UNIVERSIDAD AUTÓNOMA DE YUCATÁN

Facultad de Ciencias Antropológicas

CURSO DE ESTADÍSTICA

♏ 8	☥ 9	⌛ 1	Φ 3	🖱 5	⇑ 7	🕐 2	± 3	❄ 7	Φ 8	Ξ 2
⊖ 7	⭘ 4	⊗ 6	○ 5	Φ 2	✡ 8	⌛ 0	❒ 1	✠ 5	👌 9	⭘ 3
✡ 6	& 7	🖱 8	Ξ 9	❄ 0	Ξ 9	☝ 3	± 4	🖱 1	○ 5	⊗ 4

Carlos Augusto Evia Cervantes

Enero de 2007

ÍNDICE

INTRODUCCIÓN

El continuo aumento del uso de las técnicas estadísticas por parte de los investigadores en ciencias sociales ha producido la necesidad de que éstos se interesen tanto en el sentido concreto de esta rama de las matemáticas, como las posibilidades de su aplicación en el campo de lo social.

La intención de este libro es acercar un poco más a los antropólogos y otros estudiosos de disciplinas afines a los mecanismos matemáticos desarrollados en la historia de la Estadística. Debido a su complejidad, en muchos casos, estos mecanismos no son comprendidos en su origen ni en su aplicación; por este motivo algunos investigadores que deberían usar estas herramientas las menosprecian creando el falso antagonismo entre los análisis cuantitativos y los cualitativos, considerándose ellos mismos partidarios del segundo tipo. Una autora opina que en la práctica científica los aspectos cuantitativos y cualitativos son complementarios y forman la esencia de cualquier fenómeno a estudiar...

> "... desde una **perspectiva amplia, la opción cualitativa no se opone a la cuantitativa**. No hay demasiados argumentos como para concebir que cantidad y calidad constituyen categorías opuestas. La **investigación científica**, sea sobre los fenómenos naturales o sociales, **siempre trabaja con ambas**. Si el interés es medir algo, ese algo siempre es una cualidad, es decir, una característica o circunstancia que distingue a las personas o las situaciones que se estudian. Si por el contrario, queremos conocer la representatividad de ciertos hechos en una población, es muy probable que debemos ir más allá de la descripción o interpretación de las narraciones o comportamientos, por muy elaborados que ellas sean. Y esto es así porque se querrá conocer **si este hecho es único o se repite**, si es menor o mayor que otro, si presenta un ritmo en el tiempo, etcétera. Es decir, en algún momento surgirá la necesidad de cuantificarlo, de medirlo. Planteado en este nivel, el problema sería inexistente o falso, pues la **realidad social** se presenta como un **desorden complejo**. No es cualitativa ni cuantitativa. El **investigador** selecciona un objeto de estudio y **elige** como estudiarlo" (TARRÉS; 2001: 7).

Tampoco se pretende convencer al lector de la posición contraria que nos lleve al otro

extremo; esto es, la idea que todos los estudios antropológicos requieren del uso de la Estadística. Existen muchos temas que no exigen algún procedimiento de los que se estudiarán en este curso. Por lo tanto, la utilización forzada o al menos improvisada de los instrumentos matemáticos no proporciona validez ni rectifican un mal proyecto de trabajo. El buen uso de la Estadística se advierte en el cumplimiento cabal de su cometido: será oportuno y razonable aplicar estas técnicas será cuando lo exija la naturaleza numérica de los datos que constituyen la variable, o bien, cuando la frecuencia de aparición de los fenómenos estudiados haga necesario el uso de los instrumentos estadísticos.

Como una **definición general** se puede considerar que la Estadística es la rama de las Matemáticas que organiza, describe, analiza y, eventualmente, predice el comportamiento de las variables numéricas y no numéricas a partir de la observación de su frecuencia y de sus cambios. El **objetivo** de quien estudia y aplica la Estadística ha de ser el conocer las características de las variables en cuestión, comprender la naturaleza de sus posibles cambios para explicarlos y ofrecer, cuando se pueda, alternativas de mejoramiento o control de los fenómenos estudiados.

Aquel investigador que pretenda aplicar las fórmulas estadísticas a su información deberá considerarlo desde el inicio de su proyecto, pues desde las estrategias de captura de las variables deberán utilizarse los instrumentos adecuados para su posterior manejo. No está demás señalar que un proyecto mal concebido no se compondrá con el uso improvisado de las técnicas estadísticas.

En la actualidad ha aumentado la demanda de la Estadística en los estudios de cerámica y de restos óseos encontrados en los sitios arqueológicos. Las técnicas cuantitativas les han facilitado a los historiadores la organización de sus datos en series cronológicas de corto y largo plazo. La Estadística se ha convertido es una herramienta fundamental en los estudios de opinión y en los de preferencia del voto en tiempos electorales. Su utilización está altamente difundida en los estudios relacionados con migración, fuerza laboral y conurbación. La Estadística también propicia el análisis de la información recabada y presentación gráfica de los resultados en cualquier disciplina.

Aprovecho este espacio preliminar para expresar una reflexión en torno la difícil

relación que suele presentarse entre la mayoría de los profesionales en ciencias sociales y las ciencias matemáticas.

Es muy común encontrar, especialmente entre los estudiantes de las ciencias sociales, cierto temor a la matemática, debido e experiencias anteriores en otros cursos. Este temor se vuelve, a veces, una franca aversión hacia todo lo que sea hacer operaciones con números, dificultando el aprendizaje de cualquier rama de las matemáticas.

Esta situación se produce, en gran medida porque un gran número de maestros que imparten cursos de matemáticas dan a conocer los temas siguiendo ortodoxamente la secuencia de los mismos que están en un “buen libro” sin preguntarse si la obra es adecuada al nivel de conocimientos los que aprenden, si los temas programados confluyen con los intereses de los alumnos y lo que es más importante: enseñan sin mostrar las relaciones existentes entre los temas, puramente matemáticos, y la realidad.

Con estas dos insuficiencias el alumno se va formando una idea poco agradable de las matemáticas, además que se angustia porque sabe que de todas formas tiene que aprobar la asignatura establecida en el programa de estudios de su escuela. Esta es una causa por las que las matemáticas siguen siendo la parte difícil de las ciencias sociales.

En la formación de un estudiante de ciencias sociales deben aparecer los conocimientos estadísticos, no como los señale el índice de cualquier libro o de un solo libro, sino como lo requieran las condiciones de quienes van a aprender, procurando siempre mostrar los temas de la manera más sencilla y sin olvidar la correspondencia de cada uno con la realidad.

Otra razón importante por la que el estudioso de las ciencias sociales debe incluir en su formación los métodos estadísticos, es la que se presenta cuando el antropólogo participa en un equipo multidisciplinario. Aún que no sea él quien aplique la Estadística debe saber conocer los instrumentos principales de la metodología cuantitativa que usan los médicos, agrónomos, economistas, sociólogos, etc. y poder interpretar los resultados.

CAPÍTULO I

MUESTREO

1. CONCEPTOS FUNDAMENTALES.

Es raro que un libro de Estadística inicie su contenido con un capítulo dedicado al muestreo, ya que este tema generalmente es presentado junto con otros conceptos matemáticos más complejos. Casi siempre los autores lo abordan con detalle más allá de la mitad de sus obras. En otros casos el tema es planteado muy someramente, tanto que dejan muchas dudas al lector, quien luego tiene que recurrir a un especialista.

Por lo general un estudiante de ciencias sociales no está provisto con un amplio instrumental matemático, como para comprender a los autores del primer caso. Por otra parte tampoco cuenta con presupuesto ni tiempo para optar por la segunda alternativa.

