de orden 4

b) Momentos Reducidos:

$$q_{10} = \frac{1}{N} \sum_{i=1}^{n} t_i^0 \, r_i = 1$$

$$q_{11} = \frac{1}{N} \sum_{i=1}^{n} t_i^1 \, r_i$$

$$q_{12} = \frac{1}{N} \sum_{i=1}^{n} t_i^2 \, r_i$$

$$q_{13} = \frac{1}{N} \sum_{i=1}^{n} t_i^3 \, r_i$$

$$q_{14} = \frac{1}{N} \sum_{i=1}^{n} t_i^4 \, r_i$$

c) Momentos Centrados:

Cuando centramos la variable con respecto de la media, la llamamos variable centrada o sea desvío con respecto de X.

$$d_i = (x_i - \overline{X})$$

$$\mu_0 = \frac{1}{N} \sum_{i=1}^{n} d_i^0 x_i = 1$$

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} d_i^1 x_i = \underbrace{\frac{1}{N}\sum_{i=1}^{m} x_i r_i}_{1)} - \underbrace{\frac{1}{N}\sum_{i=1}^{m} \bar{X} r_i}_{2)}$$

1) $m_1 = \bar{X}$

2) $m_1 \frac{\sum r}{N} = m_1 = \bar{x}$ $\therefore$

$$m_1 - m_1 = 0$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} (x - \bar{X})^2 r_i = \frac{1}{N}\sum_{i=1}^{n} (x_i^2 - 2x_i\bar{X} + \bar{X}^2) r$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 r_i - \frac{2\bar{X}}{N}\sum_{i=1}^{m} x_i r_i + \bar{x}^2 \frac{\sum r_i}{N}$$

$$\boxed{\mu_2 = m_2 - m_1^2}$$

A este momento lo denominamos **Variancia** : $\mu_2 = m_2 - m_1^2$

$$V = m_2 - m_i^2$$

$$V = \sigma^2$$

desvío standard $\sigma = \sqrt{m_2 - m_1^2}$

$$\sqrt{V} = \sqrt{m_2 - m_1^2} = \therefore$$

$$\sqrt{\ } = \sqrt{m_2 - m_1^2}$$

Ahora que tenemos una idea concreta de que es numérica o algebraicamente la variación y la dispersión, las definiremos y decimos que:

3.2 - Variancia:

La variancia de una distribución de frecuencia es la media aritmética de los cuadrados de los desvíos de la variable con respecto de la media.

3.3 - Dispersión:

Es la raíz cuadrada de la variancia o también desviación standard o desvío medio cuadrático.

Características:

1) La variancia y la dispersión son siempre números positivos.

2) Por la manera como están distribuidas miden la homogeneidad de la distribución, teniendo en cuenta que cuando más se centra, los valores de la variable con respecto de su media menor será la variancia y por lo tanto menor la dispersión. Al acercarse los valores de la variable más a la media, significa que ellos están más tipificados y por lo tanto la distribución es más homogênea.

La importancia de dar la media aritmética sin la $\sqrt{\ }$ y $\sqrt{\ }^2$, es tan relativa que sin esos datos diríamos que en cierto modo la estadística dada tiene mucho menos valor si no viene acompañada de las otras.

Para confirmar lo expuesto damos un ejemplo:

Dos personas, A y B realizan diversas pesadas de una misma cosa y se obtiene ambas lo siguiente:

$\overline{X}_A = 10.03$ $\qquad$ $\overline{X}_B = 10.03$

$\sqrt{\ }_A = 0.19$ $\qquad$ $\sqrt{\ }_B = 0.11$

Luego las medias son iguales, y si no tuviéramos las $\sqrt{\ }_A$ y $\sqrt{\ }_B$ diríamos que A y B pesan iguales.

Después observamos las $\sqrt{\ }_A$ y $\sqrt{\ }_B$, y vemos que la $\sqrt{\ }_B$ es menor y concluimos diciendo que B. es mejor porque es mucho más homogêneo que A para pesar.

Es fundamental que las condiciones en que se pesa la cosa por las personas A y B, sean idênticas.

Los promedios tienen por objeto situar el valor central dentro de la serie de frecuencias, en cambio no indican nada sobre la extensión o "reparto" de los valores de la serie. Para estudiar la distribución de los valores se usa una característica llamada dispersión que se indica con la letra griega (sigma). Interesa en una saber si los valores de la serie, en conjunto, están muy próximos a la media (concentración) o muy dispersos (gran variabilidad). Se busca determinar con un valor numérico ese grado de dispersión de los valores respecto al valor central o normal.

$\sqrt{\quad}^2 =$ variancia o desvío cuadrático medio.

$\sqrt{\sqrt{\quad}^2} = \sqrt{\quad} =$ dispersión.

Si, calculando $\sqrt{\quad}$ a partir de los valores de la serie, tomamos $\sqrt{\quad}$ a la derecha y $\sqrt{\quad}$ a la izquierda de la media, se determina un intervalo donde quedan comprendido los dos tercios de todos los valores, o sea 68% de toda la población. Si se toman $\pm 3\sigma$ se incluye en tal intervalo a más del 99 % de los valores

El valor sigma es la medida de la distribución de los elementos alrededor de la media.

Partiendo del concepto de variancia o momento centrado de 2º orden.

$$\mu_2 = \sqrt{\quad}^2 = \frac{\sum (x_i - M)^2 f_i \cdot W^2}{\sum f_i}$$

En el método corto se trabaja con momentos reducidos, esto es, cambiando el origen, a la media provisoria.

$$\mu_2 = \sqrt{\quad}^2 = \frac{\sum [(x_i - M_o) - (M - M_o)]^2 f_i \cdot W^2}{\sum f_i} =$$

$$= \frac{\sum [w \cdot x_i - b]^2 f_i}{\sum f_i} = \frac{\sum x_i^2 \cdot f_i \cdot w^2}{\sum f_i} - \frac{2b \sum x_i \cdot f_i \cdot w + b^2}{\sum f_i} =$$

$$= \mu'^2 - 2b \cdot b + b^2$$

$$\sigma^2 = \mu^2 - b^2$$

la dispersión $\sigma = \mu'_2 - b^2$

Características de la Dispersión

Las medidas de variabilidad, entre las cuales se encuentra la dispersión, son más difíciles de interpretar que los promedios debido a que su uso es menos popular. No obstante, poseen características importantes bien definidas y de tal importancia que las transforman en un instrumento indispensable del análisis estadístico.

En principio, la dispersión condiciona o limita el campo de variabilidad de un promedio.

4. ANALISIS DE SERIES CRONOLOGICAS

4.1 - La Tendencia Secular

4.1.1 - Ajuste de Tendencia Lineal.

La Ecuación de la Línea Recta

-Qué significa la ecuación de la línea recta?

La ecuación $Y = a + bX$ representa una forma generalizada de una línea recta en la cual no están especificados los valores a y b. Es decir, que esta ecuación sirve para representar cualquier línea recta.

Como en un plano existen infinitas rectas, si queremos representar a una recta particular -es decir, que ajuste a la tendencia de un fenómeno determinado-necesitamos conocer los valores de a y b para dicha recta. Veamos un ejemplo:

Supongamos que: $a = 5$
$b = 2$

Por lo tanto, la ecuación de la línea recta será: $Y' = 5 + 2X$

Ella nos permite:

1. - Calcular el valor de Y' que corresponde a cada valor de x.
2. - Conocer en cuánto se incrementa la tendencia para cada unidad de incremento de X. Esto es posible por el coeficiente b. Si b es un valor positivo, la tendencia crece; si b es negativo, la tendencia decrece.
3. - Conocer el promedio de valores en todo el intervalo considerado. Es el valor de a. Representa el valor de la tendencia cuando X - 0. Localiza la altura de la tendencia respecto al eje de las Y. Por lo tanto, el problema de ajustar una línea recta consiste en encontrar los valores

de a y b que sean los más apropiados; es decir, aquellos que en la ecuación den por resultado una línea recta que ajuste lo mejor posible a los valores empíricos de la serie de tiempo que se analiza.

El método más comúnmente usado para ajustar tendencias es el de los Mínimos Cuadrados.

Este método se basa en el principio que una tendencia ajusta mejor una serie de valores, cuando las constantes de la ecuación a y b se determinan de tal modo que, la suma de los cuadrados de los desvíos (entre los datos originales y los correspondientes valores de la tendencia), sea mínimo.

Si la línea de tendencia ajusta a los datos perfectamente, cada punto caerá sobre dicha línea y entonces, la suma de los cuadrados de los desvíos será cero. En fórmula: $\sum (Y - Y')^2 = 0$ Y = valores empíricos Y' = valores en tendencia.

Mientras más se alejen los puntos respecto de la tendencia, más grandes serán los desvíos, sin embargo, puede ajustarse una línea de modo que los desvíos al cuadrado, sean mínimos.

A fin de encontrar los valores a y b por el método de los mínimos cuadrados, es necesario resolver el siguiente sistema de ecuaciones en donde la variable tiempo se representa por X y los valores empíricos de la serie, por Y. El número de años del período es :

$$N \sum Y = N\,a + b \sum X$$
$$\sum x\,Y = a \sum X + b \sum X^2$$

La solución de este sistema se simplifica grandemente si se considera que el tiempo está uniformemente espaciado sobre el eje de las x.

En vez de considerar al primer año de la serie, como origen del tiempo (año cero), se asigna un nuevo origen ubicado exactamente en el centro de la serie. De este modo, cada año calendario se asócia con un valor de la nueva escala de tiempo. Ejem.:

AÑOS	X
1955	- 3
1956	- 2
1957	- 1
1958	0
1959	1
1960	2
1961	3

De esta forma resulta: $\sum X = 0$

En consecuencia, las ecuaciones se reducen así: $\sum y = Na$

Ahora resulta muy fácil despejar los valores deseados a y b $\sum xy = b\sum x^2$

$$a = \frac{\sum y}{N} \qquad y \qquad b = \frac{\sum xy}{\sum x^2}$$

4.1.2 - Cálculo de los Valores de Tendencia

Una vez que se conocen los valores a y b, se reemplazan en la ecuación $Y' = a\ b\ X$, para calcular los valores reales de la tendencia. Para el cálculo se trabaja con la siguiente planilla.

1) Para un número impar de años

AÑOS	Consumo de Harina Per Cápita en Mendoza			
	X	Y	X^2	XY
		Kg. x habit.		
1953	-4	80,3	16	-321,2
1954	-3	82,1	9	-246,3
1955	-2	79,7	4	-159,4
1956	-1	80,0	1	- 80,0
1957	0	75,5	0	0
1958	1	78,3	1	78,3
1959	2	78,6	4	157,2
1960	3	73,2	9	219,6
1961	4	67,5	16	270,0
	0	695,2	60	-81,8

$$a = \frac{\sum Y}{N} = \frac{695,2}{9} = 77,24 \text{ kg. x hab.} \; ; \; b = \frac{\sum XY}{\sum X^2} = \frac{-81,8}{60} = -1,36$$

$$y' = a + bx = 77,24 \text{ kg.} - 1,36\ X$$

Los valores de la tendencia se encuentran sustituyendo los valores pertinentes de X en la ecuación de tendencia. Como una línea recta queda suficientemente determinada con ubicar sólo 2 puntos en el plano, basta efectuar sólo 2 reemplazos. Ejemplo:

AÑO	X	Y Kg. x habit.	Tendencia $Y' = 77,24 - 1,36\ X$
1954	-3	82,1	$Y' - 3 = 77,24 + [(-1,36)\ (-3)] = 81,32$
1960	+3	73,2	$Y_3 = 77,24 [(1,36) . 3] = 73,16$

2) Para un período con número par de años:

-3,5
-2,5
-1,5
-0,5
0,5

Uso de la tendencia como valor normal.

En muchas situaciones prácticas la tendencia se usa para:
a) predecir valores futuros.
b) determinar niveles normales o de equilibrio.
c) determinar valores promedios respecto al tiempo.

Por todo ello, la tendencia puede ser considerada como nivel de "equilibrio" , "normal" o "valor esperado" de la serie.

Cuando se usan valores anuales, los valores de la tendencia se usan como "Normal". Así por ejemplo, si en un año determinado, la producción o el consumo, alcanzan valores superiores a los de la tendencia, se dice que la producción -o el consumo- están por encima de la normal y viceversa.

Esta información es de gran utilidad para un empresario o gobernante ya que le permite juzgar la posición actual y además le puede dar alguna orientación sobre lo que puede esperarse en el futuro.

Procedimiento para usar los valores de la Normal.

Cuando una actividad-producción, consumo, trabajo, etc.- está afectada por variaciones estacionales, el conocimiento de los "picos" de máximo o mínimo, es indispensable para disponer la acción a realizar. Por ej. en una fábrica se aprovechará para licenciar al personal, efectuar reparaciones, producir stocks, etc. Si dichas variaciones son importantes, la estimación del nivel Normal o de Equilibrio de la serie, deberá tomar en cuenta al mismo tiempo los valores de tendencia y variación estacional. El

procedimiento más simple en este caso consiste en: Multiplicar los valores de la tendencia mensual por un índice de variación estacional.

4.2 - Variaciones Estacionales

4.2.1 - Utilidad e Interpretación de las Medidas de Variación Estacional.

Es obvio que con fines de análisis y control, las grandes empresas y gobiernos necesitan información sobre las variaciones estacionales de ciertos hechos. Cítanse a modo de ejemplo: la necesidad de preparar presupuestos mensuales de:

- abastecimiento de materias primas
- abastecimiento de artículos y servicios de consumo
- disponibilidad de mercaderías en existencia
- necesidad extras de mano de obra
- ocupación de equipos
- producción y distribución de artículos
- previsión de precios

Durante las estaciones o períodos de poca actividad se pueden otorgar vacaciones, reparar equipos, modificar planes de organización de propaganda, etc., ensayar la producción de nuevos artículos, etc.

Desde todo punto de vista, es razonable disponer de estimaciones precisas de las variaciones estacionales para confeccionar inteligentemente los planes internos en los negocios de gran escala.

Los indicios de variación estacional son importantes en distintos niveles de análisis: El economista, el comerciante, el hombre de la calle, frecuentemente usan varios índices como barómetros de la actividad económica.

Así, el movimiento estacional del Indice de Producción Industrial, es muy importante. Una declinación en la producción puede indicar: desempleo, fábricas ociosas, sin trabajo, pérdida de materiales, etc.

Es importante conocer si una declinación en un índice se debe a una variación normal debido a la estación o es el resultado de cambios fundamentales en la Economía. Por tal razón es costumbre ajustar los índices corrientes de producción mediante los índices de variación estacional. Suele a veces publicarse datos en forma tal que se muestra la serie ajustada y la no ajustada a las variaciones estacionales.

La serie "ajustada" muestra cómo debieran aparecer los datos en ausencia de las variaciones estacionales.

El Ajuste se efectúa dividiendo el dato original por el índice de variación estacional. (Ver ejemplo)

El elemento de estacionalidad se elimina del dato original: (Y) tal como se indica a continuación:

$$\frac{Y}{E} = \frac{T \times E \times C \times A}{E} = T \times C \times A$$

Los datos de esta naturaleza suelen también ajustarse frecuentemente por otros conceptos, tales como: el número de días de cada mes o las variaciones en el número de días laborables de cada mes.