Sin embargo, los estudiantes e investigadores tienen una necesidad inobjetable de realizar una captura de datos de una manera válida y fundamentada que permita su confiabilidad pues..."La fecundidad de todo el proceso del análisis y de inferencia estadística descansa en la exactitud de las observaciones empleadas y en la adecuación de esas observaciones a los fines que se utilizan" (MILLS; 1969: 682). Por esta razón es útil repasar algunas definiciones pertinentes al tema del muestreo y establecer los conceptos que van a ser usados frecuentemente en este curso. Los conceptos fundamentales del muestreo pueden resumirse en las definiciones y ejemplos siguientes:

Población: Se llama población o universo a una colección finita o infinita de unidades, sean éstas elementos o eventos definidos por una característica preestablecida.

Ejemplos:

a) El conjunto de municipios del Estado de Yucatán (finita).

b) El conjunto de sitios arqueológicos en el área maya (finita teórico).

c) La colección de variaciones de un mito en una zona indígena (infinita).

Muestra: Cuando para fines específicos se separa una parte de esa colección, esta fracción recibe el nombre de muestra y constituye un subconjunto de la población.

Censo: Se le denomina Censo al procedimiento mediante el cual se registra, mide o verifica las características a estudiar en todas y cada una de las unidades que conforman la población.

Ejemplo:

a) El Censo Nacional que se verifica cada diez años en México y considera a todas las personas y familias como las unidades de la población.

b) El Atlas Arqueológico del Estado de Yucatán que pretende ubicar todos los sitios arqueológicos comprendidos en esa entidad federativa.

c) El registro de todas las haciendas y su valor en el año 1846 ubicadas en la vía Mérida-Campeche.

Muestreo: Es un procedimiento mediante el cual se toma una parte de la población con el fin de conocer o estimar las características de dicha población.

Cuando se realiza un estudio sobre cualquier tipo de fenómeno, encontramos que para conocerlo mejor es necesario observar todas sus partes o elementos. Aunque esta situación es la ideal, difícilmente podrá lograrse en todas o en la mayoría de las veces que lo necesitemos. Las razones de este impedimento son muchas y de distintos tipos, por ejemplo, si se trata de estudiar el avance de una enfermedad de un árbol, la posición de las hojas, las características del árbol, ya sea la altura o el ancho de la copa, etc., será casi imposible el conteo y análisis de las hojas. Otro ejemplo ilustrativo es el siguiente: cuando se revisan las cargas de los explosivos, pues para comprobar que éstos están en buenas condiciones, se tendrían que estallar todos. Entonces habrá que recurrir a un muestreo económico y eficaz.

En el campo de las ciencias sociales las dificultades son de otra índole, pero en esencia, el problema de abarcar todos los casos, es el mismo. Veamos algunas situaciones:

1. Si se tratara de estudiar las condiciones socioeconómicas de la fuerza de trabajo en la industria del Estado expresadas por los obreros contratados en un momento determinado, sería muy difícil conocerla en su totalidad ya que todos los días hay altas,

bajas, nuevas plazas y otros eventos que hacen variar a la población de trabajadores en cada día.

2. Si se tiene la necesidad de conocer las características de la afluencia turística en nuestro país en el momento presente para calcular los ingresos obtenidos gracias a esa actividad, habría que informarse acerca del número de turistas, la vía por la cual llegaron, sus destinos principales y días de estancia. Esta labor requiere de mucho tiempo y personal que acopie la información oportunamente.

3. Otro ejemplo nos lo proporcionan las encuestas de opinión que se realizan en algunos países con el fin de conocer la preferencia de los votantes hacia un candidato o un partido. Si esta encuesta abarcara a toda la población ya no tendría caso llevar a efecto el sufragio oficial.

4. Para aproximar el uso de las cuevas y cenotes por parte de los antiguos mayas en el estado de Yucatán, habría que investigar en el interior de estas cavidades después de ubicarlas geográficamente. Si se considera que hay unas 3000 cavernas o cenotes en Yucatán, los requerimientos en tiempo, presupuesto y personal rebasarían los parámetros acostumbrados.

En todos los casos mencionados se percibe la necesidad de un muestreo económico, eficaz y oportuno. El **objetivo del muestreo** es tratar de conocer las características de una población a través de una parte de la misma. El proceso de muestreo elimina las dificultades que se presentan al tratar de conocer la población completa, pero al mismo tiempo nos plantea otro problema que es necesario afrontar:

¿Cómo podemos saber si la muestra elegida para nuestro estudio contiene determinadas características de la población si sólo estamos observando una parte del total?

Para que este problema se solucione positivamente debemos elegir una muestra representativa de las características de la población; esto se puede lograr utilizando los procesos establecidos de muestreo. Entonces la característica indispensable de toda muestra para lograr resultados válidos y confiables es la **representatividad**

Sin embargo, a pesar del rigor del muestreo, paradójicamente, no siempre se puede garantizar la representatividad de la muestra, porque para comprobarla

tendríamos que compararla con las características exactas de la población, no siendo esto siempre posible. Y si fuera posible entonces no tendría caso la labor de muestreo.

Por otra parte, entre las **ventajas** que el muestreo llega a proporcionar, la del **costo** resulta sin duda alguna, significativa. Por ejemplo, en la aplicación de una encuesta a una población cuyos elementos alcanzan cifras considerables, la inversión en dinero que tendría que soportar en gastos de sueldos de encuestadores y de materiales sería bastante elevada en contraste a los costos ocasionados por una muestra tomada de esa misma población.

La reducción del costo no es la única ventaja que nos proporciona el muestreo, sino que la reducción del **tiempo** es otra de las ventajas. A veces la **oportunidad** de la información es vital para su uso; como el caso de las encuestas preelectorales basadas en muestras, tiene el fin de expresar la tendencia de la votación antes de las elecciones. Un estudio censal que nos revelara la tendencia del voto después de las elecciones ya no tendría caso alguno.

Además de los requerimientos de tiempo, dinero, personal y otros factores, un censo completo no elimina el error, dice Lohr; con frecuencia, la subcobertura, la carencia de respuestas y los descuidos en la recolección de datos, produce inexactitudes sin control. El muestreo no es solamente una sustitución de una cobertura total por una parcial, sino que es la ciencia y arte de controlar y medir la confiabilidad de la información estadística útil a través de la teoría de la probabilidad (LOHR; 2000: 16-17)

Ahora bien, al trabajar con un número reducido de elementos, dentro de un margen racional, se puede conseguir o producir una mayor **exactitud relativa** en los resultados, con respecto a la que obtendríamos con un mayor volumen de datos. Esto puede explicarse porque el personal encargado de la recolección de datos y de su posterior tratamiento, tiene la oportunidad de una mayor especialización en el campo de trabajo. La etapa que sigue a la recolección, esto es, la del procesamiento, se puede desarrollar en forma más eficaz debido a que se trabaja con un menor número de datos. En conjunto estas circunstancias permiten disminuir la influencia de las fuentes de error aumentando el control sobre ellas, y en consecuencia, se puede obtener

relativa ventaja en cuanto a la exactitud de la muestra (COCHRAN; 1976: 21).

Sin embargo, esto no quiere decir que una muestra aportará información más exacta en la medida que la hagamos más pequeña; porque a pesar de que los métodos bien realizados nos aseguren una representatividad aceptable y una exactitud considerable o significativa en el manejo de los datos, nunca se debe olvidar que estudiar una muestra es analizar solamente una parte de la población, y por lo tanto, siempre habrá de surgir una diferencia entre las conclusiones que obtengamos de ellas y las de la población. A estas diferencias se les llama **error de muestreo**.

Teóricamente y siguiendo una lógica al planteamiento anterior, se puede afirmar que el error de muestreo será más pequeño en cuanto la muestra sea mayor, porque las características descritas en las muestras serán inferidas a partir de un mayor número de elementos en común con la población. Si por el contrario, se toma un tamaño menor de muestra, entonces el error de muestreo aumenta. Por ejemplo:
Si tomamos experimentalmente tres muestras (N1, N2 y N3) a partir de una población cuyo número total de elementos es de 15000 deberíamos esperar que el error de muestreo se comporte de la siguiente manera:
Si N1 = 500, entonces el error de muestreo fuese del 10.0 %
Si N2 = 1000, entonces el error de muestreo fuese del 7.5 %
Si N3 = 1500, entonces el error de muestreo fuese del 5.0 %

Por supuesto que la relación entre el tamaño de la muestra y la magnitud del error de muestreo no puede darse de esa manera proporcionalmente exacta, porque en todo proceso de muestreo hay **fuentes de variabilidad** tales como la dispersión de los datos, la precisión del instrumento de captura, la confiabilidad del proceso cuantificador y la calidad del desempeño por parte del personal.

2. ASPECTOS PRÁCTICOS DEL MUESTREO.

Hasta aquí se han tocado los aspectos más generales y teóricos del muestreo, así como sus ventajas y desventajas resultantes con la intención de sopesarlos para cuando llegue el momento de elegir o construir un diseño apropiado. Hay otros aspectos un poco más prácticos cuya mención no puede evadirse pues están ligados

con el proceso de investigación social. He aquí algunos de ellos:

a) El interés del investigador.

Un mismo fenómeno social podría generar dos estrategias de muestreo distintas y con diferentes tamaños de muestra. Supongamos que un psicólogo y un antropólogo se han avocado al estudio del sincretismo religioso maya católico. Si el psicólogo apunta sus objetivos al estado actual de las prácticas de ese sincretismo quizá con un cuestionario aplicado a una muestra ampliamente distribuida entre las familias o individuos que forman la población satisfaga sus objetivos. En cambio, si el antropólogo se interesa por conocer el desarrollo de las instituciones sociales que permitieron y conformaron ese mismo sincretismo, podría ser que con una breve selección de informantes claves, apoyado con alguna información de archivo, pueda cubrir sus metas.

b) Las condiciones de la investigación.

Si por su trascendencia una investigación dispone de un amplio margen de tiempo y presupuesto abundante, no cabe duda que la instrumentación puede llegar a tal rigor que cubra todas las posibles fuentes de error y el acopio de los datos arrojen una información muy confiable. Pero en otras ocasiones los investigadores no necesitan más que una idea general de las características de la población o quizás su presupuesto esté limitado por otras instancias, o bien, sea porque resulte apremiante la resolución de un problema, entonces es conveniente diseñar un muestreo que cubra las condiciones más elementales de objetividad y, por lo tanto, pueda avalarse cierto grado de representatividad. Siempre será preferible tomar una decisión con un grado mínimo de garantía que obtener una muestra sin alguna base razonable y sustentada.

c) Las condiciones de la población a estudiar.

Es conveniente que el investigador explore cuidadosamente la naturaleza de la población en la que se hará el muestreo. Si la población no presenta mucha variabilidad, quizá no se requiera de una elaboración compleja de estrategias aleatorias o de una muestra grande. Por ejemplo: Si en una comunidad de hablantes monolingües mayas se estudia únicamente las variaciones gramaticales de ese idioma, pues quizá con una muestra simple aleatoria se resuelva el problema de la toma de datos. Pero si lo que se intenta es conocer cómo esas variaciones se corresponden con variables de

escolaridad e ingreso económico en la misma comunidad, entonces habrá que considerar el grado académico alcanzado y el estrato socioeconómico de los entrevistados, es decir, si los valores de la población son muy heterogéneos con respecto a esa variable, entonces el diseño muestral será más complicado y el tamaño de la muestra necesariamente será mayor.

3. TÉCNICAS DE MUESTREO.

Las técnicas o tipos de muestreo son en realidad una serie de procedimientos que permiten al investigador seleccionar un subconjunto de elementos llamado **muestra** de la **población**, en los términos en que ya se han definido. Puede expresarse también como un proceso en el que elegimos un sector 'X' de una población, con el fin de conocer las características de esa parte e inferir por medio de ellas, las de la población de donde se ha tomado.

Para que el estudio de la muestra sea válido y funcional, ésta debe ser **representativa**, es decir, las características de la muestra deben aproximarse en el mayor grado posible a las características de la población de la cual se está extrayendo.

Existen dos tipos fundamentales de muestreo: los **probabilísticos** y los **no probabilísticos**. En el primer tipo de muestreo se debe proceder de tal forma que todos y cada uno de los elementos de la población tengan la misma probabilidad de ser seleccionados. Por otra parte, en el segundo tipo de muestreo, los elementos son escogidos en circunstancias especiales por lo que no se garantiza la igualdad ya mencionada.

3.1 Muestreos Probabilísticos

Para evitar confundir el significado del término **probabilidad** utilizado en Estadística con el sentido que se le da en la comunicación coloquial repasaremos las dos concepciones conocidas:

Definición clásica: La probabilidad de que suceda un evento es igual número de casos favorables dividido entre el número de casos posibles.

Definición estadística: La probabilidad de aparición de un evento es igual a la

frecuencia relativa que haya presentado ese evento.

En la primera definición la fórmula se aplica antes de que el fenómeno ocurra y presupone el conocimiento de los casos posibles. La segunda definición sólo se puede aplicar cuando el evento ya se ha presentado determinado número de veces y su magnitud está en función del número total de casos registrados. Cualquiera que sea el concepto que se aplique para conocer la probabilidad de que un elemento sea seleccionado, la magnitud de ésta deberá ser la misma en cada extracción.

A su vez, los muestreos probabilísticos se subdividen en simple aleatorio y aleatorio estratificado, en tanto que los muestreos no probabilísticos tiene sus propias subdivisiones y son denominadas: de porcentajes o proporciones, sistemático o mecánico y el muestreo a juicio. Existen otras variaciones en el proceso de muestreo tales como el muestreo polietápico, el polifásico, el muestreo por áreas y la muestra repetida.

3.1.1 Muestreo Simple Aleatorio.

En este proceso de selección se extrae un elemento a la vez del total de la población, asegurándose que todos los elementos que han de construir la muestra hayan sido seleccionados utilizando un mecanismo impersonal, esto es, que no intervengan las preferencias e intereses del investigador. Existen dos mecanismos principales; el sorteo y la tabla de números aleatorios.

a) Sorteo. Para realizar este proceso es necesario que se enumeren individualmente las 'N' unidades que conforman la población. Los 'N' números se copian sobre tarjetas, fichas, etc.; se mezclan después en urnas o recipientes, y se extraen al azar las 'n' unidades que conformarán la muestra simple aleatoria.

b) Las tablas de números aleatorios.

Si los 'N' elementos de una población total, se enumeran en serie de 1 a 'N' se puede extraer una muestra aleatoria más fácilmente y con mayor fiabilidad mediante la utilización de tablas de números aleatorios ya preparadas. Esas tablas capacitan al investigador para seleccionar aleatoriamente 'n' números de la lista ya completa de los números de la serie de 1 a 'N'. Actualmente los números aleatorios se pueden obtener

por medio de programas de cómputo. (Ver Tabla al final)

3.1.2 Muestreo Aleatorio Estratificado

Si la población presenta sectores bien definidos y necesitamos que los elementos que los forman estén proporcionalmente representados en la muestra, ésta se puede formar con la extracción de elementos en números preestablecidos, siempre con un mecanismo de azar. Ejemplo. Se necesita conocer la opinión de algunos grupos de profesionales para tomar una decisión sobre la disyuntiva de legalizar o no el uso de la marihuana. Pero dada la magnitud de la población sólo se tomará una muestra que abarque el 10% del total.