Ejemplo:

INDICE DE PRODUCCION INDUSTRIAL DE EE. UU. EN 1953

Base 1947 - 49 = 100

MES	No Ajustado	Ajustado
Enero	132	134
Febrero	136	134
Marzo	138	135
Abril	136	136
Mayo	136	137
Junio	136	136
Julio	129	137
Agosto	136	136
Septiembre	135	133
Octubre	136	132
Noviembre	130	129
Diciembre	124	126

Una vez que se dispone de ambas series, se grafican juntas con distinto trazo. Las diferencias que quedan entre ambas después del ajustamiento, probablemente se deberán en gran parte a factores erráticos o al hecho de que el índice típico de variación estacional determinado, no represente precisamente la variabilidad estacional en un año particular.

No hay que olvidar además, que el impacto de las estaciones sobre el sistema económico puede cambiar en el tiempo, por diferentes circunstancias. Por ejemplo, durante la segunda guerra mundial cuando muchas industrias podían vender todo lo que producían prescindiendo de las estaciones, el índice de variación estacional tenía poca importancia. Por otra parte, los cambios a la tecnología y en los gustos y hábitos de los consumidores, pueden operar modificaciones en los patrones estacionales de la industria.

4.2.2. Metodología para la Determinación de Indices de Variación Estacional.

Los índices de variación estacional se calculan en el supuesto que los datos primarios de cada mes, son el producto de cuatro tipos de movimientos:

. Tendencia	T
. Variación cíclica	C
. Variación estacional	E
. Variación accidental	A

Por lo tanto, los datos de cada mes pueden ser representados por:

$$Y = T \times C \times E \times A.$$

Para determinar variaciones estacionales se procede de la siguiente manera:

1o. Arreglar los datos en una columna continua, sin espacios entre años columna (a) de la Tabla Nº 1.

Tabla No. 1

Años y Meses	Datos originales	Total de 12 meses	Promedio móvil (b): 12 = (c)	% del Promedio móvil (a): (c) =d
	(a)	(b)	(c)	(d)

2o. A partir del primer mes de la serie (que debe tener alrededor de 10 años, o sea, 120 datos mensuales) se calcula la suma de 12 meses para tener un promedio móvil de 12 meses que se anota en otra columna paralela, frente al mes de julio. Columna (b).

3o. Obtener la suma de otros doce meses, restando a la suma anterior el mes de enero y agregando el mes de enero del año siguiente, la nueva suma se coloca frente a agosto. Seguir sumando siempre grupos de 12 meses restando uno y agregando el similar del año siguiente.

4o. Calcular los promedios móviles dividiendo por 12 cada una de las sumas obtenidas en la columna anterior. Columna c.

5o. Dividir los datos originales por el promedio móvil y multiplicar por 100 para expresar las cifras como por cientos. (Columna d). Mediante estos co-

cientes, se eliminan las componentes de tendencia y variación cíclica, según la fórmula.

$$\frac{T \times C \times E \times A}{T \times C} = E \times A$$

Graficar estos porcientos para observar si existe una estabilidad razonable. Si se advierte estacionalidad, continuar con el paso siguiente:

6o. Calcular un promedio para cada uno de les meses (o trimestrastros) de los datos estacionales en porcientos obtenidos en el paso 5o. Este valor medio o promedio puede obtenerse por el procedimiento para calcular media aritmética. No obstante, es más aconsejable calcular la mediana.

Para obtener la mediana se procede así:

a) - Se ordenan los valores (los de un mismo mes en todos los años, ejemplo: todos los porcientos de enero, febrero, etc.

Tabla No. 2

AÑO	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Set.	Oct.	Nov.	Dic.

b) Ordenar de menor a mayor los datos de cada mes

Tabla No. 3

ORDEN	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Set.	Oct.	Nov.	Dic.
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
-Total de líneas de 4 a 10 -Mediana -Medianas niveladas.												

c) - Eliminar, si existen, un número dado de valores extremos. Este número de cifras a eliminar debe ser determinado a criterio del investigador; después de un examen de la conducta año a año de las cifras específicas estacionales.
Por ejemplo, podría eliminarse las 3 filas de estos extremos; tanto los 3 más bajos como los 3 más altos.

d) - Sumar los porcientos restantes para cada mes (Ej. líneas desde la 4 a 10 si se habían tomado 13).

e) - Dividir por el número de porcientos tomados en consideración y obtener el promedio correspondiente a cada mes. (Como una alternativa, puede tomarse directamente el valor central de la serie de cada mes).

7o. Nivelar las medias. Esto tiene por objeto eliminar los componentes accidentales y dejar sólo la estacional. Para ello proceder así:

a) - Sumar todos los datos de Medianas Mensuales (o medianas trimestrales).

b) - Dividir este total por el número de medianas incluidas. (12 si se trabajó con meses y 4 si se trabajó con trimestres).

c) - Dividir cada mediana por el promedio anual obtenido en el paso anterior. Con esta operación se obtienen las medianas niveladas e índices de variación estacional típica.

Es necesario verificar este cálculo de modo tal que, promediando estos índices se obtengan 100 como promedio del año.

8o. Graficar las medianas niveladas.

4.2.3.- Medida de la Dispersión

Los índices mensuales son promedios muestrales de un período más o menos largo y, por tanto, están sujetos a aviación. Para calcular la medida de esta variación es necesario determinar el error standard que se obtiene con la fórmula:

$$\text{Error standard} = \sqrt{\frac{\sum x^{-2} - \frac{(\sum x)^2}{n}}{n(n-1)}}$$

donde: X = índice medio del mes
X = índice de un mes en un año particular
n = número de años para los cuales se observó el índice.

El error standard para todos los meses y todos los indices, se presenta en la tabla N° 4

Tabla N° 4

Errores Standard de Indices Estacionales

ARTICULO	Ene.	Feb.	Marz.	Abr.	May.	Jun.	Jul.	Ago.	Set.	Oct.	Nov.	Dic.
Faena de vacas	2,0	1,8	1,6	1,9	2,3	2,0	1,2	1,1	2,3	3,1	3,0	2,6

Deben ser interpretados como sigue: Existe una probabilidad de 95% que el verdadero índice de cada mes, caiga en un intervalo determinado por dos errores standards a ambos lados del índice mensual observado. Es decir:

$$P \cdot y\ (X + 2\ Es) = 95\%$$

Ejemplo:

El índice observado para la faena de vacas en noviembre es134, según la tabla 3.

El error standard en la tabla No. 4 es 3, 0.

Dos veces este error standard es 6, 0.

Luego hay una probabilidad de 95 en 100 que el intervalo 134 + 6 contenga el verdadero índice del mes de enero. O sea, que estará comprendido entre 128 y 140.

El índice verdadero, es el que resultaría si se observan todos los ítems de la población. En el muestreo, las observaciones se han hecho solamente sobre una parte de los datos del universo.

Economic Division, Canada Department of Agriculture, Seasonal Variations in the hivestock Industry. (Otawa, Ontario), 1961. Biblioteca I.I.E.T. No. 390

Paden D. and Lindquist, Statistics for Economics and Business (New York 1956).

5. NUMEROS INDICES

Indices de la Actividad Industrial

El objeto que se persigue con los números índices es el de medir rápidamente y a cortos períodos los cambios que experimente la producción industrial. Se tiene así, a mano, un poderoso recurso para saber cómo marcha la industria permitiéndonos poseer en forma continua y actualizada un inapreciable elemento de comparación, no sólo entre las diversas industrias entre sí, sino también en cualquier otra manifestación económica con la que esté conexa.

En primer análisis, el índice de la industria permite evaluar qué importancia y significado tienen los cambios producidos en el volumen de la producción en relación con los que se operan en la población, la renta bruta, el comercio exterior, los precios, la producción, etc. Se infiere de aquí el valor excepcional que tiene un índice de producción industrial en cualquier estudio económico, ya que aquella conforma su elemento más dinámico y de carácter más variable.

Las Naciones Unidas por medio de sus oficinas especializadas, recomienda que todos los países que tienen industrias importantes computen un índice de producción industrial, calculado según un criterio uniforme, en beneficio de la comparabilidad en el campo internacional. Procediendo así se podrá tener incluso un índice de la producción industrial del mundo, que es el objeto final que en definitiva, persigue la mencionada institución.

Elementos Integrantes

La razón de que un índice de producción conocido también como de "volumen físico", es el más importante, se extractará seguidamente el planteo y modo de cálculo que ha adoptado la Dirección General de Estadística y Censos de Argentina.

El índice de actividad industrial de la Argentina se calcula mensualmente para los cuatro conceptos:

- obreros ocupados
- horas obrero trabajadas
- monto de salarios pagados
- volumen físico de la producción industrial.

Cada uno de ellos se elabora para las tres grandes divisiones de la actividad industrial y sus diversas ramas:

1o. Explotación de minas y canteras. (Industrias Extractivas).
2o. Industrias manufactureras y,
3o. Producción de Energía. (Electricidad y Gas).

Están excluidos por consiguiente, la producción agrícola y forestal, las construcciones civiles y caza marítima.

No obstante que las recomendaciones internacionales propenden a que se incluya en el índice de producción industrial las construcciones civiles, la Dirección Nacional de Estadística y Censos ha mantenido y mantendrá su criterio de excluir esa actividad en los cómputos ya que su consideración introducirá complejidad e incertidumbre en los cálculos en virtud de la modalidad tan particular que tienen las industrias de las construcciones.

El segundo de los sectores mencionados (el de las industrias manufactureras) incluye:

1. Industria de la alimentación y bebidas.
2. " del tabaco
3. " textil
4. " de la confección
5. " de la madera
6. " del papel y cartón
7. " imprentas y publicaciones
8. Fabricación de productos químicos
9. Obtención de productos derivados del petróleo
10. Fabricación de productos de caucho
11. Industrias del cuero
12. " de la piedra, vidrio y cerámica
13. " de metales, excluída maquinaria
14. " de la fabricación de vehículos y maquinaria no eléctrica.
15. Fabricación de maquinaria y artefactos eléctricos
16. Industrias manufactureras diversas.

Esta clasificación difiere algo de la Clasificación Internacional Uniforme de todas las actividades Económicas para la Industria, vigente en Naciones Unidas, pero era la usual en la época de la estructuración del índice.

Fórmula Adecuada.

El cómputo del número índice se efectuará por aplicación de la conocida fórmula de Laspeyres (promedio aritmético de relativos de cantidades ponderados por valores monetarios de la producción del año base).

Su expresión analítica resulta ser un cociente de dos sumatorias en las que intervienen: La producción actual en relación con la que correspondió al año base, afectados por los valores monetarios de la producción de la base, en el numerador, y en el denominador, la suma de todos los valores básicos de la producción en la rama que comprende el producto.

En la mecánica del cálculo, el número que se obtiene por razón entre el valor de la producción del artículo para la base, y el valor de la producción de toda su rama, es lo que corrientemente denominamos como coeficiente de ponderación del relativo de artículos.

Simbólicamente se le representa:

$$\equiv \frac{\sum \frac{q_n}{q_0} p_0 q_0}{\sum p_0' q_0}$$

en la que "q" tiene el significado de la cantidad de trabajo efectuado o volumen de producción y "p" el valor aportado por unidad de trabajo efectuado.

Información Básica

En la disponibilidad de los valores resultantes del producto de los "p" por las "q" estriba la gran dificultad de la aplicación de la fórmula.

En efecto: en la medición de la producción industrial es posible seguir dos caminos: producción bruta o neta.

La primera se refiere a la cantidad de productos terminados por los establecimientos, tomados a precios finales.

La producción neta mide el valor aportado por las industrias en función de la diferencia entre el valor final del producto elaborado y los valores recibidos de otras industrias (materias primas, combustibles, energía eléctrica).

Un índice de producción bruta que abarque la totalidad de la industria está sujeto a duplicaciones, por cuanto la producción terminada de un establecimiento suele ser materia prima de otros que la utilizan tal como la reciben o la someten a otros precesos de elaboración.

A su vez, la medida de la producción neta de una industria, resulta prácticamente irrealizable ya que no es posible disponer del valor agregado correspondiente a cada serie de productos elaborados.

La separación de los factores comunes que concurren a determinar el precio de costo de un producto, en los casos de establecimientos de producción diversificada, implicaría una tarea tan ardua como incierta en sus resultados. Por tal motivo, el índice que se elabora en la Dirección Nacional de Estadística y Censos, tiene características de producción bruta para ramas, y neta para subgrupos, grupos y nivel general.

Emergente del concepto establecido, en los índices de ramas los referidos valores monetarios que se usan como ponderación, son los de la producción bruta. En cambio,

son netos (valor aportado) para las ponderaciones en las agrupaciones de orden superior.

Como no siempre es posible medir la producción por la cantidad de productos elaborados, en algunos casos se utilizaron series de sustitución, tomando en cuenta, en lugar de la producción, el atributo que mejor podía representarla (Consumo de principales materias primas, horas-obrero trabajadas, etc.).

Los establecimientos de la muestra que han servido para la obtención de los datos de producción, ocupación, salarios, horas-obrero trabajadas y ausentismo, se seleccionaron del total de los establecimientos existentes en el año 1943 (dato censal) y se consultan actualmente, 1909 establecimientos correspondientes a 91 ramas de la industria.

Con los informes solicitados se construyen los siguientes índices: 258 de producto, 91 de índices de rama, 16 índices de subgrupos, 3 índices de grupo y 1 de nivel general.

Los números índices se computan y publican mensualmente por la Dirección Nacional de Estadísticas y Censos.

Esquema de Ponderaciones

A modo de ejemplo, se indica a continuación las ponderaciones que indican la importancia relativa de cada grupo, subgrupo y rama, en el conjunto industrial de Mendoza según el Censo Económico de 1954.

ENCUESTA INDUSTRIAL

BASES DE PONDERACION

NIVEL GENERAL, GRUPOS, SUBGRUPOS Y RAMAS

NIVEL GENERAL	GRUPO	SUB GRUPO	RAMAS
	Producción Neta	Producción Neta	Producción Bruta
NIVEL GENERAL	100,00		
I - INDUSTRIA EXTRACTIVA	14,0	100,00	
A - Petróleo, Yacimientos		100,00	
II - INDUSTRIA MANUFACTURERA	82,8	100,00	
A - Alimentos y bebidas		71,8	100,00
1 -Aceites comestibles,fab.y ref.			1,4

2 -Aguas gaseosas, bebidas s/alcohol y cerveza			3, 9
3 -Frutas y leg. secas y conser (*)			17, 4
4 -Harina y/o produc. de molienda de trigo			11, 5
5 -Pan y/otros produc. elab. en panaderías			2, 7
6 -Sidra			1, 3
7 -Vinos, elaboración (bodegas)			71, 8
B - Confecciones		1, 6	100, 00
1 -Ropa exterior p/homb. o niño (sastrería)			100, 00
C - Madera		3, 8	100, 00
1 -Cajones para envase y embalaje			40, 1
2 -Muebles inclusive los de mimbre			28, 0
3 -Puertas y ventanas, marcos, etc.			31. 9
D - Productos Químicos		4, 3	100, 00
1 -Alcohol, destilerías y desnatu.			30, 1
2 -Subst. y produ. quím. y farmacéuticos, no mencionados especialm.			
E - Piedras, vidrio y cerámicas		7, 8	100, 00
1 - Mosaicos, Cal y Cemento			82, 7
2 - Vidrios y crist. en diversa form.			17, 3
F - Metales, inclusive maquinarias		1, 8	100, 00
1 -Talleres mecánicos (excepto auto)			100, 00
G - Vehículos y maq. exc. la eléctri.		8, 9	100, 00
1 -Máq. y motores, armado y reparac.			34, 9
2 -Talleres mecánicos p/automotores y vulcanización.			65, 1
III - ELECTRICIDAD Y GAS	3, 2	100, 00	
A - Fábricas de electricidad		100, 00	

FUENTE: Censo Industrial año 1954.