Profesión	Totales	10%
Médicos	5000	500
Abogados	3000	300
Químicos	2500	250
Antropólogos	400	40
	10900	1090

Si el investigador ha decidido utilizar un muestreo estratificado, queda por resolver otra cuestión ¿Cómo se hará la selección por cada estrato? Conviene establecer de una vez que la porción que habrá de tomarse por cada estrato es denominada **afijación** y existen tres tipos: afijación porcentual, afijación uniforme y afijación óptima (COCHRAN; 1976: 149-161).

Se usa la **afijación porcentual** para garantizar en la muestra la misma proporción que los estratos tienen en el universo. Se requiere el conocimiento preciso o muy aproximado de la magnitud de cada estrato, para determinar la afijación en términos absolutos. Ejemplo: Obténgase en el siguiente caso una muestra aleatoria estratificada con afijación porcentual (10%) agricultores en la comunidad de Texán de Palomeque, Hunucmá.

Agricultores	N	N = 10%
Independiente	30	3
Jornalero	80	8
Ejidatario	140	14
	250	25

La **afijación uniforme** se usa en muestreos exploratorios y/o cuando se pretende hacer comparaciones entre categorías de igual magnitud. Se descarta la importancia de mantener alguna proporción entre los estratos ya que probablemente no se conozca la magnitud de cada uno y sea difícil su estimación.

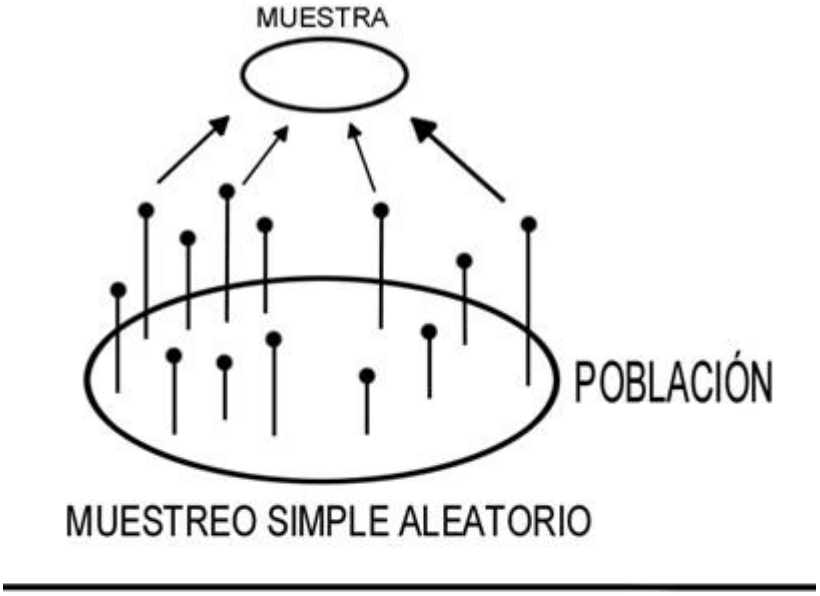
Ejemplo: En el año 2003 una organización internacional que promovía el desarrollo económico y social aplicó una evaluación a 40 mil estudiantes mexicanos de nivel básico para saber cual es la calidad de su desempeño en las áreas de lectura, matemáticas y ciencias naturales. La cantidad de cuestionarios por cada uno de los 32 estados de la república fue de 1250 pruebas, sin considerar sus distintos números de habitantes. De acuerdo con ésta información se trató de un muestreo con una afijación de tipo uniforme; en este caso, la población estudiada fueron todos los estudiantes mexicanos del nivel básico del país y los estratos estuvieron conformados por las entidades federativas.

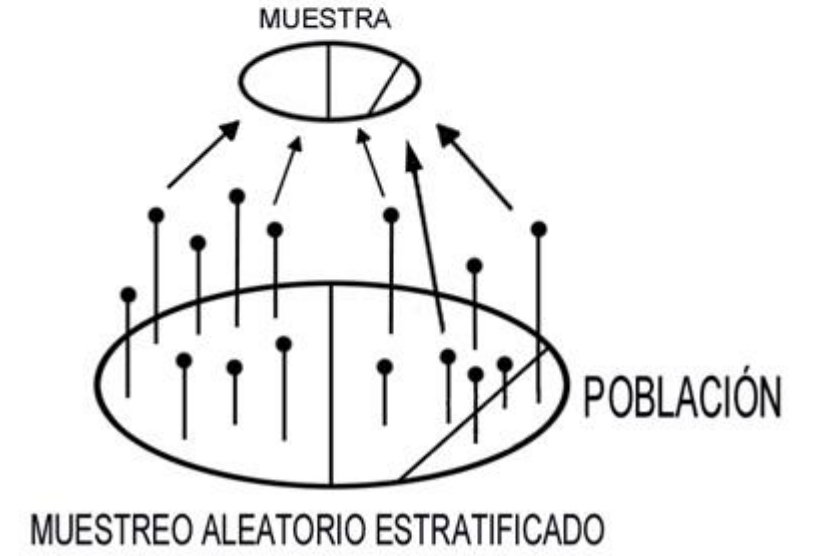
Si las categorías de la población presentan distinto grado de heterogeneidad en los valores que contienen, es preferible usar una **afijación óptima** que resuelva el problema de la variabilidad de cada estrato. Entonces, el investigador determinará una afijación de distinto tamaño para cada estrato. En el siguiente cuadro se expresan cuatro categorías con distinto grados de heterogeneidad interna pero se necesita conocer su opinión sobre la política económica de su país. La primera está formada por personas que ganan el salario mínimo (máxima homogeneidad), luego siguen los burócratas de bajo rango (poca heterogeneidad), la clase política (mucha heterogeneidad) y los empresarios (máxima grado de heterogeneidad).

	Población	0.01% N %	0.01% n Op
S. mínimo	100000	1000	964
B. b. r.	6000	60	60
Políticos	1000	10	16
Empresarios	3000	30	60
	110000	1100	1100

En otras ocasiones la afijación óptima se vuelve obligatoria porque algunos estratos de

la población son muy pequeños y no pueden ser representados en forma estrictamente proporcional en la muestra.





3.2 Muestreos No Probabilísticos.

Es el tipo de muestreo cuyo procedimiento se aplica en circunstancias especiales tales como en aquellos casos en los que hay una gran homogeneidad entre los datos del universo, un limitado acceso a la población y una diferente intensidad en los elementos que conforman la variable estudiada. Este procedimiento excluye el uso del azar estadístico o los mecanismos impersonales de selección de datos. A continuación explicaremos estas circunstancias especiales a la que nos hemos referido.

Homogeneidad. A veces los elementos de las poblaciones se encuentran en alto grado de homogeneidad, por lo que resulta innecesario el uso de un muestreo probabilístico. Esta condición puede provenir de un proceso mecánico o químico como sucede con los productos fabricados en serie. En ciencias sociales es posible encontrar situaciones similares ante procesos legales (ser o no casado, estar o no divorciado, ser ciudadano de un país o extranjero, etc.) o naturales como la viudez y la ancianidad.

Acceso a la población. En ocasiones la muestra se restringe por causas insalvables o sumamente difíciles de superar. Por ejemplo, en los casos de conductas delictivas los posibles informantes están en condiciones difíciles de abordar y por lo tanto de colaborar; en otros casos de escasa colaboración puede ser cuando se trabaja con residentes ilegales, drogadictos y otra clase de personas que no desean hacer pública su conducta. Como último ejemplo se cita la temática del suicidio en la que se realiza entrevistando a los parientes, amigos (informantes indirectos) y a las personas que sobrevivieron al intento.

En cuanto a la última circunstancia especial, el distinto grado de **intensidad** de los elementos que integran la variable, debe entenderse que se procura seleccionar a aquellas unidades que presenten con mayor intensidad la característica a estudiar. Hay una situación en el tema de la medicina tradicional que puede ejemplificar este caso. Algunos practicantes de la herbolaria curativa suelen tener características que los hacen más auténticos que otros: el hecho que hablen el idioma indígena, que no cobren por sus servicios y que estén avecindados en la comunidad en donde desempeñan sus servicios. La presencia de estas tres particularidades los define como elementos más confiables respecto a aquellos que no tienen estos atributos. Entonces no se debe

correr el riesgo de una selección al azar, la cual podría tomar los casos menos representativos.

Al revisar la literatura sobre el muestreo no probabilístico encontramos que existen las siguientes modalidades: muestreo de cuotas, muestreo sistemático, muestreo a juicio y muestreo sin norma.

3.2.1 El Muestreo de Cuotas.

Es un tipo de muestreo no probabilístico que puede ser llamado también de porcentajes o proporciones. Se efectúa en encuestas de opinión acerca de temas generales y/o fases exploratorias. Sus categorías generalmente son de sexo, edad, ocupación, nacionalidad y nivel socioeconómico. Las variables más frecuentes son filiación política, consumo, opinión sobre servicios públicos, etc.

En esencia el muestreo de cuotas es un tipo de muestreo estratificado que puede llevarse a efecto con las afijaciones porcentual y uniforme. No se descarta el uso de directorios y otra clase de registros que permita tener una magnitud aproximada de la población o de los estratos (DUVERGER; 1978: 198 - 203).

Un ejemplo de este tipo de muestreo es el que se efectuó en los tres países participantes en el Tratado de Libre Comercio (TLC) para América del Norte con el fin de conocer la aceptación del acuerdo entre los ciudadanos. Desde 1990 se aplicaron 2000 cuestionarios con 20 preguntas a un número igual de personas en cada país y se observó una tendencia general de aceptación; pero las mismas encuestas aplicadas en 1991 y 1992 revelaron un cambio de opinión a tal grado que los expertos detectaron que conforme pasaba el tiempo menos personas apoyan el TLC. En este caso, los estratos son representados por los países en cuestión, cada uno con una población distinta y la cuota de 2000 personas por nación es una **afijación uniforme**. Cada una de las porciones se convierte en poblaciones a comparar entre sí, de las cuales se obtienen datos a partir de un mismo cuestionario.

3.2.2 Muestreo Sistemático o Mecánico.

Consiste en un procedimiento sencillo siempre que se disponga de una población con elementos identificados en una lista. Ya numerados se procede a su selección tomando el primero aleatoriamente y los siguientes con un intervalo denominado "Razón de Muestreo" (RM). Esta (RM) es un simple cociente que se encuentra dividiendo el tamaño de la población (N) entre el tamaño de la muestra (n).

Ejemplo: Con una población de 1500 elementos se requiere tomar una muestra de 75 casos; por lo tanto la (RM) es de 1500 entre 75 = 20. Si se toma como primer caso el número 8 habrá que tomar, por consecuencia, el 28, el 48, el 68 etc. hasta llegar al 1488; para entonces se habrán completado los 75 elementos de la muestra.

Al aplicarse el muestreo sistemático es importante verificar si los elementos de la lista están o no ordenados en cuanto a su magnitud. En caso afirmativo, sea el orden ascendente o descendente, la validez de la muestra será cuestionable pues sus cualidades dependerán del primer dato que se tome; si es un dato de alto valor, por consecuencia, los siguientes también lo serán y sucederá lo correspondiente si se empieza con un valor bajo (STEVENSON; 1985: 197).

3.2.3 Muestreo a Juicio o Criterio.

En este tipo de muestreo se establece de antemano un atributo que tendrá que cumplir cada elemento de la muestra de acuerdo con algún un criterio o juicio que el investigador considere válido o suficiente. Normalmente se utiliza para conformar muestras pequeñas de elementos o eventos poco frecuentes con los que no se puede aplicar los principios de aleatoriedad. Por ejemplo, si el interés del investigador reside en definir los rasgos determinantes de las personas que son líderes, entonces deberá tomar como elemento de muestra aquellos que simplemente presenten ese atributo y se desecharán los casos que no lo tengan o de los que haya dudas. El peor inconveniente de este tipo de muestreo es que la composición de la muestra está totalmente influida por la subjetividad del investigador y por lo tanto no puede sustentarse su confiabilidad por medio de métodos estadísticos.

3.2.4 Muestreo Sin Norma

Se realiza tomando, del total de la población, una parte sin ningún procedimiento que discrimine a unos elementos, porciones o eventos de otros porque se cuenta con el supuesto de que todos sean iguales. Otra característica que se le puede atribuir es que no pueda analizarse toda la población por razones prácticas. Por ejemplo, cuando una persona necesita un análisis sanguíneo, no puede sacar toda su sangre para tal propósito; una pequeña muestra es suficiente para el caso. De manera similar, un catador de vino no necesitará probar todo un tonel para conocer la calidad de la bebida; con una cantidad mínima será suficiente.

3.3 Variaciones

Debido a la diversidad de la naturaleza de cada población, quienes diseñan el muestreo recurren a estrategias o variaciones que les permitan acceder a la población sin menoscabo en la calidad de la muestra. Estas variaciones son llamadas también muestreos pero en realidad son procedimientos combinados, los cuales son requeridos ante la especificidad de cada caso pero siempre regidos por los principios de objetividad y representatividad. Los más usuales son los siguientes:

3.3.1 Muestreo Polietápico

Cuando la población es compleja y para llegar a cada unidad de muestreo es necesario realizar el procedimiento de selección en dos o más etapas. Por ejemplo, si se quiere obtener datos para estudiar la economía doméstica y calidad de vida de las familias en una ciudad, es posible que primero sea necesario seleccionar los barrios o colonias y después elegir las manzanas o cuadras incluidas en esos barrios. Por último, se seleccionará un determinado número de casas de las manzanas antes escogidas para entrevistar a las familias que allí vivan.

3.3.2 Muestreo Semiprobabilístico.

Este muestreo puede considerarse una variación del Polietápico, porque es una estrategia que se aplica para tomar una muestra bien distribuida sin correr el riesgo de

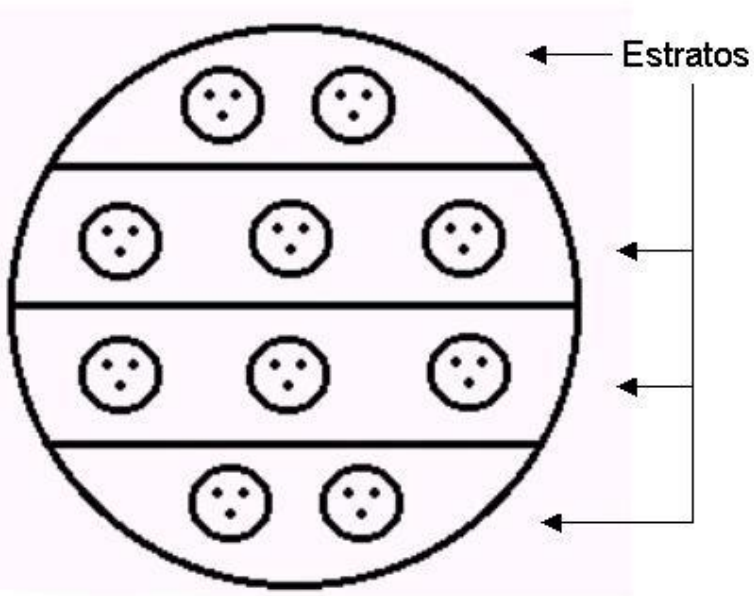
azar y sin renunciar a la objetividad de un muestreo. Es una combinación de los procedimientos probabilísticos y no probabilísticos. Citando el ejemplo anterior de la economía doméstica en la ciudad, el investigador podría escoger a su juicio los barrios o colonias que estén distribuidas estratégicamente en el área urbana; pero al escoger las manzanas o cuadras convendría utilizar el mecanismo impersonal de selección. Esta sería la modalidad del muestreo semiprobabilístico inferior. Si la selección de los barrios es aleatoria y la de las manzanas o las casas se realiza con un criterio específico del investigador, entonces estaríamos ante la modalidad del muestreo semiprobabilístico superior. Otro caso, más frecuente por cierto, es cuando se necesita la opinión de los estudiantes de una escuela acerca de un determinado asunto, entonces se seleccionan al azar los grupos escolares y se entrevista a cierto número de estudiantes de cada grupo escogidos arbitrariamente (AZORÍN; 1972: 8 y 17).