(*) Incluye: Dulces, mermeladas y jaleas.

(**) Comprende: Carburo de calcio, Fósforos, Sulfato de cobre, Acido tartárico, Cremor tártaro, Sulfuro de calcio, Tartrato de cal y Fertilizantes.

Capítulo III

CORRELACION Y ASOCIACION

1. CORRELACION

"Si dos fenómenos están relacionados de tal forma que para determinadas modalidades de uno, el otro necesariamente presenta determinados valores, se dice que entre ellos existe correlación. Por el contrario, si para la presentación de las modalidades de uno es indiferente la condición de las del otro, decimos que son independientes.

Todos los grados de relación intermedios entre la independencia y la dependencia funcional, constituyen correlación, admitiéndose por tanto, distintas intensidades para ésta; así podemos hablar, por ejemplo, de una correlación muy débil o de una correlación muy fuerte, según que el grado de relación sea muy débil o muy intenso". (M. García y J. Ayuso).

Estos fenómenos pueden ser por ejemplo:

a) El nivel de precios mayoristas y el nivel de precios minoristas.
b) Consumo de combustible y producción industrial.
c) Edad del esposo y de la esposa en la distribución de cuplas matrimoniales.
d) Proporción de jóvenes menores de 18 años y gastos en educación.
e) Tasas de Natalidad y Mortalidad.
f) Edad y producción de vacas lecheras.

Cuando analizamos las relaciones entre dos variables solamente, hablamos de CORRELACION SIMPLE. Cuando interesan las relaciones entre más de dos variables -por ejemplo, entre cantidad consumida de carne, el precio y el ingreso de los consumidores- hablamos de CORRELACION MULTIPLE.

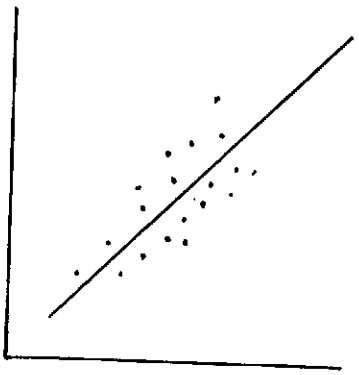
1.1. Correlación Simple

1.1.1. -Correlación de Series Simples (Datos no Agrupados)

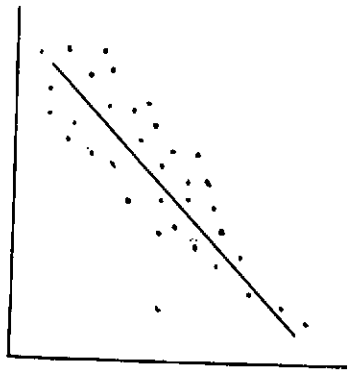
Si indicamos con X y Y las dos variables en consideración, un diagrama de dispersión nos mostrará la localización de los puntos (X, Y) sobre un sistema de ejes coordenados.

Si todos los puntos parecen estar alrededor de una línea recta, la correlación se llama lineal y en tales casos, si se ajusta a dichos puntos, una línea mediante la función de la recta, podremos "estimar" los valores de una variable con sólo conocer los de la otra.

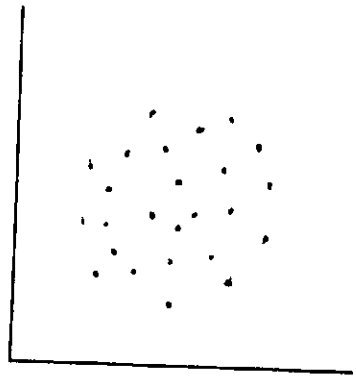
Cuando un crecimiento de Y está asociado con un crecimiento de X, se dice que la correlación es POSITIVA. En cambio, si un decrecimiento de Y está asociado con un crecimiento de X, tendremos una CORRELACION NEGATIVA. Si los puntos se ubican desordenadamente, sin una dirección determinada, se dice que las variables no están correlacionadas.



CORRELACION POSITIVA



CORRELACION NEGATIVA



NO CORRELACION

Si los puntos parecen agruparse, no alrededor de una recta, sino siguiendo alguna curva, se dice que la correlación es NO LINEAL, o CORRELACION CURVILINEA, y entonces, se necesitará la función de una curva para poder efectuar las estimaciones.

a) Rectos de Regresión

Si dos variables X e Y están relacionadas por una relación lineal o ley en línea recta, la ecuación que exprese dicha relación será de la forma:

$$Y = aX + b$$

donde:

a..... es el parámetro que indica la pendiente de la recta.
b..... es el parámetro que indica en qué punto la recta corta el eje de las Y.

El valor de "a" o pendiente, expresa: "cuánto crece Y para cada aumento unitario de X".

Siempre que exista una tendencia en línea recta entre 2 variables X e Y, estaremos en condiciones de encontrar valores para los dos parámetros a y b, que nos den la recta de ajuste de los puntos del gráfico.

Debemos encontrar la recta que mejor ajuste, es decir, la que tenga el mejor valor de pendiente y el mejor de intersección.

Hay muchos criterios para definir al "mejor de los ajustes", pero el más generalmente aceptado es el de los "mínimos cuadrados".

La recta de regresión por mínimos cuadrados de Y sobre X, es:

$$Y = a_0 + a_1 X$$

donde, los parámetros a_0 y a_1 se obtienen de las ecuaciones normales:

$$\begin{cases} \sum Y = a_0 N + a_1 \sum X \\ \sum XY = a_0 \sum X + a_1 \sum X^2 \end{cases} \quad (1)$$

Resolviendo el sistema, podemos encontrar los valores:

$$a_0 = \frac{\begin{vmatrix} \sum Y & \sum X \\ \sum XY & \sum X^2 \end{vmatrix}}{\begin{vmatrix} N & \sum X \\ \sum X & \sum X^2 \end{vmatrix}} = \frac{\sum Y \cdot \sum X^2 - \sum X \cdot \sum XY}{N \sum X^2 - (\sum X)^2}$$

$$a_1 = \frac{\begin{vmatrix} N & \sum Y \\ \sum X & \sum XY \end{vmatrix}}{\begin{vmatrix} N & \sum X \\ \sum X & \sum X^2 \end{vmatrix}} = \frac{N \cdot \sum XY - \sum X \cdot \sum Y}{N \cdot \sum X^2 - (\sum X)^2}$$

También existe una línea que explica la relación de X sobre Y cuya expresión es:

$$X = b_0 + b_1 Y \quad (4)$$

donde: b_0 y b_1, se obtienen de las ecuaciones normales:

$$\begin{cases} \sum X = b_0 N + b_1 \sum Y \\ \sum XY = b \sum Y + b_1 \sum Y^2 \end{cases}$$

resolviendo el sistema de ecuaciones, se obtienen los parámetros:

$$b_0 = \frac{\begin{vmatrix} \sum X & \sum Y \\ \sum XY & \sum Y^2 \end{vmatrix}}{\begin{vmatrix} N & \sum Y \\ \sum Y & \sum Y^2 \end{vmatrix}} = \frac{\sum X \cdot \sum Y^2 - \sum Y \cdot \sum XY}{N \cdot \sum Y^2 - (\sum Y)^2} \quad (5)$$

$$b_1 = \frac{\begin{vmatrix} N & \sum X \\ \sum Y & \sum XY \end{vmatrix}}{\begin{vmatrix} N & \sum Y \\ \sum Y & \sum Y^2 \end{vmatrix}} = \frac{N \cdot \sum XY - \sum X \cdot \sum Y}{N \cdot \sum Y^2 - (\sum Y)^2} \qquad (6)$$

El cálculo de los parámetros puede simplificarse mucho si adoptamos como artificio de cálculo, un origen distinto para cada variable.

Dicho artificio consiste en trasladar el origen de la variable de cálculo, al valor medio y luego, trabajar con los desvíos:

$$x = X - \bar{X}$$
$$y = Y - \bar{Y}$$

Sabemos, por una propiedad de la media aritmética, que la suma de los desvíos es igual a cero. En consecuencia, en las ecuaciones (2), (3) (5) y (6), se anulan todos los términos $\sum X$, $\sum Y$

En consecuencia, las ecuaciones de las rectas de ajuste (1) y (4) pueden escribirse:

$$Y = \left(\frac{\sum xy}{\sum x^2} \right) X \qquad (7)$$

$$X = \left(\frac{\sum xy}{\sum y^2} \right) Y \qquad (8)$$

donde: $x = X - \bar{X}$
$y = Y - \bar{Y}$

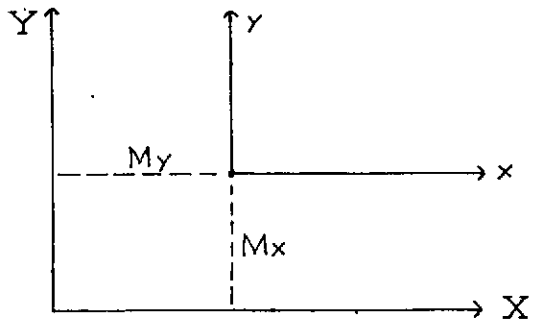
Las ecuaciones de regresión son idénticas si, y sólo si todos los puntos del diagrama de correlación caen sobre una línea. En tal caso, hay una correlación perfecta entre X e Y.

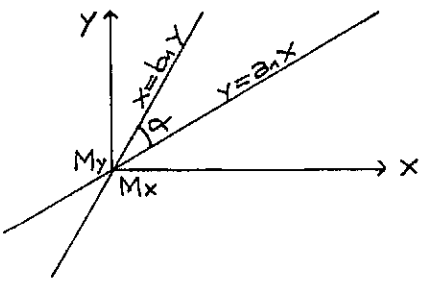
Las rectas de regresión fueron empleadas por Galton en sus estudios de relación entre la estatura de los padres y de los hijos.

Las dos rectas de regresión forman un ángulo. Cuando no existe correlación, el ángulo que se forma entre ambas es un ángulo recto.

Cuando las dos rectas coinciden, la correlación es funcional y en ese caso, el ángulo es cero. Entre estos dos extremos se presenta la mayoría de los casos. Mientras menor sea la abertura del ángulo mayor es la correlación.

Si se trabaja con desvíos, las ecuaciones de regresión se cruzan en el baricentro del sistema que es un punto formado por (X Y). Este es el origen del nuevo sistema de ejes y por ello, las rectas pasan por el origen y no tienen término independiente.





b) Coeficiente de correlación:

El coeficiente de correlación r es la media geométrica de los coeficientes de regresión.

$$r = \sqrt{a_1 \cdot b_1} \qquad (9)$$

También se expresa con la forma más usual del "momento-producto".

$$r = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} \qquad (10)$$

Esta fórmula da automáticamente el signo propio de r.

El coeficiente r, no puede tener un valor que exceda de + 1 ni tampoco ser menor que - 1. Por lo tanto; si:

r = + 1 . . . Hay relación funcional perfecta directa entre X e Y.

r = - 1 . . . " " " " inversa o negativa.

r = 0 . . . No hay correlación.

Cualquier otro valor intermedio de r indica que, aunque no se observa una relación funcional estricta entre las variables, existe sin embargo una tendencia, o sea que, en algún grado están asociadas.

No interesa cuál sea el valor de X o de Y que se tome como origen ni la clase de unidades que se utilicen. El valor de r se conserva constante para una distribución dada a pesar de cualquier cambio que se haga en las variables.

El coeficiente r, no prueba la causalidad entre X e Y, sino que la mide.

El cálculo de este coeficiente puede hacerse:

a) antes del ajustamiento de las líneas de regresión, usando la fórmula del momento producto

b) o bien, después del ajustamiento de las líneas de regresión usando la media geométrica de los coeficientes angulares de ambas rectas.

Existen también otros modos de expresar el coeficiente de correlación r, partiendo de las fórmulas anteriores:

-Dividiendo numerador y denominador en la fórmula del producto-momento, expresamos a r, en función de la covarianza y varianzas de X e Y.

Siendo: $\frac{\sum xy}{N} = \frac{1}{N}\sum (X-\overline{X})(Y-\overline{Y})$ covarianza de x e y.

$$\sigma_x^2 = \frac{\sum x^2}{N}$$ varianza de x

$$\sigma_y^2 = \frac{\sum y^2}{N}$$ varianza de y

Resulta:

$$r = \frac{\sum xy}{N\,\sigma_x \cdot \sigma_y} \qquad (11)$$

-Generalmente se usa como fórmula de cálculo, la equivalente que use los valores naturales en vez de los desvíos.

$$r = \frac{N\cdot\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}} \qquad (12)$$

c) Error standard de la estimación.

Si representamos con Y_e los valores de Y, estimados para cada valor de X, mediante la recta de regresión, se puede calcular una medida de la dispersión de los valores alrededor de la recta de ajuste mediante la fórmula:

$$\sigma_{yx} = \sqrt{\frac{\sum (Y - Y_e)^2}{N}} \qquad \ldots\ldots\ldots\ldots \quad (13)$$

Esta fórmula nos da el ERROR STANDARD de la estimación de Y respecto a X.

Del mismo modo, el error standard de la estimación de X sobre Y, queda definida por la fórmula:

$$\sigma_{xy} = \sqrt{\frac{\sum (X - X_e)^2}{N}}$$

La ecuación (13) puede escribirse también así, para facilitar el cálculo:

$$\sigma_{yx}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

d) Propiedades del error standard

El error standard de la estimación tiene propiedades análogas a las de la desviación standard.

Por ejemplo, si se construyen líneas paralelas a la línea de regresión de Y sobre X, tomando como distancia vertical:

$1\sigma_{yx}$ si N es bastante grande estarán incluidos 68% de los puntos
$2\sigma_{yx}$ " " " " " " " 95% " " "
$3\sigma_{yx}$ " " " " " " " 99,7% " "

Como para las pequeñas muestras se usa como desviación standard modificada, la fórmula $\hat{\sigma} = \sqrt{\frac{N}{N-1}}\sigma$ para el caso de correlación de dos variables, en pequeñas muestras la fórmula modificada es:

$$\hat{\sigma}_{yx} = \sqrt{\frac{N}{N-2}}\sigma_{yx}$$

También suele usarse -en las pequeñas muestras- N-2 en lugar de N en la fórmula (13).