3.3.3 Muestreo Polifásico

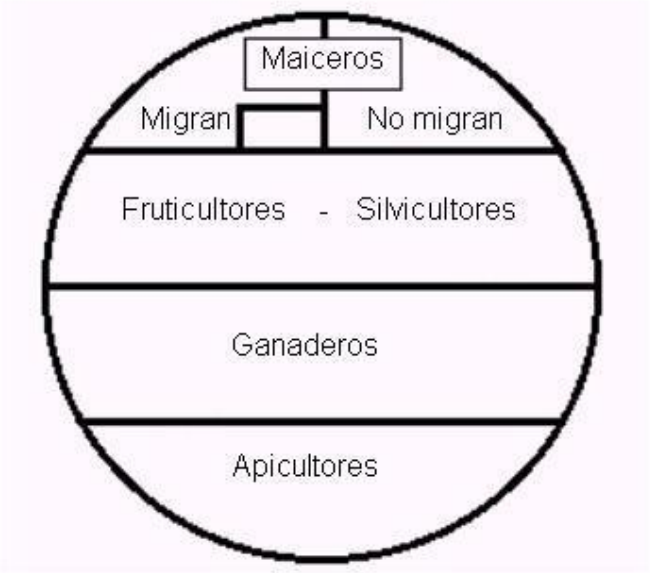
Se trata de una estrategia en la que primero se toma una muestra grande con la perspectiva de obtener de ella otras submuestras en una o varias fases posteriores. Por ejemplo, si se pretende analizar cuáles son los motivos que hacen migrar a los habitantes vinculados a las actividades agropecuarias y en especial a los cultivan maíz, se deberá proceder de la siguiente forma: primero se obtiene un censo o una muestra grande con alguna forma de registro disponible, después se descartan los estratos que no son parte del objetivo de la investigación. Luego, se apartan a los maiceros que no migran y finalmente se toma una submuestra o simplemente muestra en la categoría de maiceros migrantes.

Los resultados de esta última fase se podrían comparar con otras submuestras de otros estratos aprovechando que ya estarían disponibles los datos contenidos en la muestra inicial gracias a este procedimiento.

Polietápico



POLIFÁSICO



3.3.4 Muestreo por Áreas

Cuando no existe algún registro confiable de los elementos que conforman la población y sus elementos resultan difíciles de enumerar, se recurre a un tipo de muestreo en el que las unidades son conglomerados o áreas. Un conglomerado es un grupo de unidades de muestreo del cual sí se puede extraer una muestra. Quizá el ejemplo más frecuente sea el de las manzanas que contienen a las casas y a la vez, a los sujetos de estudio, como se explicó en el caso del Muestreo Polietápico. Es posible que esos conglomerados sean superficies o áreas en que sea dividido el terreno ocupado por la población estudiada. En estos casos el investigador se ve precisado a usar mapas, planos y otros medios para delimitar la población y sus unidades de muestreo. Ocasionalmente este tipo de muestreo ha sido utilizado en exploraciones arqueológicas en donde el área investigada es relativamente grande y se vuelve necesaria una selección de los puntos específicos previa a la excavación (GARDUÑO; 1979: 38-41).

3.3.5 Muestra repetida.

En el caso de las poblaciones que presentan cambios a intervalos regulares y su dinamismo impida la reunión de todas para su estudio, pueden diseñarse muestras repetidas bajo condiciones establecidas previamente. Por ejemplo, si se necesita estimar la variación en la afluencia diaria de los usuarios a un medio de transporte público, el investigador deberá tomar varias muestras que considere los horarios de trabajo, de las escuelas y otros factores que puedan incidir en las observaciones. En este caso la población es infinita porque estaría constituida por un número indeterminable de observaciones realizadas en el horario total de servicio.

TIPOS DE MUESTREO

	Simple	Mecanismos impersonales	Sorteo y N° Aleatorios
Probabilístico			
	Aleatorio	A. Proporcional	
	Estratificado	A. Uniforme	
		A. Optima	
	De Cuotas	A. Proporcional	
		A. Uniforme	
No Probabilístico	Sistemático o Mecánico		
	A Juicio		
	Sin norma		
	Polietápico		
	Semiprobabilístico		
Variaciones	Polifásico		
	Por Áreas		
	Muestra Repetida		

4. TAMAÑO Y DISEÑO DE LA MUESTRA.

4.1 La relación entre el tamaño y la representatividad.

Sabemos que hay diversas fórmulas confeccionadas con el objetivo de medir el grado de confianza de una muestra. Estas fórmulas tratan de encontrar un equilibrio entre el tamaño de la población (N), el tamaño de la muestra (n) y la desviación típica (s). Finalmente se obtiene un tamaño de muestra para trabajar con cierto grado de confianza bajo el supuesto de que el fenómeno tenga un comportamiento probabilístico igual a aquellos que describen una curva normal o sus aproximaciones binomiales.

Claro está, que en cada fenómeno, habrá algunas inexactitudes respecto a la distribución teórica de probabilidades de que se trate, pero despreciando estas pequeñas anomalías el modelo funciona. Sin embargo es necesario reflexionar sobre lo apropiado o no que puede ser el tratar de establecer un tamaño de muestra con base a aquellas fórmulas probabilísticas y más aún cuando se presume que pueden medir el nivel de confianza y el porcentaje de error. Las razones son las siguientes:

Primera. La naturaleza de las poblaciones en las ciencias sociales es tan diversa en lo general y tan específica en los casos concretos que en muchas ocasiones no se ajustan a un modelo de distribución normal, como en el tema de los salarios nominales que nunca muestran un comportamiento parecido al normalizado. Además estos modelos se crearon en función del comportamiento de una variable y en el campo de la investigación social siempre se relacionan y estudian simultáneamente más de una variable que permitan explicar o matizar la hipótesis principal. No es posible concebir que un investigador elabore un costoso diseño de muestreo solo para preguntar sobre la escolaridad de sus informantes sin abordar la edad, sexo, condición económica, etc.

Segunda. Si todos los autores coinciden en la existencia invariable del error de muestreo, es decir que a un menor número de datos corresponde un mayor error de muestreo, sobra decir que la **mejor muestra** será la del **mayor tamaño posible** y en este caso es innecesario conocer sus probables niveles de confianza porque una muestra mayor obviamente nos conferirá un menor error de muestreo.

Nunca deberá descartarse la posibilidad de que se presente una situación tal que

una muestra de mayor tamaño pueda ser menos representativa que otra por defecto de concepción en el diseño del muestreo o por falta de control en las fuentes de error como lo sería un cuestionario mal construido, datos falseados, información omitida, etc.

Uno de los aspectos del muestreo que chocan contra la intuición, es que lo mismo puede tomar una muestra de 500 casos para una población de 10000 unidades que para otra de 100,000. Esto se debe a que el tamaño de la población no tiene gran influencia en la estimación del tamaño de la muestra como puede verse en diversas fórmulas. Más adelante, podrá observarse que no es la magnitud, si no la variabilidad de la población, la que tiene que ver con el tamaño de la muestra y la precisión de las estimaciones (MONTEMAYOR; 1973: 356).

4.2 El objetivo de la investigación y el diseño de muestreo.

El **diseño de muestreo** es un plan preconcebido de acciones sucesivas encaminadas a definir concretamente el proceso de selección de unidades que habrán de conformar la muestra. El diseño debe sujetarse en primer término a dos factores inherentes al proceso, el primero es el **objetivo de la investigación** y segundo, la **naturaleza de la población**. Este plan de acciones debe exponer los pasos que se realizarán para extraer los elementos de la muestra y si es necesario, deben presentarse los criterios que justifiquen las acciones. Por estas razones debe tenerse presente que no hay dos diseños de muestreo iguales ya que cada investigación tiene sus propósitos específicos y cada población es distinta con base al objetivo de la investigación.

En el diseño de muestreo se deben considerar las distintas situaciones especiales que pueden presentarse en la captura de los datos, las cuales podrían afectar la magnitud de la muestra y las estrategias previamente fijadas. Al respecto, un punto importante a destacar es la forma preestablecida para obtener las **unidades de reposición** cuando el caso lo amerita. Ejemplo, si en una de las casas seleccionadas para aplicar un cuestionario, acaba de ser desocupada y por tanto no se puede contar con ese dato, debemos establecer previamente cuál será el mecanismo para reponer el faltante y no se altere el tamaño de la muestra. Entonces debemos establecer si aplicamos el cuestionario en la casa siguiente o en la anterior. Por último, para elaborar

un diseño de muestreo debe tenerse presente los principios de aleatoriedad y objetividad.

4.3 El muestreo en los estudios históricos.

Un comentario final acerca del muestreo es acerca de la precaución que se debe tener en las investigaciones históricas. A menudo un investigador se pregunta si conviene usar algún tipo de muestreo para conocer la información de un archivo en particular, que dicho sea de paso, por lo general son grandes y su revisión lleva mucho tiempo.

Representa un riesgo aplicar un muestreo a un conjunto de documentos porque podría no ser seleccionado aquellos que hablen de un hecho relevante. Por ejemplo, supóngase que en un estudio de la historia de Yucatán en el siglo XIX no se consulte el año de 1847 cuando se inició la Insurrección Indígena. Si el investigador no revisa los datos de ese año, entonces será más difícil entender los acontecimientos posteriores.

Si se toma la decisión de aplicar un muestreo, debe pensarse en un procedimiento que no descarte la información básica para que no se distorsione la realidad histórica.

5. ETAPAS PRINCIPALES DEL MUESTREO

5.1. Definición del tema y la variable estudiar.

Antes de realizar el diseño de muestreo es preciso señalar el tema y la variable que se va a estudiar. Ejemplo: La juventud como tema de estudio puede tener diversas variables:

a) Escolaridad
b) Participación política
c) Drogadicción

5.2. Definición de la población.

Es de primordial importancia determinar la población sobre la cual se aplicará el muestreo. Esta determinación se realiza a través de los criterios previamente establecidos que delimitan el alcance de las conclusiones que tendrá la investigación;

por ejemplo se puede estudiar:

a) La juventud de México
b) La juventud de Yucatán
c) La juventud campesina de Maxcanú (Municipio)
d) La juventud campesina del barrio San Juan de la comunidad de Maxcanú

5.3. Técnicas de captura y medición de la variable.

En esta etapa se debe establecer cuáles serán las técnicas de captura de los datos pues de éstas depende el adecuado manejo de los mismos. Además debe visualizar las formas de registro más apropiadas para el tipo de datos y el análisis que se pretenda. Por ejemplo si el investigador pretende establecer promedios de participación política, le será preciso crear una escala lo mejor graduada posible para que esta variable sea agrupada en una serie de intervalos numéricos y no conformarse con una serie de categorías nominales a las cuales no les podrá aplicar la media aritmética.

5.4. La definición de la unidad y el marco del muestreo.

Para una correcta constitución de la muestra es fundamental definir cual será la **Unidad de Muestreo**, es decir cuál es la clase y magnitud del fenómeno que se considerará como elemento de la muestra. Este elemento deberá encontrarse en toda la extensión de nuestra población. Por ejemplo:

El estudio de la participación política (variable) en la juventud (tema) de los habitantes del barrio de San Juan de Maxcanú (población) será analizado a partir de las respuestas de los cuestionarios (técnica) que cada uno de los jóvenes campesinos no mayores de 21 ni menores de 15 años (unidad de muestreo) den al técnico de la encuesta.

Siempre que sea posible se tratará de obtener un registro escrito o lista de todos los elementos de la población que sean susceptibles de ser seleccionados; esta lista que contenga a las posibles unidades de muestreo se le denomina **Marco de Muestreo** y la ventaja adicional que nos ofrece es que nos permite conocer la magnitud de la población ya depurada de los casos imperfectos, lo que aumenta el grado confiabilidad

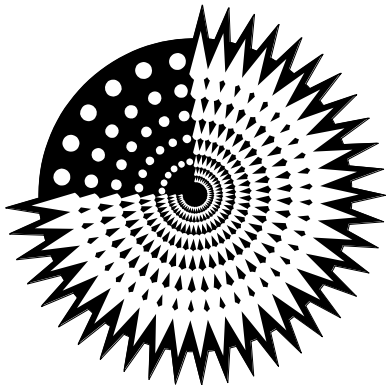
estimada de la muestra.

5.5 La obtención de la muestra.

Para elegir una técnica adecuada de muestreo es necesario conocer y evaluar los marcos o listados de elementos disponibles de la población. Por ejemplo, una sociedad local de agricultores puede tener una lista aparentemente completa de sus integrantes, pero al evaluar se advierte la falta de actualización hallándose nombres de personas que han fallecido o abandonaron la comunidad para dedicarse a otras actividades económicas. Peor aún, si en esta lista falta nombres de socios que recientemente se han incorporado a la asociación.

En otros casos el investigador se encuentra con el inconveniente que los marcos de muestreo existentes le son inaccesibles por lo que tendrá la tarea de elaborarlos con base a sus necesidades. Entonces será preciso recabar la mayor información posible, ya sea oral o documental, para aproximarse al conocimiento real de la población.

Posteriormente generará, a partir del paso anterior, un Marco de Muestreo confiable cuya naturaleza le permitirá decidirse por un tipo de muestreo específico. Es indispensable señalar que el apego riguroso al diseño de muestreo establecido nos llevará a las condiciones objetivas necesarias para fundamentar la representatividad de la muestra.



CAPÍTULO SEGUNDO

ORGANIZACIÓN DE DATOS

1. GENERALIDADES.

Cuando se lleva a cabo una investigación, afluyen a los analistas una gran cantidad de datos de distinta naturaleza, por lo que el investigador debe saber distinguir y seleccionar aquellos que estén relacionados con las hipótesis fundamentales de su trabajo.

Para poder hacer un análisis de los datos que se han obtenido de la realidad, es muy importante saber ordenarlos de forma tal, que nos muestren sus características de una manera relativamente fácil. Así por ejemplo, si estudiamos las condiciones físicas de un grupo de niños de un estrato socioeconómico determinado y necesitamos describirlos, será difícil comprender las características del conjunto si mencionamos una a una las condiciones de cada niño. En cambio, si las agrupamos en estratos y clases, podemos hacer generalizaciones y también compararlas con otros grupos.