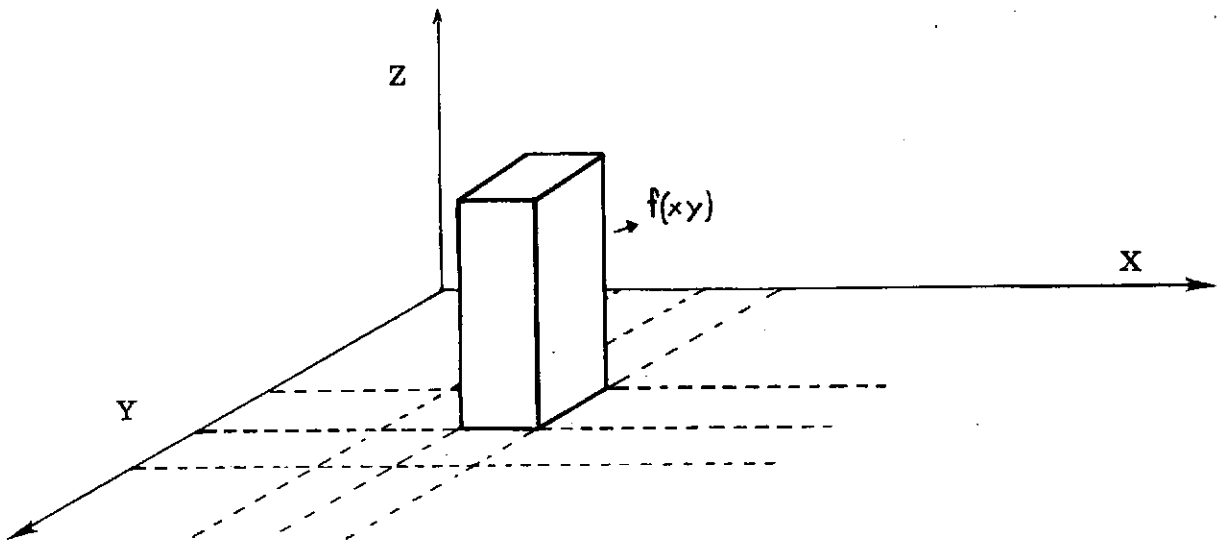
e) Variación explicada y no explicada.

La variación total de Y se define como $\sum (Y-\overline{Y})^2$ es decir, "la suma de los desvíos al cuadrado de los valores de Y con respecto a la media". También puede escribirse:

$$\sum (Y-\overline{Y})^2 = \sum (Y-Y_e)^2 + \sum (Y_e-\overline{Y})^2$$

1.1.2. Correlación en Distribución Bidimensional (datos agrupados)

Si existe una cantidad apreciable de observaciones simultáneas para los valores de X e Y, se construye una distribución bidimensional en una tabla de doble entrada. Los datos contenidos en ella se pueden representar gráficamente mediante el esterograma. Para ello se utiliza una terna de ejes ortogonales. Sobre el eje de las X y sobre el de las Y, se ubican los intervalos de cada variable y en el eje Z, las frecuencias.



En el cuadro de correlación, las columnas y las filas reciben el nombre de "array" (arreglo). Cada array es una serie simple de frecuencias y como a tal se le puede calcular la media aritmética.

Cada array horizontal y cada array vertical tiene una media aritmética. Su cálculo se efectúa sumando los productos de: puntos medios de cada intervalo, por las respectivas frecuencias marginales y dividiendo por el total de las frecuencias observadas en el array.

Representando las medidas de cada array (por ejemplo, ubicamos un pequeño círculo en el lugar que corresponde a las medias de las filas, y una cruz para las medidas de las columnas), tendremos una serie de puntos que pueden ajustarse con dos rectas, que reciben el nombre de rectas de regresión.

El cálculo de los parámetros de estas rectas se efectúa en la misma forma ya descripta, con la única diferencia que los valores de la variable deben ir siempre multiplicados por las respectivas frecuencias.

1.2. Aplicaciones Prácticas.

(Desarrollo completo de las órdenes planteadas en los ejercicios que como anexo forman parte de esta cartilla).

2. ESTADISTICA DE ATRIBUTOS : ASOCIACION

Correspondería llamarla con más propiedad: "estadística cualitativa".

Se refiere a la variación de atributos que no son medibles pero sí son medibles sus frecuencias.

Pearson hizo un estudio minucioso de este tipo con el color de los ojos para determinar las leyes de herencia. Es un atributo que no se puede medir por grados, por ejemplo; en el color de los ojos sólo pueden considerarse las variaciones negro, castaño, celeste, verde, etc.

En laboratorios médicos se usa mucho para los experimentos sobre vacunas y drogas en sus distintas aplicaciones y resultados.

Ej.: aplicación de vacuna y salud. . . . . A B
aplicación de vacuna y muerte. . . . . A b
no aplicación de vacuna y salud . . . . a B
no aplicación de vacuna y muerte . . . a b

Llámase atributo dicotónico al que puede tomar dos características solamente, varón o mujer, argentino o extranjero.

Se indica con A cuando se cumple
a cuando no se cumple

Ej.: B argentino C soltero
b extranjero c casado

Cuando se desea expresar las frecuencias de un determinado atributo, se escribe la letra entre paréntesis (A) , (a).

TEOREMA: Un atributo más su contrario debe dar el total
(A) más (a) igual N

Podemos formar un atributo de segunda categoría o combinaciones entre distintos atributos.

(A B c) + (A B C) = (A B)
(A B C) + (A b C) + (A B c) + (A b c) = (A)

argentino hombre soltero
" mujer " "
" hombre casado
" mujer " "
= argentinos (total)

Esto es muy importante para el control de los censos.

Cualidades: 1) Independencia. Cada atributo se cumple independientemente del o de los otros.

Tenemos A y a
B y b, interesa saber si son o no independientes.

Vale decir, el hecho de cumplirse uno no influye en el cumplimiento del otro.

$$\frac{(AB)}{(aB)} = \frac{(A)}{(a)} = \frac{(Ab)}{(ab)}$$

La proporción de favorables sobre contrarios es la misma, ya sea que se cumpla o no B. Así podría ser la relación arg. varones, es la misma que arg.

Si la proporción dependiera del sexo, no serían independientes, pero como el sexo no influye (mayormente) los atributos son independientes.

Existe otro concepto semejante al de correlación, es el de contingencia.

Supongamos dos atributos en una tabla en que las frecuencias tienen 4 valores posibles

	(A)	(a)
(B)	(A B)	(aB)
(b)	(A b)	(ab)

Las exteriores (A) (B) (a) (b) son las frecuencias simples, las otras son las dobles.

Si hubiera variación funcional directa al cumplirse A , forzosamente se cumple B, por lo tanto (aB)= (Ab) = 0 , y el coeficiente resulta igual a 1. Si la correlación fuera funcional inversa, los atributos son excluyentes; si se cumple b, forzosamente se cumple A, y si B, a. Entonces:

$$(AB) = (ab) = 0 \quad \text{coef. } -1$$

Pero buscamos los casos intermedios cuando no hay correlación funcional. Buscamos un coeficiente capaz de medir esos grados intermedios. Elegimos:

$$Q = \frac{(AB)(ab) - (aB)(Ab)}{(AB)(ab) + (aB)(Ab)} \qquad (1)$$

Debe cumplir las condiciones : 1 contingencia funcional directa
-1 contingencia funcional inversa
0 no hay contingencia.

Veamos: Conting. funcional directa: (aB) = (Ab) = 0 reemplazando en (1)

$$Q = \frac{(AB)\ (ab)}{(aB)\ (Ab)} = -1$$

Satisface cuando A y B son dependientes.

Contingencia funcional inversa.

$$Q = \frac{-(aB)\ (Ab)}{(aB)\ (Ab)} = -1$$

No hay correlación en

Recordando que: $\frac{(AB)}{(aB)} = \frac{(A)}{(a)} = \frac{(Ab)}{(ab)}$ por la condición de independencia, es:

$$(AB)\ (ab) = (Ab)\ (aB)$$

luego Q = O

En el caso expuesto por Charlier del experimento de las vacunas, si el coeficiente resulta positivo y próximo a uno, revela la eficacia de las vacunas, en cambio si resulta próximo a cero no es tan eficaz y si resulta negativo es absolutamente ineficaz.

Veamos un ejemplo:

	Infectados	No inf.	
No inoculados	10	117	127
Inoculados	3	144	147
Total	13	261	274

Aplicando el coeficiente de asociación de Yule.

$$Q = \frac{(AB)\ (ab) - (Ab)}{(AB)\ (ab) + (Ab)} = \frac{10 \times 144 - 3 \times 117}{10 \times 144 + 3 \times 117} = 0,61$$

El error típico se calcula así:

$$E = \frac{1 - Q^2}{2} \sqrt{\frac{1_+ 1_+ 1_+ 1}{a \quad b \quad c \quad d}} = 1 - \frac{0,61^2}{2} \sqrt{\frac{1_+ 1_+ 1_+ 1}{10 \quad 117 \quad 3 \quad 144}}$$

$$E = 0,21$$

Es razonablemente cierto (alrededor del 95% de probabilidad) que el valor verdadero se encuentra dentro de un intervalo formado por $\pm$ 2 E , respecto al valor calculado.

Resultaría que el coeficiente de Asociación se encontraría entre 0, 61 $\pm$ 2 x 0, 21 o sea, entre 0, 19 y 1, 03. Pero como Q $\geq$ 1 , podemos decir que el verdadero valor se encuentra entre 0, 19 y 1, 0.

Bibliografía:

Yule y Kendall, Introducción a la Estadística Matemática (Aguilar)

Spiegel Murray, Statistics, Theory and Problems

Moroney, Hechos y Estadísticas (Eudeba)

Capítulo IV

ELEMENTOS DE CALCULO DE PROBABILIDADES

1. VARIABLE ALEATORIA.

La ganancia o pérdida que puede derivarse de un juego, es el ejemplo más sencillo de una variable (magnitud, cantidad) que puede tomar distintos valores con probabilidades dadas.

En estadística utilizaremos las variables aleatorias. La variable X es una letra susceptible de tomar diversos valores:

$$X_1 \ , \ X_2, \ X_3, \ \ldots\ldots \ X_k$$

Además de dar los valores de X es condición, en las variables aleatorias, de que a cada valor de la variable le corresponda una probabilidad:

$$P_1 \ , \ P_2, \ P_3, \ \ldots\ldots \ P_k$$

En este caso, en qué tema X valores se denomina variable discreta. K es el orden de la variable aleatoria. También podría tomar valores infinitos o continuos. Dar los valores y probabilidades de una variable aleatoria, es dar una ley de probabilidad.

Ejemplo: Si utilizamos un dado, las variables aleatorias serían:
1, 2, 3, 4, 5, 6 y sus respectivas probabilidades: 1/6, 1/6.

1/6, 1/6, 1/6, 1/6 y decimos que es una variable aleatoria de orden 6º, definida por los pares $X_i \ P_i$ de la magnitud X_i y la probabilidad P_i.

Si en lugar de uno arrojamos dos dados y sumamos los puntos X_i obtenidos en ambos, la suma X es una variable aleatoria de orden 11º definida totalmente por los pares $X_i \ P_i$, en la forma siguiente:

X_i = 2; 3; 4; 5; 6; 6; 7; 8; 9; 10; 11; 12

P_i = 1/36; 2/36; 3/36; 4/36; 5/36; 6/36; 5/36; 4/36; 3/36; 2/36; 1/36.

Supongamos jugar a cara o cruz, conviniendo en pagar 1 peso si sale cara; y $ 0 si sale cruz. En cada jugada definimos una variable aleatoria X de 2º orden (la variable es la ganancia del jugador) por medio de los 2 pares:

X = 1 ; o

P = 1/2 1/2

Si jugamos diez partidas, la ganancia total al cabo de las 10 partidas, es una variable aleatoria de 11º orden definida totalmente por los pares siguientes:

$$X = 0; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10$$

$$P = P_e ; P_1 ; P_2 ; P_3 ; P_4 ; P_5 ; P_6 ; P_7 ; P_8 ; P_9 ; P_{10}$$

Siendo tales probabilidades fáciles de calcular (problema de las pruebas repetidas). Podemos entonces, establecer la siguiente definición: de una manera general una variable aleatoria X es de orden n si puede tomar n valores diferentes, según que se presente uno de los acontecimientos excluyentes E_1; E_2 . . . . . E_n de probabilidades dadas:

P_1 P_2

P_1 P_e . . P_n , probabilidad cuya suma es igual a 1.

El conjunto de los pares X_i P_i define la ley de distribución o de probabilidad de la magnitud considerada.

Puede ser una variable continua como las estaturas de las personas. Así pueden tomarse todos los valores comprendidos entre 1,50 m y 2 m.

$$150 \leq X \leq 200$$

En esta función de variable continua, tenemos que tener la función de probabilidad f (X).

Las variables aleatorias son los elementos que utilizaremos tanto en el cálculo de probabilidades como en la Estadística. La condición para sustituir está en el teorema de Bernouilli.

2. IDEA DE MOMENTO:

Es una idea que resume casi todos los índices. Su nomenclatura es:

m_1 momento primero desde el origen

m_2 " segundo " " "

$$m_i = \sum_{i=1}^{K} x_i P_i = x_i P_i + x_2 P_2 + x_3 P_3 + \cdots x_F P_K$$

En el caso del dado sería:

$$m_i = \sum_{i=1}^{K} x_i P_i = 1.1/6 + 2.1/6 + 3.1/6 + 4.1/6 + 5.1/6 + 6.1/6$$

$$= 1/6\ (1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}$$

$$m_2 = \sum_{i=1}^{K} x_i^2 \cdot P_i = x_1^2 P_1 + x_2^2 P_2 + x_3^3 P_3 + \ldots . x_K^2 P_K$$

$$= 1^2 . 1/6 = 2^2 . 1/6 + 3^2 . 1/6 + 4.^2 1/6 + 5.^2 1/6 + 6.^2 1/6$$

$$= \frac{1}{6} \quad \frac{4}{6} \quad \frac{9}{6} \quad \frac{16}{6} \quad \frac{25}{6} \quad \frac{36}{6} = \frac{91}{6}$$

$$m_S = \sum_{i=1}^{K} x_i^3 . P_i =$$ concepto general de momento.

El primer momento suele llamarse esperanza matemática cuyo símbolo es:

$m_i = E(x)$ es el valor más probable de X.

Esperanza matemática, acostumbra definirse como: "producto de cierta suma: S por la probabilidad de ganarla".

Si X puede tomar los valores:

X_1 , X_2 , X_3 . . . . . X_n con probabilidades:

P_1 , P_2 , P_3 . . . . . P_n , siendo $\sum_{i=1}^{K} P_i = 1$

Se dice esperanza matemática de las X_i a la expresión:

$$E(x) = \sum_{i=1}^{K} x_i p_i = \frac{\sum x_i p_i}{\sum p_i}$$

En el caso del dado teníamos:

$$\sum (x) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$$

En el juego, la esperanza matemática es lo que uno espera ganar. Se dice que un juego es equitativo, cuando lo que se paga es la esperanza matemática. En ese caso, se multiplica la ganancia esperada por la probabilidad. En el caso de la ruleta, el juego no es equitativo porque lo que se paga es distinto que la ganancia esperada:

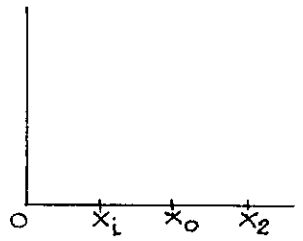
$$E(x) = \frac{36}{37}$$

Momento reducido. Estos índices suelen usarse referidos a otro origen, en la forma de X menos un cierto valor fijo; se los llama momentos reducidos y su símbolo es μ (mu).

μ_1 = momento primero
μ_2 = " segundo
μ_S = " subs

$$\mu_S = \sum (x - x_o)^S P_i$$

Representación gráfica:



En lugar de tomar como origen el 0 tomo el X_0. A veces tomamos el primer momento para referirnos a los demás, como origen y no el X_i para poder obtener la esperanza matemática.

$$\mu_S^{\cdot} = \sum (x_i - m_i)^S P_i = \text{momento centrado}$$

El momento reducido de orden O, es la suma de las probabilidades y habitualmente es la unidad (es 1 cuando se trata de casos exhaustivos, o sea que se agoten todas las posibilidades).

Si utilizamos el 1o. momento centrado, S = 1, tenemos:

1er. momento centrado

$$\mu_1 = \sum x_i p_i - \sum m_i p_i = m_i - m_i \sum p_i = 0$$

2o. " "

$$\mu_2 = \sum_{i=1}^{K} (x_i - m_i)^2 p_i = \sqrt{\ }^2 \quad (1)$$

Esto tiene importancia primordial. Se le llama desviación standard, dispersión o desvío medio cuadrático).

$$\mu_3 = \sum_{i=1}^{K} (x_i - m_i)^3 p_i$$

$$\mu_S = \sum_{i=1}^{K} (x_i - m_i)^S p_i$$

En realidad los momentos centrados son los mismos anteriores pero con otro origen. Hay tres momentos en los cuales se toma como origen una media provisoria (Charlier). Se denominan momentos reducidos.

μ'_1 = momento primero desde la media provisoria

μ'_2 = " segundo " " " "

La esperanza matemática suele llamarse media ponderada y en estadística es la media aritmética.
Hemos visto en (1) que el 2o. momento centrado es:

(Dispersión) La palabra dispersión no se utiliza para μ_2 sino para su raíz cuadrada o sea $\sqrt{\sqrt{}^2} = \sqrt{}$

$$\mu_2 = \sqrt{}^2 = \sum \underbrace{(x_i - m_i)}_{1}^2 p_i \quad \Big| \; (x_i - m_i)^2 = 2x_i m_i + m_i^2$$

desarrollando el 2o. miembro tenemos:

$$\sqrt{}^2 = \overbrace{\sum x_i^2 p_i}^{m_2} - 2m_i \overbrace{\sum x_i p_i}^{m_1^2} + \overbrace{m_i^2 \sum p_i}^{1} =$$

$$\boxed{= m_2 - 2m_1^2 + m_1^2 = m_2 - m_1^2}$$

$$\boxed{\mu_2 = \sqrt{}^2 = m_2 - m_i^2}$$ también se escribe así:

$$\boxed{\mu_2 = \sqrt{}^2 = E\left[X - E(x)\right]^2}$$

3. DISTRIBUCION BINOMIAL: PRUEBAS REPETIDAS

Consideremos una serie de pruebas análogas al juego de cara o cruz; o a una extracción de bolillas de una urna. Se trata de un acontecimiento que tiene las siguientes posibilidades: se cumple o no se cumple.

El acontecimiento favorable (se cumple) será por ejemplo la extracción de una bolilla blanca de probabilidad P.

Se cumple $1 = P$

No se cumple $0 = q$ luego $q = 1 - p$

Suponemos que se repone cada vez la bolilla extraída y que se hacen s pruebas (pruebas repetidas).

A la primera prueba asignaremos una magnitud X_i que tomará el valor 1, en caso de presentarse la bolilla blanca; O en caso contrario. Igualmente demos a las otras pruebas las magnitudes aleatorias

$$X_2 \quad X_3 \ldots\ldots\ldots X_s$$

Estas magnitudes son independientes. Su suma Z tiene por valor el número de acontecimientos favorables en una serie de S pruebas.

$$Z = x_1 + x_2 + x_3 + \cdots x_s \quad \text{(frecuencia absoluta)}$$

trataremos de hallar la esperanza matemática de Z, que significa el número de veces que se cumplió el hecho favorablemente.
Si llamamos F a la frecuencia relativa

$$F = \frac{Z}{n}$$

Si queremos establecer el momento primero o m_1 o esperanza matemática y el μ_2 o dispersión de Z y de F, debemos conocer los siguientes teoremas o propiedades de la esperanza matemática:

1) la esperanza matemática de una suma de variables aleatorias, es igual a la suma de sus respectivas esperanzas matemáticas.
2) la esperanza matemática de un producto de variables independientes es igual al producto de las esperanzas matemáticas.

   Lo aceptaremos sin demostración (ver Darmois)

   Cuál es la esperanza matemática de un valor de X?

$$x_i : E(x_i) = 1p + 0q = p \quad (1)$$

$$x_i^2 : E(x_i^2) = 1^2p + 0^2q = p \quad (2)$$

$$x_i - p : E(x_i - p) = (1-p)p + (0-p)q = qp - pq = 0$$

$$(x_i - p)^2 : E(x_i - p)^2 = E(x_i^2) - 2p\,E(x_i) + p^2$$

efectuando operaciones y reemplazando los valores de X_i y X_i^2 por los de su igual en (1) y (2)

$$E(x_i - p)^2 = p - 2p^2 + p^2 = p - p^2 = p(1-p) = p \cdot q$$

$$\boxed{E(x_i - p)^2 = pq}$$ Aplicando el primer teorema:

$$\boxed{m_i(Z) = E(Z) = E(x_1) + E(x_2) + E(x_3) + \cdots E(x_n) = np}$$

Z es la frecuencia absoluta

F es la frecuencia relativa

Si queremos encontrar la esperanza matemática de F

Veamos ahora el momento 2o. de Z, o sea $\mu_2(X)$

$$\mu_2(Z) = E\left[Z - E(Z)\right]^2 = E(Z - sp)^2 =$$

$$= E(x_1 + x_2 + x_3 + \cdots + x_s - sp)^2$$

$$= E\left[(x_1 - p) + (x_2 - p) + \cdots + (x_s - p)\right]^2$$

efectuando el desarrollo del cuadrado de un polinomio, nos da la suma de los cuadrados

$$\mu_2(Z) = E\left[\sum (x_i - p)^2 + \sum\sum (x_i - p)(x_j - p)\right] = \qquad i \neq j$$

$$= \sum E(x_i - p)^2 + \sum\sum E(x_i - p)(x_j - p) = s\,pq$$

pues, todos los dobles productos son nulos desde que la esperanza matemática del producto, es el producto de las esperanzas matemáticas y

$$E(x_i - p) = 0$$

tenemos luego el resultado fundamental:

$$\mu_2(Z) = spq = \sigma^2(Z) \qquad \text{o bien} \qquad \sigma(Z) = \sqrt{spq}$$

El desvío medio cuadrático (o dispersión) de la variable aleatoria Z es proporcional a la raíz cuadrada del número de pruebas.

Resumen:

$$Z = x_1 + x_2 + \cdots + x_s \quad \text{(frecuencia absoluta)}$$

$$F = \frac{x_1}{s} + \frac{x_2}{s} + \cdots + \frac{x_s}{s} \quad \text{(frecuencia relativa)}$$

$$m_1(Z) = np = E(Z)$$

$$\sigma^2(Z) = npq \qquad \text{o bien} \qquad \sigma(Z) = \sqrt{npq}$$

$$m_1(F) = E(F) = p$$

$$\sigma^2(F) = \mu_2(F) = \frac{pq}{n} \quad \text{o también} \quad \sigma(F) = \sqrt{\frac{pq}{n}}$$

Capítulo V

MUESTREO

1. INFERENCIA ESTADISTICA

La investigación estadística, que tiene por finalidad principal establecer las características de un universo o población, tropieza en la mayoría de los casos ante la imposibilidad física de poder tomar en cuenta todos los casos. La obtención de medidas cuantitativas de un universo que sean su expresión representativa, se hace entonces imposible. En épocas aun no muy lejanas, se creyó que tales determinaciones solo podrían calcularse en base al conocimiento total de las unidades del universo por censos o encuestas totales. Actualmente sabemos que no es necesario encuestar a todos los elementos de una población para conocer sus características. Basta elegir un pequeño número de individuos para obtener valores representativos de ese universo o población.

Para el estudio de un universo (población de personas, o cosas) contamos con dos procedimientos:

la encuesta total: (c e n s o)

la encuesta parcial: m u e s t r a

El primero consiste en la enumeración completa de cada uno de los elementos de la población considerada, llevado a cabo sólo de tiempo en tiempo. Consiste en relevamientos planteados en base a una buena cartografía, formularios adecuados, numeroso personal adiestrado en la recolección de datos y su inserción en los cuestionarios, oficinas de verificación, codificación y depuración, equipos de mecanización para las tabulaciones, y por último, estadígrafos para su ulterior análisis. Todo esto implica grandes organizaciones, gran cantidad de personal, vastos recursos y un tiempo más o menos largo para conocer los resultados.

Para obviar estos inconvenientes, se han ideado métodos más rápidos de mejor control y de costo más reducido. Son los métodos de muestreo.

Se llama m u e s t r e o a la técnica de seleccionar algunos casos de un universo, de tal manera que se obtenga estadísticas sobre las cuales podamos, por inferencia, construir un panorama de la población derivado de las características observadas en los individuos de la muestra. Es decir, que podemos "estimar" las características de una población o universo con sólo analizar las características de unos pocos individuos pertenecientes a dicha población. A este procedimiento se le llama: Inducción estadística. La gran ventaja que existe, es que, la teoría del muestreo posee técnicas especiales para "medir" el grado de seguridad de dichas "estimaciones".

El método censal y el de muestreo, no son procedimientos excluyentes, por el contrario, la mayoría de las veces se complementan mutuamente.

Los censos detallados, frecuentes y regulares, siguen siendo convenientes y muy útiles. Si bien no son indispensables para tomar una muestra, son en cambio, la base del desarrollo de todas las ventajas potenciales del muestreo. Mediante la teoría del muestreo, el estadístico está en condiciones de "computar" la probabilidad de que un acontecimiento haya ocurrido razonablemente "sólo por azar" o bien que dicho acontecimiento, no pudo razonablemente haber sido causado "sólo por azar", sino debido a cualquier otra causa.

Hay ciertos valores de probabilidad que funcionan como línea divisoria entre los casos que pueden explicarse como debidos al azar únicamente, y aquellos que no pueden explicarse por el azar sino por otras causas.

A menudo se usan valores de probabilidad de 1 en cada 20, o sea 5%, o bien 1 en cada 100 veces.

2. VENTAJAS, LIMITACIONES Y OPORTUNIDAD DEL USO DE MUESTREO

2.1 - Ventajas

Como la inferencia siempre supone un riesgo, es útil resumir en qué casos conviene obtener muestras en lugar de censos.

La técnica del muestreo se usa porque algunas veces,

a) es el único medio posible de investigación, o porque el proceso de medida es destructivo.

b) es el procedimiento más práctico, porque la población es infinita.

c) es el procedimiento más práctico, porque la población es homogénea.

d) es la forma más eficiente.

a) Como único medio posible:

Se usa cuando no se puede realizar un censo completo; por falta de recursos, por falta de personal adiestrado, por tratarse de zonas escasamente pobladas (Groenlandia) o muy densamente pobladas (India) o por ser zonas con accidentes geográficos muy desfavorables. También se utilizará el muestreo cuando se deba realizar un análisis de materiales para examinar su calidad y este examen implique su destrucción. Por ejemplo si se trata de constatar la resistencia de cierto tipo de tejido o materiales de construcción, como: cemento, ladrillos, baldosas, etc.; no se someterá a la prueba toda la partida de material, sino que se elegirá un número determinado, por lo que se está obligado a usar el muestreo, ya que el proceso de medida o investigación de las características de cada elemento es destructivo. En encuestas sociales, se considera procedimiento destructivo a las encuestas agotadoras que disminuyen la eficacia de las respuestas por desagrado de los encuestados.

b) Como procedimiento más práctico: (población infinita)

Por lo general se trata de investigar las características de un universo compuesto por miles o millones de casos. Mediante la muestra será posible obtener un redu-

cido conjunto de ellos y captar sus características para tabular los resultados y analizar los con el propósito de inducir y generalizar las características del total.

Por ejemplo: 1) Deseamos conocer el mercado del aceite para automotores. Como no podremos conversar con cada uno de los conductores que existen en una población, preparemos una encuesta y sacaremos una muestra con suficiente cuidado para que pueda ser representativa de los hábitos de compra del total de los conductores.

2) Un estudio de la población total es a veces imposible. Por ejemplo: el número de niños de raza blanca que han nacido ciegos, será un número desconocido. Prácticamente, se trata de una población infinita de niños. Cualquier estudio de esta población, por ejemplo: por ciento de varones o niñas nacidas ciegas debe hacerse por muestreo.

3) El número de votantes de EE. UU. constituye una población de varios millones. Si un instituto de opinión pública deseara conocer las tendencias políticas del país, debiera reunir a éstos varios millones de votantes y preguntarles su opinión favorable a tal o cuál candidato. Este "voto falso" necesitaría una preparación previa considerable y sería altamente costosa. Necesariamente debe conocerse la opinión pública por medio de una muestra cuidadosamente seleccionada.

c) Cuando la población sea suficientemente homogénea.

Es decir, cuando los elementos sean tan similares entre sí, que cualquier muestra dé una buena representación del universo.

d) Como forma más eficiente

Es, generalmente, el método más eficiente -desde el punto de vista físico y financiero- que se puede llevar a cabo para tener una idea del total, pues: tiene ventajas del: menor costo total, de realizarse en menos tiempo, de reducir las molestias al público y usar menor cantidad de equipo y personal.

Menor costo: Aunque el costo unitario por familia (o unidad inspeccionada) entrevistada, o vivienda visitada, suele ser mayor en una muestra que en un censo, en definitiva, el costo total disminuye considerablemente por la menor cantidad de personal y fichas a emplearse. Se reducen también grandemente los gastos de movilidad y propaganda.

Menor tiempo: entre el comienzo de la investigación y la publicación de los resultados. (El censo de población de 1947, aun no se ha publicado en su totalidad. El tomo correspondiente a población -sólo una parte de los datos referidos a ella- se ha publicado 10 años más tarde: 1957). Si las características estudiadas sufren rápidas mutaciones, el censo tendrá sólo valor histórico.

En el control de un proceso, las conclusiones deben obtenerse rápidamente a fin de salvaguardar la producción en casos de fallas.

Reducción de molestias al público: cuando él tiene que informar, ya que reduce el número de unidades visitadas.

Menor cantidad de equipos: en material y personal, lo que permite que éste último sea más adiestrado en la recolección de una información más depurada y mejor controlada.

En general: el muestreo es conveniente aun cuando no siendo estrictamente imposible el estudio de la población por el número de sus elementos, sí resulte muy difícil la realización práctica de dicho estudio. O bien, que la población sea aceptablemente homogénea, o los elementos investigados en las muestras no lleguen a inutilizarse, pero sí a disminuir su valor o a producir inconvenientes de otro tipo, como el del desagrado de las personas a las que se someten a excesivas encuestas.

2.2.- Limitaciones del Muestreo

1º) Debido a que sólo se observa una parte de la población, el muestreo no puede utilizarse cuando se necesita un inventario de cada uno de los elementos. Cuando la información deba hacerse sobre grupos o áreas muy pequeñas de la población, tampoco es aconsejable el muestreo. Las muestras excesivamente pequeñas provocan incertidumbre en los resultados que pueden afectar gravemente las consecuencias basadas en las estimaciones. Por ejemplo: al publicar estimaciones muestrales relativas al número total de centenarios (o número de establecimientos) en una localidad de pocos habitantes, puede ocurrir que la fácil comprobación de diferencias entre el número estimado y el real, origine una desconfianza en el público aunque los resultados de la muestra sean aceptables.

2º) Existencias de errores: En los censos hay un solo tipo de errores: los sesgos; en las muestras existen dos clases: los sesgos y los propios del muestreo.

3º) Exige un personal estadístico muy especializado. Requiere en comparación con el censo, menor cantidad de trabajo, pero exige mayor refinamiento y preparación.

4º) Requiere gran conocimiento de las teorías matemáticas del muestreo y gran experiencia práctica de la materia que se va a investigar y además, un buen entrenamiento de agentes, inspectores y supervisores.

El muestreo requiere material y planteos de alta eficiencia y complicada estructura.

3. CONDICIONES DE LAS MUESTRAS

La muestra debe poseer dos características para que pueda servir de base a generalizaciones o inferencias sobre la población de la cual proviene.

La muestra debe ser : Representativa
Aleatoria

La representatividad es el problema más importante en el muestreo. Existen universos que son homogéneos, por estar formados por unidades o partículas idénticas y bien mezcladas, en los cuales cualquier muestra que de ellos se extraiga es suficientemente representativa: vino, cemento, jabón, cereales, tejidos, piezas manufacturadas, etc. En general, la muestra debe ser tal que los estimadores que produzca sean capaces de revelar las características del conjunto lo más aproximadamente posible. Si el universo no es homogéneo, es necesario lograr la representatividad por medio de la estratificación o zonificación.

La aleatoriedad implica que en la formación de la muestra, cada persona o elemento del universo, tenga la misma probabilidad o posibilidad de ser elegido. Por ello, se han ideado sistemas en que la elección de las unidades de la muestra se realice sin la más mínima intervención del investigador o encuestador. Por esto se llaman muestras probabilísticas o muestras al azar.

Sin embargo, hay encuestas en las que los investigadores deciden qué unidades deben tomarse mediante su apreciación estrictamente personal y por ello pierden su carácter de azar. A este tipo de muestras se les llama muestras opináticas o causadas.

Las muestras de azar tienen la particularidad de que es posible determinar, mediante fórmulas especiales, los errores propios del muestreo y evitar los sesgos de la elección. En cambio, en las muestras opináticas o dirigidas no se pueden calcular los errores del muestreo ni los sesgos. (Se llaman segos o bias, o inclinación viciada, a la diferencia existente entre el valor esperado de la estimación y el verdadero de la población). Si tal diferencia es cero, se dice que la muestra es insesgada.

Los sesgos son constantes y difíciles de descubrir y calcular. Los censos tienen un tipo de error: la tendencia viciosa cuya corrección es muy difícil.

Las muestras tienen dos tipos de error: los sesgos o bias, y los errores propios del muestreo. Los errores propios del muestreo pueden ser controlados mediante el tamaño y forma de la muestra.

Los segos que originan las tendencias viciosas, son comunes a las muestras y a las enumeraciones censales y son errores difíciles de calcular pero fáciles de evitar con una buena planificación. Son causa principal de estoserrores: Mala definición del universo; Falta de respuestas; Información equivocada; etc.

Los errores del muestreo al azar son tratados por la teoría del cálculo de probabilidades; la tendenciosidad del muestreo es en cambio un error de índole estrictamente personal debido a la incapacidad de cumplir con las instrucciones al hacer la muestra o al generalizar los resultados de la muestra al universo.

4. PROCEDIMIENTOS TECNICOS DE SELECCION

Para obtener muestras aelatorias, se dispone de varios procedimientos técnicos de selección de los elementos:

- SISTEMAS MECANIZADOS O BOLILLEROS
- TABLAS DE NUMEROS ALEATORIOS
- ELECCION SISTEMATICA

4.1 - Sistema Mecanizado o Bolilleros:

Se asigna a cada elemento de la población, un número que al mismo tiempo figura en una bolilla, ficha o ruleta o cualquier otros instrumento de juego. Al azar se van extrayendo las bolillas, fichas o números hasta la cantidad que se desee hacer intervenir en la muestra.

Tiene el inconveniente de fallas mecánicas que produzcan sesgos, es decir, mayor frecuencia de ciertos números o bien que, si se introducen las bolillas o fichas en cilindros, o urnas, y una persona debe sacarlas a ciegas, puede influir en ella la suavidad, la forma, o la impresión, si las mira, los números más grandes o los más chicos, etc.

4.2.- Manejo de la Tabla de Números Aleatorios :

Las tablas de números al azar se construyen con procedimientos no sistemáticos y muy delicados, luego se someten a test para comprobar que no existen tendencias en las filas, columnas o grupos de números.

El procedimiento práctico para la selección de una muestra simple al azar de extensión n de una población que consta de N unidades, es la utilización de una cualquiera de las varias tablas de números aleatorios que se han publicado: Tippet, Babington-Smith, Fisher-Yates, para lo cual:

1o. Se asocia a cada una de las N unidades de la población, uno de los números desde 1 a N. Es decir, se hace un listado o padrón de las unidades que integran la población y se las numera consecutivamente.

2o. Se eligen n números diferentes de la tabla de números al azar, siendo $n<N$.

3o. Se toman como constituyentes de la muestra, aquellas unidades de la población que tienen asignados números iguales a los obtenidos en el segundo paso.

Supongamos tener una población de 350 unidades de muestreo, de la que se desea extraer una muestra simple al azar de 10 unidades.

Entramos en un punto cualquiera de la tabla, por ejemplo, pág. 3, fila 11, columna 13 y tomamos para incluir en la muestra aquellas unidades de la población que tienen asignadas números iguales a los que se leyeron en tabla y menores de 351 a partir de la fila 11 hacia abajo y están formados por las cifras que pertenecen a las columnas 13, 14 y 15. Procediendo así, encontramos los números: 49-12-105-10-152-102-275-347-146-53.

Las unidades de la población:

Ej.: columna 13, 14, 15

fila 11 :	0 4 9	
" 12:	666	} éstos no se pueden tomar por ser mayores que 350
" 13	432	
" 14	463	
" 15	0 1 2	
" 16	1 0 5	

También puede elegirse en diagonal o tomando ciertas columnas o filas de distintas páginas, etc.

Suele ocurrir a veces que la población total está subdividida en grupos, cada uno de los cuales contiene N_1^o, N_2^o, $N_3^o \ldots N_n^o$, unidades de muestreo (por cierto $N_i=N$. Por ejemplo, las viviendas están agrupadas en manzanas en una ciudad, los departamentos o partidos agrupados en provincias, los abonados telefónicos se agrupan en las páginas de la guía telefónica, etc.

Tratándose de obtener una muestra simple al azar puede utilizarse un procedimiento que no requiere numerar todas las unidades.

Para ejemplificar supongamos una población que consta de 371 unidades que se distribuye en 14 grupos, cada uno de los cuales consta del número de unidades anotado en la segunda columna del cuadro siguiente:

Grupo No.	Unidades	Acumulado
1	25	25
2	17	42
3	5	47
4	59	106
5	64	170
6	22	192
7	38	230
8	16	246
9	21	267
10	12	279
11	14	293
12	38	331
13	17	348
14	23	371
	371	

Debiendo elegir una muestra al azar de 10 unidades, se construye la columna 3 acumulando las unidades contenidas en cada grupo y se toman de una tabla de números al azar, 10 números menores que 372; supongamos que se encuentran los siguientes:

29 72 128 96 326 199 202 58 117 33

Las unidades 29 y 33 caen en el grupo 2
" " 58, 72 y 96 caen en el grupo 4
" " 117 y 128 " " " " 5
" " 199 y 202 " " " " 7
" " 326 " " " " 12

4.3. Elección Sistemática

Es un procedimiento mediante el cual sólo el primer elemento se elige al azar y los restantes se seleccionan automáticamente, según un determinado orden de espaciamiento.

Si deseamos el 2% de una población de 10.000 individuos, es decir, 200 de ellos, elegimos una unidad cualquiera al azar, por ejemplo: 35 y luego, tomando las siguientes con intervalos regulares de 50 hasta completar los 200 elementos.

Por ejemplo:

35 - 85 - 135 - 185 - 235 . . . . . .

Este método se usa mucho por su costo reducido y por su simplicidad.

Además tiene la ventaja de poder organizar controles sobre la marcha de la investigación.

Sin embargo, si las unidades se han ordenado por mes, por edad, o por zonas, etc., la selección sistemática en vez de azar, se convierte en muestra estratificada con el inconveniente de que se pueden producir sesgos en el arreglo de las unidades.

Resulta así que solo será necesario numerar parte de las unidades de 5 de los 14 grupos. Del 2º grupo, entran en la muestra las unidades Nº 4 y Nº 8 (29 -25 = 4; 33 - 25 = 8); del 4º grupo, entran en la muestra las unidades 100.

58 - 47 = 11
72 - 47 = 25
96 - 47 = 49

Elección Sistemática

"Cuando los miembros de una población siguen un orden espacial o temporal, (por ejemplo, las personas ordenadas en la guía de teléfonos, según orden alfabético, los

precios dados regularmente cada semana, o las plantas que crecen en los surcos de un campo) suele ser conveniente elegir una muestra escogiendo los individuos a intervalos iguales a lo largo de la ordenación establecida: por ejemplo,

-elegir nombres de individuos a intervalos de 100 en 100.
-elegir las plantas de un surco de 5 en 5.

Sin embargo, si existe periodicidad entre los elementos del universo, este procedimiento no es aconsejable.

Por ejemplo, supongamos que se desea tomar una muestra de los habitantes de una calle. Estos se encuentran ya distribuidos por casas. Seleccionaremos un cierto número de casas cuyos habitantes constituirán la muestra.

Adoptemos como procedimiento la selección de la última de cada diez casas contadas a partir de un origen arbitrario. Pareciera que atributos como la renta o el número de miembros de la familia no están agrupados de manera periódica a lo largo de la calle.

Sin embargo, si la calle estuviese dividida en calles transversales, en manzanas de 10 casas y la última casa de cada manzana fuese la décima, y formase esquina, estando dedicada generalmente a un negocio, resultarían viciadas las condiciones de azar.

En este caso, como el método de selección tiene el mismo período, resulta que el método y las propiedades que se investigan no son independientes.

Cuando se presentan ritmos en la población -por ejemplo en las series cronológicas oscilatorias, o en el terreno que ha sido cultivado con máquinas- este método de selección sistemática, puede dar resultados que no son dignos de confianza.

Sólo se debe usar este método cuando, basándose en motivos apriorísticos, tenemos la seguridad de que el intervalo entre los individuos de la muestra no guarda relación con cualquier propiedad sistemática de la población". (1)

a) Muestreo de fichero con tarjetas inútiles (2)

Si el fichero contiene un cierto número de tarjetas NO ELEGIBLES, por ejemplo, tarjetas en blanco, o tarjetas pertenecientes a individuos que no interesan a los efectos de la investigación muestral, dichas tarjetas NO DEBEN SUSTITUIRSE por tarjetas elegibles.

Por ejemplo, en una investigación referente a empleados de una empresa, manejando el fichero de personal, pueden salir tarjetas correspondientes a empleados que ya no pertenecen a ella.

Supongamos que debemos sacar una muestra de 600 empleados. Pero, si sacamos 600 tarjetas, algunas serán de ex empleados y entonces, no tendremos las 600 tarjetas útiles. Para lograrlo, procedemos así:

1o. Sacar una muestra piloto.

2o. Calcular los porcientos de cada categoría o atributo en la muestra piloto.

3o. Sumar los porcientos que representan los elementos no utilizables.

4o. Calcular el complemento a 100% y obtener así el porciento de fichas útiles.

5o. Multiplicar este porciento por el intervalo de muestreo k,

A continuación veremos un ejemplo:

Se deseaba investigar el control de vacuna antituberculosa B C G en personas menores de 20 años.

Para ello, se necesitaba elegir una muestra en el fichero de la Liga Antituberculosa (de Mendoza) que, a la sazón, contaba con 60.000 fichas logradas por vacunación con B C G en los últimos 25 años.

Se observó que en el fichero habían fichas no elegibles, por lo que se extrajo una muestra piloto que permitió detectar las siguientes características:

Características	Fichas	%
-Personas mayores de 20 años	82	12,2
-No cumplieron tiempo de control de vacuna	127	19,0
-Dudosa o con datos incompletos	50	7,5
-No concurrieron a control	26	3,9
-Concurrieron 2, 3 o más veces a control	385	57,4
	670	100,0

La muestra debía ser de 1.100 elementos. Se procedió así:

a) Determinar la fracción e intervalo de muestreo.

$$\text{Si } N = 60.000 \text{ fichas de vacunados}$$
$$n = 1.100$$

$$f = \frac{n}{N} = \frac{1.100}{60.000} = 0,018333$$

Luego, si de cada 1.000 fichas corresponde sacar f=18, el intervalo k de selección, es:

$$k = \frac{1.000}{18} = 54,55$$

o sea que, de cada 55 tarjetas hay que sacar 1.

II) Eliminación de TARJETAS NO ELEGIBLES

Por corresponder a personas mayores de 20 años, se debe descartar un porciento de:

12, 2%

Los vacunados que aun no cumplieron con el tiempo de control, significan:

19, 0%

Las tarjetas dudosas o incompletas representan:

7, 5%

En consecuencia, el total de tarjetas no elegibles es:

$$12,2 + 19,0 + 7,5 = 38,7\%$$

III) Determinación de TARJETAS ELEGIBLES

El universo de fichas elegibles, que contiene datos útiles para los objetivos de la investigación, quedaba reducido en un 38, 7% y en esa proporción había que reducir también el intervalo de selección de las tarjetas. De tal modo:

$$55 \times (100 - 38,7) = 33,7$$

Por lo tanto, de cada 34 tarjetas deberá sacarse 1.

Se obtendrán así, 1765 fichas de las cuales solo serán útiles 1.100.

La selección debe hacerse entonces, eligiendo un "arranque" al azar entre los dígitos 0 a 9. Si por ejemplo, sale elegido el 6, la selección de las tarjetas será:

No. 6	142
No. 40	176
74	etc.
108	

Si no se hace la depuración previa del fichero, por el procedimiento indicado, la muestra estará sujeta a error de muestreo y no simplemente a error en la selección.

b) Tratamiento de TARJETAS DUPLICADAS

Al trabajar con ficheros se debe tener la seguridad de que cada elementos a muestrear esté representado en 1 y sólo 1 tarjeta. Si algunas personas tuvieran 5, otras 3 y otras 1, entonces la representatividad de la muestra sería de 5 a 3 a 1.

Esto implicaría que la muestra tendría una sobrerepresentación de las dos primeras clases comparadas con la última y los resultados estarían viciados seriamente.

Para obviar esto, deben excluirse todos los duplicados o usar alguna técnica especial para cada persona.

c) Tratamiento de TARJETAS NUMERADAS

Si las tarjetas están numeradas correlativamente en el fichero, estos números pueden ser utilizados para facilitar la selección de la muestra, eligiendo determinados dígitos y sus secuencias.

De este modo, se puede estar relativamente seguro de que no hay tendencia sistemática respecto a algunas tarjetas que puedan tener mayor representatividad.

Si las tarjetas no estuvieran numeradas correlativamente, pero tuvieran otro número, por ejemplo el de afiliación a las Cajas Jubilatorias y se deseara una muestra del 5%, se elegirán a las fichas cuyo número de afiliación termine por ejemplo, en 14, 34, 54, 74 y 94, o en otras combinaciones de dígitos que representen al 5% de los dígitos entre 00 y 99.

d) Tratamiento de FICHEROS MUY NUMEROSOS

Los métodos anteriores se basan en la idea de que la extracción de fichas de un fichero no resulte gravosa. Pero, si el universo es muy grande y no hay numeración en las tarjetas, esta tarea puede resultar muy costosa o muy lenta.

En tales casos, puede recurrirse a otro procedimiento, que es el de conglomerados.

Mediante este procedimiento, puede considerarse que cada cajón o bandeja del fichero es un conglomerado; elegir una muestra de ellos y seleccionar una submuestra de tarjetas en los conglomerados de la muestra.

Sin embargo, debe recordarse que el muestreo por conglomerados es aceptable sólo si no existe -o es muy pequeña- correlación entre la característica a estimar en las sucesivas fichas.

e) Tratamiento de ficheros con información parcial

A veces, se dispone sólo de listas o padrones de personas, establecimientos, familias, etc. pero, no tienen otra información aparte de nombre y domicilio.

Esto es sólo un medio para elegir la muestra. La información deberá requerirse por encuesta de campo, por correo, o por teléfono.

También puede utilizarse como método de selección, el expuesto anteriormente, pero pueden también utilizarse otros procedimientos alternativos a veces más eficientes (por ejemplo, el muestreo por conglomerados, que introduce grandes economías, particularmente se selecciona una muestra pequeña en un área grande.

Si los registros están por orden geográfico, puede extraerse una muestra

por conglomerados geográficos muy fácilmente. Por ejemplo, supongamos que se desea una muestra de manera que los conglomerados elegidos tengan,

$$\tilde{n} = 5 \text{ unidades}$$

Se procederá así:
Observar las fichas en grupos de ñ,
tomar 1 en k de tales grupos.

Supongamos que $\tilde{n} = 5$, y también:

$$f = 1/20$$

a) Si las fichas están numeradas correlativamente, se agrupan juntas:

1 - 5
6 - 10
11 - 15
12 - 20 etc.

se elige 1 en 20 de estos grupos.

b) Si las fichas son muchas y están ordenadas geográficamente por cajones, se muestrean los cajones y luego se saca una submuestra de tarjetas. Este procedimiento será ventajoso cuando haya muchos más cajones que el tamaño de la muestra.

5. ESTIMACION DE CARACTERISTICAS

Cuáles son los objetivos del muestreo ?
Qué se desea investigar mediante las muestras?

Se trata:

- de estimar los parámetros de la población ?
- de probar una hipótesis ?
- de probar el grado de significación de un acontecimiento?
- de determinar niveles de confianza?

Estos son en general, los problemas típicos que resuelve el muestreo con distintas técnicas.

El problema que más comúnmente se debe afrontar en los estudios socio-económicos, es:

"Estimar los parámetros de la población partiendo de los datos de la muestra".

Estas estimaciones pueden ser:

media (promedios)
porcentajes
medianas
totales de la población
dispersión

Por ejemplo:

1. - Determinar el promedio de años de escolaridad de un determinado grupo social (contratistas, obreros, etc.).
2. - Determinar el % de personas a favor de tal o cuál partido político.
3. - Estimar el total de viviendas deshabitadas en una zona determinada.
4. - Calcular la variación entre los gastos para asistencia médica de las familias de cierta zona o cierto grupo (y ferroviaria).

NOTACION

Usaremos la siguiente notación

Característica	Para la Muestra (estimador)	Para la Población (parámetro)
1. Media	$\bar{X}$	μ
2. Porciento	p	P
3. Número de elementos	n	N
4. Total	X	X
5. Variancia	S^2	V^2
6. Dispersión	S	V
7. Error standard de la media	$S\bar{x}$	$V_{\bar{x}}$
8. Error standard del porciento	Sp	V_p

5.1 - Distribución de la Media

Cuando de un universo se extraen todas las muestras posibles, y a cada una de ellas se le calcula la media, la media de las medias muestrales coincide con la media de la población.

Supongamos un sencillo ejemplo:

De un universo hipotético de sólo 6 personas, debemos calcular la edad media.

Efectuaremos el cálculo por dos vías distintas para comprobar al fin que en ambos casos los resultados coinciden.

1º - Considerando a todos los elementos de la población

A	18 años
B	22 "
C	19 "
D	21 "
E	15 "
F	25 "
Total	120 años

Luego, la edad media de la población es: $\mu = \frac{120}{6} = 20$ años

2º - Considerando todas las muestras posibles

Si aplicamos muestreo, elegiremos una de todas las muestras posibles -en este caso: 15, que puedan formarse con los 6 elementos de esta población.

El número de muestras posibles para: N = 6
n = 2

es: $T = C_{N,n} = \frac{N!}{n!\,(N-n)!} = \frac{6!}{2!\;4!} = \frac{6.5.4.3.2.1.}{2.1 \times 4.3.2.1} = 15$

Las 15 muestras posibles y sus correspondientes valores son:

Muestra Nº	Elementos	Edad Media Muestral $\bar{X}_i$ años	Desvío Respecto a la Media Verdadera μ = 20 años
1	A y B	20	0
2	A y C	18,5	-1,5
3	A y D	19,5	-0,5
4	A y E	16,5	-3,5
5	A y F	21,5	1,5
6	B y C	20,5	0,5
7	B y D	21,5	1,5
8	B y E	18,5	-1,5
9	B y F	23,5	3,5
10	C y D	20,0	0
11	C y E	17,0	-3,0
12	C y F	22,0	2,0
13	D y E	18,0	-2,0
14	D y F	23,0	3,0
15	E y F	20,0	0,0
	Total	300,0	0,0

Luego, la media de las medias muestrales es:

$$X = \frac{\sum \bar{X}}{T} = \frac{300}{15} = 20 \text{ años}$$

O sea, que:

$$\bar{X} = \mu$$

Obsérvese que:

1º - En las muestras 1, 10 y 15, el estimador de la muestra (estimación de la edad media) coincide con el parámetro. En ellas no hay error de muestreo.

2º - La mayoría de las medias muestrales resultaron muy próximas a la media parámetro. Ello se debe a que la población es bastante homogénea respecto al atributo que se estudia (edad).

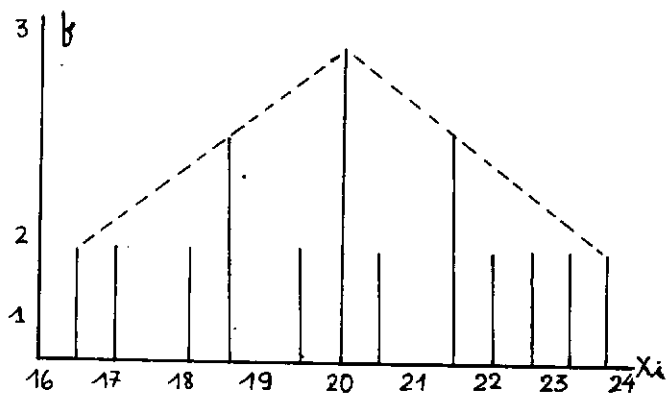
3º - Los errores o desvíos se han presentado en sentido positivo y negativo de tal modo que, en definitiva, se compensan y así resulta: $\bar{X} = \mu$

4º - Si representamos la distribución de los desvíos ($\bar{X}_i - \mu$) o errores de muestreo, se logra una distribución de tipo campanular. Esto ocurrirá no sólo para este ejemplo.
Generalizando, podemos decir que los errores grandes (desvíos o medias muestrales muy alejados del valor verdadero) se presentan con poca frecuencia.
En cambio, son muy frecuentes los valores próximos al valor esperado (media parámetro).

Esto también se observa si se representa la distribución de las medias muestrales, también llamada distribución de probabilidades de $\bar{X}$.

Distribución de las medias muestrales o distribución de probabilidades de $\bar{X}$

Media	Muestral $\bar{X}_i$	Frecuencia A f	Distribución muestral
	16,5	1	
	17,0	1	
	17,5	0	
	18,0	1	
	18,5	2	
	19,0	0	
	19,5	1	
	20,0	3	
	20,5	1	
	21,0	0	
	21,5	2	
	22,0	1	
	22,5	1	
	23,0	1	
	23,5	1	



La distrubución corresponde siempre a una población dada y a un tamaño dado de muestra. Cualquiera de ambos que cambie, hará variar la distribución.

En general, se comprueba que la Media de las Medias Muestrales es igual a la Media Parámetro.

Lo que acabamos de comprobar con un ejemplo, lo demostraremos ahora analíticamente con la siguiente demostración:

Si una población está compuesta por N elementos

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

de la cual extraemos muestras al azar de tamaño n,

$$X_1 \ , \ X_2 , \ X_3 , \ldots\ldots\ldots X_n$$

el número de muestras distintas que se pueden obtener, es:

$$C_{N,n} = \frac{N!}{n! \ (N - n)!} = T$$

Se considera que 2 muestras son distintas cuando por lo menos tienen 1 elemento diferente.

La media de la población o MEDIA PARAMETRO se calcula:

$$\mu = \frac{\sum_{}^{N} x_i}{N} = \frac{1}{N}\left(x_1 + x_2 + x_3 + \cdots + x_n\right)$$

La media de cada muestra o estimador, es:

$$\bar{X} = \frac{\sum_{}^{n} x_i}{n} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \cdots + x_n\right)$$

Si consideramos un número grande de muestras tomadas en la misma forma todas, obtendremos una serie de medias muestrales.

$$\bar{X}_1 \quad \bar{X}_2 \quad \bar{X}_3 \ \ldots\ldots\ldots\ldots\ldots\ldots \bar{X}_T$$

Como son variables aleatorias, les son aplicables las propiedades de la Esperanza Matemática. Es decir, "La esperanza de las medias es el promedio de las medias".

$$E(\bar{X}) = \frac{\sum_{}^{T} \bar{X}_i}{T} = m(\bar{X})$$

Por ser la esperanza matemática de la suma, igual a la suma de las esperanzas; y la esperanza de una constante por una variable aleatoria, igual a dicha constante por la esperanza de la variable, resulta:

$$m(\bar{X}) = E(\bar{X}) = E\left[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)\right] \frac{1}{n} \sum^{n} E(x_i)$$

Pero, cada una de las variables x_i tiene por esperanza matemática $E(x_i)$, la suma de los valores que puede tomar, es decir, todos los de la población (por la condición de equiprobabilidad) por sus respectivas probabilidades. La probabilidad de cada valor es:

Así, para cada una de las x_i

$$E(x_i) = \sum_{i=1}^{N} x_i \frac{1}{N} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

Reemplazando (2) en (1) tenemos:

$$m(\bar{X}) = E(\bar{X}) = \frac{1}{n} \sum^{n} \mu = \frac{1}{n} \cdot n\mu = \mu$$

Por lo tanto,

$$m(\bar{X}) = \mu$$

Es decir que: "La media (o esperanza) de las medias muestrales es igual a la media parámetro o media de la población.

La seguridad de una media muestral como estimación de la media de la población, depende de la variabilidad de la distribución de las medias muestrales.

En el ejemplo ya tratado, las medias oscilan entre 16,5 y 23,5. Si las $\bar{X}_i$ son todas muy semejantes, cualquiera de ellas puede tomarse como una buena estimación de la media parámetro.

En cambio, si las $\bar{X}_i$ difieren mucho entre sí, una cualquiera de ellas tendría un valor muy escaso para representar a la media de la población.

En resumen: la confianza de la media basada sobre valores de muestras al azar, depende de la variabilidad de su distribución muestral.

Si la dispersión de las $\bar{X}_i$ es pequeña, cualquier $\bar{X}_i$ es eficaz.

Inversamente, si la dispersión es grande, las medias no son una buena estimación de la media parámetro.

5.2 - Error Standard de la Media

La desviación standard o dispersión de las medidas de una distribución muestral es útil porque nos permite:

a) Estimar la precisión efectivamente lograda en la investigación.

b) Estimar el tamaño de la muestra requerido para alcanzar una determinada precisión.

c) Comparar la precisión de la estimación obtenida mediante el muestreo simple al azar con la lograda por otros métodos de muestreo.

La desviación standard de las medias muestrales se llama ERROR STANDARD DE LA MEDIA, y se indica con la notación $\sigma_{\bar{x}}$, donde el subíndice $\bar{x}$ sirve para recordar que se trata de la distribución de los valores de las medias muestrales en lugar de los x_i de la población.

Si el error standard de la distribución muestral es grande, las medias no son representativas (ello se debe a que difieren grandemente unas de otras).

Si el error standard es pequeño, se dice que las medias son representativas (ello se debe a que son muy parecidas, por lo cual tienen poca diferencia entre sí).

Veamos a continuación los fundamentos del cálculo del error standard:

Desviación Standard de las Medias Muestrales.

La desviación standard de las medias de todas las muestras posibles de una misma población, es el error standard de la media.

Por definición de variancia, siendo $\bar{X}_i$ la variable aleatoria:

$$\sigma_{\bar{x}}^2 = E(\bar{X}_i - \mu)^2 = \quad \text{sustituyendo } \bar{X}_i \text{ por su igual}$$

$$= E\left(\frac{\sum x_i}{n} - \mu\right)^2 = \quad \text{desarrollando la suma multiplicando y dividiendo por } n \text{ al valor de}$$

$$= E\left[\frac{x_1}{n} + \frac{x_2}{n} + \frac{x_3}{n} \cdots - n\frac{\mu}{n}\right]^2$$

aplicando propiedad asocitiva.

$$= E\left[\left(\frac{x_1}{n} - \frac{\mu}{n}\right) + \left(\frac{x_2}{n} - \frac{\mu}{n}\right) + \cdots + \left(\frac{x_n}{n} - \frac{\mu}{n}\right)\right] =$$

Resolviendo el cuadrado del polinomio:

$$= E\left[\sum_x \left(\frac{x_i}{n} - \frac{\mu}{n}\right)^2 + \sum_i \sum_j \left(\frac{x_i}{n} - \frac{\mu}{n}\right)\left(\frac{x_j}{n} - \frac{\mu}{n}\right)\right]$$

Pero, como la "esperanza de una suma es igual a la suma de las esperanzas":

$$\sigma_{\bar{x}}^2 = \sum^{n} E\left(\frac{x_i}{n} - \frac{\mu}{n}\right)^2 + \sum_i \sum_j E\left(\frac{x_i}{n} - \frac{\mu}{n}\right)\left(\frac{x_j}{n} - \frac{\mu}{n}\right)$$

Por ser variables aleatorias independientes, sabemos que E (x-μ1) = 0, lo cual anula al 2o. sumando. Además, la suma de n veces, el 1o. sumando es lo mismo que multiplicar por el factor n.

Por lo tanto:

$$\sigma_{\bar{x}}^2 = n\, E\left(\frac{x_i}{n} - \frac{\mu}{n}\right)^2$$

Sacando al denominador de las fracciones como factor común,

$$\sigma_{\bar{x}}^2 = n\, E\left[\frac{1}{n^2}\left(x_i - \mu\right)^2\right]$$

Pero, la esperanza de una variable por una constante, es igual a la constante multiplicada por la esperanza de la variable,

$$\sigma_{\bar{x}}^2 = n\,\frac{1}{n^2}\, E\left(x_i - \mu\right)^2$$

Reemplazando por su igual:

$$\sigma_{\bar{x}}^2 = n\,\frac{1}{n^2}\,\sigma_{x_2}^2 = \frac{\sigma_{x_i}^2}{n}$$

Extrayendo raíz cuadrada

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma_{x_i}}{\sqrt{n}}} \qquad (3)$$

O sea que, la dispersión de las medias $\sigma_{\bar{x}}$ o error standard es igual a la dispersión de los elementos de la población (σ_x parámetro) dividido por la raíz cuadrada del número de elementos de la muestra.

Es decir : el error que se comete en la estimación, varía en proporción inversa al número de elementos de la muestra.

El valor $\sigma_{\bar{x}} < \sigma_{x_i}$ pues, en σ_{x_i} intervienen todos los valores de la población (hay mayor variabilidad) y en σ_x intervienen sólo los promedios.

La fórmula (3) vale sólo para poblaciones infinitas o bastante grandes.

En cambio, si la población es relativamente pequeña, o el muestreo se hace sin reposición de los elementos elegidos para la muestra, se aplica un factor de corrección para poblaciones finitas. (c.p.f.) y así, el error standard en tales casos se calcula mediante:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

En general, ocurre que no se conoce el parámetro σ, por lo cual hay que estimarlo partiendo del valor de la dispersión de la muestra s. Hay que reemplazar entonces por la relación:

$$\sigma = s \sqrt{\frac{n}{n-1}}$$

La corrección $\sqrt{\frac{n}{n-1}}$ se hace innecesaria cuando $n > 100$, en cuyo caso $\sigma = s$ así:

n	$\sqrt{\frac{n}{n-1}}$
2	1,414
3	1,225
. . .	....
10	1,054
. . .	.....
20	1,026
. . .	.....
30	1,017
. . .	.....
60	1,008
. . .	.....
100	1,005

El valor s, o dispersión muestral, se calcula partiendo de la cuasi variancia, fórmula muy semejante a la de la variancia excepto que se divide por n-1 en lugar de n.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Esta estimación de s, reemplaza en las fórmulas (3) y (4) a la desviación standard de la población σ_x. La fórmula de trabajo para poblaciones grandes, queda entonces:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Esta fórmula nos indica que: mientras mayor sea el tamaño de la muestra, menor será el error standard de la media, o sea que la media muestral será una estimación más confiable.

Por ejemplo, si cuadruplicamos el tamaño de la muestra, reduciremos a la mitad el error standard, con lo cual duplicamos la seguridad de la media.

Para $\sqrt{}$ = 5
n = 25

$$\sqrt{\bar{X}} = \frac{5}{\sqrt{25}} = 1$$

Para n = 100

$$\sqrt{\bar{X}} = \frac{5}{\sqrt{100}} = 0{,}5$$

Estimación del Total

Llamando X' al total estimado en base a relaciones muestrales, será

$$X' = N \cdot \bar{X}$$

donde: X' = total estimado
N = total de elementos
$\bar{X}$ = media muestral

5.3. Estimación de la Proporción P y su Error Muestral

Una proporción es simplemente un caso especial de una media aritmética, donde la variable -de tipo cualitativo- toma sólo los valores 1 o 0.

Se trata de acontecimientos con dos eventualidades excluyentes A y B, con sus respectivas probabilidades P y Q de encontrarse en la población. Veamos algunos ejemplos para una población N = 1.000.

Característica	Elementos A		Elementos B	
	No.	P %	No.	Q %
argentinos	800	80 %	.	.
extranjeros	.	.	200	20 %
varones	510	51 %	.	.
mujeres	.	.	490	49 %
con opinión favorable a la ley X	250	25%	.	.
con opinión desfavorable	.	.	750	75 %

En general, cualquier población o conjunto de características cuantitativas (variable) constituye también una población de características cualitativas (atributo) clasificando sus elementos en: inferiores y en: iguales o mayores que un cierto valor dado.

A este procedimiento se le denomina "dicotomizar la variable" y es de gran aplicación en estudios sociales.

A veces, se agrupan los elementos de la población en más de dos clases, por ejemplo; varios partidos políticos, distinto estado civil, grupos diferentes de opiniones, etc. (tricotomización).

Volviendo a nuestro ejemplo hipotético de una población de 6 individuos, podría interesarnos determinar la proporción de hombres que hay dentro del grupo. Asignamos el número 1 a los hombres y 0 a las mujeres. Tendremos·

		X_i
Sr.	A	1
Srta.	B	0
Srta.	C	0
Sr.	D	1
Sr.	E	1
Srta.	F	0
	Total hombres ...	3

$$P = \frac{\sum X}{N} \qquad P = \frac{3}{6} = 0,50 \qquad \text{o sea: } 50\%$$

Esto es la media de los valores o proporción P.

La desviación standard de estos valores, puede encontrarse en la forma usual o más fácilmente por la forma siguiente:

$$\sigma = \sqrt{P\,(1 - P!)} \qquad \sigma = \sqrt{0,50 \times 0,50} = 0,50$$

Supongamos que se desea calcular la proporción de hombres del país X, y que para ello tomamos una muestra de 400 individuos seleccionados al azar en el total de la población.

Si en la muestra aparecen 204 hombres y 196 mujeres, la proporción de varones resulta:

$$\frac{204}{400} = 0,51 \qquad \text{o sea, el } 51\%$$

Pero, este resultado no nos capacita para inferir que los hombres constituyen el 51% de la población, sino sólo para decir que el verdadero porciento se encuentra "cerca" del 51%. Para poder generalizar los resultados de esta muestra, debemos conocerla

Distribución Muestral de las Proporciones

La distribución muestral de las proporciones es simplemente una distribución de las proporciones obtenidas de un gran número de muestras al azar (idealmente, todas las muestras posibles) siendo todas del mismo tamaño y seleccionadas de la misma población.

La forma de una distribución muestral de una proporción P, para un gran número de muestras al azar, se aproxima bastante a una distribución normal.

Esto tiende a ser cierto, excepto cuando la verdadera proporción en la población difiere considerablemente del 0, 50. El desvío, respecto a este valor, debe asegurarse previamente.

A fin de describir la seguridad de una proporción muestral, necesitamos una medida de la variabilidad de las proporciones obtenidas.

Como en el caso de la distribución muestral de las medias, la medida que se usa para describir la variabilidad de una proporción, es la dispersión o desviación standard. Se llama ERROR STANDARD DE UNA PROPORCION y se lo designa por el símbolo $\sqrt{p}$

El error standard de una proporción, es la desviación standard de la distribución de las proporciones para un gran número de muestras al azar del mismo tamaño y de la misma población.

Error Standard de una Proporción

Sustituyendo la dispersión de una distribución de proporciones,

$$\sqrt{p} = \sqrt{P\,(1-P)}$$

por $\sqrt{\ }$ en la fórmula correspondiente a las medias $\sqrt{\bar{X}} = \frac{\sqrt{\ }}{\sqrt{n}}$, obtenemos la expresión del error standard de la proporción, que está dada por la fórmula:

$$\sqrt{p} = \sqrt{P\,(1-P)} \qquad (5)$$

Por lo tanto, la verdadera variabilidad $\sqrt{p}$ de la proporción P para un número grande de muestras al azar de igual tamaño y relativamente grandes, depende:

1o. del número de elementos n de cada muestra.

2o. de la verdadera proporción P que existe en la población total (a menos que la verdadera proporción esté próxima a cero o uno).

La fórmula (5) confirma lo que intuimos por el sentido común: "para cualquier valor de P, a medida que se aumente el tamaño de la muestra n, se aumentará la confianza de la estimación, ya que al crecer n, se obtendrán valores más pequeños del error $\sqrt{p}$

Ejemplo: Para P = 0, 50, el error standard con distintos tamaños de muestra es:

n = 625 $\sqrt{p} = \sqrt{\frac{0,50 \times 0,50}{625}} = \frac{0,50}{\sqrt{625}} = 0,02 \ldots\ldots 2\%$

n = 2.500 $\sqrt{p} = \sqrt{\frac{0,50 \times 0,50}{2,500}} = \frac{0,50}{\sqrt{2,500}} = 0,01 \ldots\ldots 1\%$

n = 10.000 $\sqrt{p} = \sqrt{\frac{0,50 \times 0,50}{10.000}} = \frac{0,50}{\sqrt{10.000}} = 0,005 \ldots\ldots$ 1/2 de 1%

Otro aspecto de gran interés en esta fórmula, es el siguiente: "La magnitud del error tiende a crecer a medida que P se acerca a 0, 50".

Por ejemplo, si mantenemos constante el tamaño de la muestra en n = 2.500 tendremos:

Si P = 0,50 $\sqrt{p} = \sqrt{\frac{0,50 \times 0,50}{2.500}} = 0,01 \ldots\ldots\ldots\ldots\ldots\ldots 1\%$

Si P = 0,20 $\sqrt{p} = \sqrt{\frac{0.20 \times 0.80}{2,500}} = 0,008 \ldots\ldots\ldots\ldots\ldots\ldots 0,8\%$

Esta propiedad tiene un extraordinario interés cuando se intenta hacer predicciones sobre los resultados de una elección.

Con un tamaño dado de muestra, la seguridad del pronóstico disminuye grandemente a medida que P tiende a 0, 50. Es el caso de políticas más o menos equivalentes.

6. TAMAÑO DE LA MUESTRA

Una vez definido el objetivo de la investigación, la primera incógnita que se plantea es:

> Qué tamaño debe tener la muestra para que los resultados no sean imprecisos y el costo no sea excesivo.

Para determinarlo, se deben establecer previamente:

1o. La característica o características a estimar (promedio, porcentaje total, etc.).

2o. El grado de confianza con que se desea la estimación.

3o. El error que será admisible (la precisión requerida).

4o. La variabilidad de la población respecto a la característica a estudiar.

Con respecto al punto 4o. conviene observar que:

a) mientras más dispersos -valor grande y variabilidad- están los valores de la variable, más arriesgado será utilizar una muestra de tamaño pequeño.

b) el diseño del tamaño óptimo sólo podría conseguirse a partir del conocimiento de la población para determinar el valor correcto de la varianza. (Pero esto es lo que se denomina la "paradoja de Friedman", y que no es más que un aspecto de la posible representación espiral de las fases del conocimiento científico).

Pero, el problema del muestreo se presenta casi siempre en términos de "conocer las características de la población a través de la muestra, precisamente porque no se conoce o no es posible o conveniente conocer a toda la población".

En consecuencia, el valor de la dispersión se deberá estimar en base al conocimiento que se tenga de la población:

a) por alguna investigación anterior o bien,
b) por una muestra piloto previamente seleccionada
c) por el conocimiento de poblaciones parecidas cuyas características de variabilidad se conocen.

Planteo del Problema:

6.1. Estimación de la Media en una Distribución con Variable cuantitativa.

Se trata de estimar la media de la población M mediante la media muestral X, con un error máximo admisible y un coeficiente de confianza k (1).

Para determinar el tamaño de la muestra, partimos de la ecuación fundamental siguiente:

Este valor k, tomará valores distintos, según que la distribución de la variable sea de tipo Normal o Gausiana; o del tipo de distribución t de Student, o que convenga partir de la desigualdad de Chebyshev.

$$\text{Error absoluto} = \text{desvío } k \cdot \text{error de muestreo} \qquad (1)$$

El error absoluto se expresa en las mismas unidades de la variable en estudio.

Se denomina también "semi intervalo confidencial" y se determina como el producto

del error de muestreo o error standard de la estimación por un desvío k que corresponda al coeficiente de confianza requerido. (2)

Partimos de la fórmula del error de muestreo de la media.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}}$$

Reemplazando en la relación (1), tenemos:

$$e = k\,\sigma_{\bar{x}}$$

$$e = k \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \quad \text{donde, eliminando raíces}$$

$$e^2 = k^2 \cdot \frac{\sigma^2}{n} \cdot \frac{N-n}{N}$$

$$e^2 = k^2 \frac{\sigma^2}{N} \left(\frac{N}{n} - 1 \right)$$

$$\frac{e^2 N}{k^2 \sigma^2} = \frac{N}{n} - 1$$

$$\frac{e^2 N}{k^2 \sigma^2} + 1 = \frac{N}{n}$$

$$\frac{e^2 N + k^2 \sigma^2}{k^2 \sigma^2} = \frac{N}{n}$$

$$\boxed{n = \frac{N k^2 \sigma^2}{e^2 N + k^2 \sigma^2}}$$

En distribuciones normales, para un grado de confianza del 95%, k = 1,96
Generalmente, se toma: k 2. Si se desea una seguridad del 99%, k = 3

6.2. Estimar la Proporción de un Atributo (variable cualitativa) dentro de la población total.

El número de elementos a incluir en la muestra, para el caso en que se desee investigar porcientos de un determinado atributo, se puede conocer consultando la tabla especial preparada por la Universidad de Harvard.

Los valores de dicha tabla se han calculado para distintas combinaciones posibles de p y de q, usando la fórmula:

$$n = \frac{4pq}{E^2}$$

que se deduce de la fórmula del error standard de muestreo, en la siguiente forma:

Si admitimos una precisión en los resultados (margen de confianza) del 95%, se toma como base el intervalo de confianza k = 2, o sea, trabajamos con $2\sqrt{\ }$. Por ello, siendo:

$$\sqrt{\ } = \sqrt{\frac{pq}{n}}$$

$$2\sqrt{\ } = 2\sqrt{\frac{pq}{n}}$$

$$\left(2\sqrt{\ }\right)^2 = \frac{4 \cdot p \cdot q}{n}$$

$$n = \frac{4 \cdot p \cdot q}{\left(2\sqrt{\ }\right)^2}$$

$$\boxed{n = \frac{4 \cdot p \cdot q}{E^2}}$$

Ejemplo:

Límites de error E $2\sqrt{\ }$ en %	Valores presumibles de p y q en % (p + q = 100)				
	1/99	5/95	10/90	20/80	30/70 .... 50/50
0,1	39.600	190.000	360.000	640.000	840.000 ... 1.000.000
0,5	1.584	7.600	14.400	25.600	33.600 40.000
1,0	396	1.900	3.600	6.400	8.400... 10.000
10,0	4	19	36	64	84 100

El presente trabajo se terminó de imprimir
en los talleres del C. F. I.
en la Primera Quincena de Febrero de 1966.