2. SERIES ESTADÍSTICAS.

Las series o tablas estadísticas constituyen una técnica, a la vez un instrumento que hace disponible a una variable para su descripción de acuerdo con los valores y frecuencias que presente el fenómeno estudiado.

Las series estadísticas resumen la gran cantidad de detalles que provienen de las variables y facilitan el análisis ya que en este proceso los datos se convierten en información y con ello permite establecer relaciones internas de las variables (STEVENSON; 1978: 14).

La expresión más simple de una serie estadística se compone de dos partes dispuestas generalmente en columnas:

La primera parte se integra con las categorías que tienen características

cualitativamente distintas y quedan abarcadas por la variable en estudio.

La segunda parte se forma con el registro de las frecuencias de los casos observados en cada categoría o nivel. Conviene definir claramente los límites numéricos o conceptuales de cada categoría pues éstas deben ser mutuamente excluyentes; esto significa que cada dato debe contabilizase en sólo una categoría de la serie.

Desde esta perspectiva una serie estadística resulta ser la forma más elemental de describir a una variable. Existen cuatro tipos de series estadísticas: nominales, ordinales, cronológicas o históricas y de intervalos.

2.1. Series nominales.

Las nominales constituyen las series más sencillas para ordenar una variable. Sus categorías generalmente provienen del lenguaje común y se crean dando nombre a los objetos de estudio para dividirlos en clases genéricas. Después se cuenta el número de elementos que posee cada categoría y de esta manera se le asigna su frecuencia. Sus características son:

1ª El orden de las categorías de la serie no tiene un propósito determinado.

2ª Las nominales sirven para agrupar variables cualitativas o discretas, es decir que sus elementos presentan atributos que no se miden, sino se cuentan. Por ejemplo: la nacionalidad, el color de la piel, estado civil y sexo.

3ª Como consecuencia de las características anteriores, el número de operaciones estadísticas es muy limitado. Ejemplos.

RELIGIÓN	F
Católicos	150
Protestantes	70
Mormones	40
T. de Jehová	40
Otras	8
	308

PARTIDOS POLÍTICOS	F
PRI	1200
PAN	1800
PRD	2500
PT	700
PVEM	900
No especificado	60
	7160

ESTADO CIVIL	F
Casado	1200
Soltero	800
Viudo	50
Divorciado	150
Unión libre	300
Concubinato	100
	2600

A veces se presentan datos con una mínima frecuencia y sin embargo el investigador desea representarlos en el conjunto general de la información, entonces se usan las categorías complementarias tales como Otros, Varios, Indefinido y No suficientemente especificado. Debe tratar de usarse lo menos posible este tipo de categorías o tener un mínimo de datos en ellas pues su presencia introduce un margen de incertidumbre en las interpretaciones finales.

2.2. Series ordinales.

Cuando los datos que se están trabajando tienen la característica de presentarse con mayor o menor magnitud y es posible ordenarlos de acuerdo a esta magnitud, debe usarse una serie tipo ordinal. En estos casos el enlistado de las categorías debe de hacerse de manera que refleje ese orden (LEVIN; 1979: 20-21).

Los datos que cumplen las condiciones anteriores resultan más valiosos que aquellos que forman las escalas nominales, porque el orden implicado en ellos representa una ventaja en información que se puede aprovechar en análisis posteriores. He aquí algunos ejemplos:

TIPO DE CERÁMICA

Nivel	A	B	C	D	E
I	3	0	14	5	0
II	5	1	9	10	0
III	8	9	3	6	4
IV	12	0	0	3	2
V	4	0	0	1	0

ESTRATOS SOCIALES

ESTRATO	SUBESTRATOS	F
	Alta	100
ALTA	Media	150
	Baja	250
	Alta	820
MEDIA	Media	11300
	Baja	12280
	Alta	44350
BAJA	Media	60500
	Baja	48470

ESTRATIFICACIÓN SOCIAL DE INGLATERRA EN EL AÑO DE 1688 (Basado en FLOUD; 1975:25).

Estratos	f
Señores laicos	160
Señores eclesiásticos	26
Barones	800
Caballeros	600
Escuderos	300
Hidalgos	12000
Cargos oficiales	5000
Mercaderes	2000
Clérigos menores	8000
	28886

OPINIÓN ACERCA DE LA POLÍTICA INDIGENISTA

OPINIÓN	F
Muy acertada	35
Acertada	58
Poco acertada	70
Mala	99
Muy mala	85
TOTAL	347

2.3 Series cronológicas o históricas.

Cuando los cambios de una variable se registran en unidades de tiempo ya establecidas como días, meses, años, décadas, siglos, etc. entonces se constituye una serie cronológica. Con estas series se pueden registrar los incrementos de la población por décadas, las fluctuaciones anuales en la producción maicera, los cambios en el nivel de precios al público; en fin, toda clase de variable que manifieste su fluctuación a lo largo del tiempo. Las características de las series cronológicas son las siguientes:

1ª Las categorías en este caso están formadas por períodos de tiempo y tienen un orden convencional invariable.

2ª Las series cronológicas son un tipo especial de serie ordinal pero en este caso las categorías son consideradas equidistantes y de una misma magnitud.

3ª En este tipo de series la columna de las frecuencias puede ser ocupada por otra variable analizada precisamente por los cambios que registra al paso del tiempo (Nº de habitantes, toneladas de producción, etc.,). Ejemplos:

Nacimientos

Meses	F
Enero	13
Febrero	18
Marzo	29
Abril	12
Mayo	24
Junio	15
Julio	18
Agosto	20
Septiembre	26
Octubre	26
Noviembre	22
Diciembre	19

Producción Maicera de Yucatán

Año	Tons. miles
1997	250
1998	210
1999	190
2000	170
2001	150
2002	145
2003	130

POBLACIÓN DE YUCATÁN

Año	No. de habitantes
1900	309,652
1910	339,613
1921	358,221
1930	386,096
1940	418,210
1950	516,899
1960	614,049
1970	758,355
1980	1 063,733
1990	1 363,540
2000	1 658,210

2.4 Series progresivas de intervalos.

Antes de entrar en detalles acerca de las series progresivas de intervalos es necesario hacer una precisión en torno a los tipos de variables de acuerdo a los valores que pueden tomar las mismas:

“Una variable continua, es aquella que puede tomar cualquier valor numérico dentro de

un recorrido específico. Cuando las observaciones sobre esa variable se disponen por orden de magnitud, los valores sucesivos pueden diferir en incrementos infinitesimales".

"Una variable discontinua toma solamente valores discretos. Las observaciones sobre esa variable, ordenada por magnitudes, cambian de valor únicamente por cantidades definidas" (MILLS; 1969: 65).

Para aclarar más este punto, a continuación se expone la opinión de Spiegel: Una variable que teóricamente puede tomar cualquier valor entre dos valores dados, se llama variable *continua*, si no es así, se llama variable *discreta*.

Ejemplo 1. En una familia el número de hijos puede tomar cualquiera de los siguientes valores 0, 1, 2, 3,... pero no podrán 2.5 o 3.842; es pues una variable discreta.

Ejemplo 2. La estatura de un individuo puede ser 1.68 o 1.6857 metros, dependiendo de la exactitud de la medida; es una variable continua.

"Los datos que vienen definidos por una variable discreta o continua se llaman *datos discretos* o *continuos*, respectivamente. El número de hijos en cada una de 1000 familias es un ejemplo de datos discretos, mientras que las alturas de 100 universitarios es ejemplo de datos continuos. En general, las *medidas* dan origen a datos continuos, mientras que las *enumeraciones* o *conteos* originan datos discretos" (SPIEGEL; 1978: 1-2).

Cuando ya se han recopilado los datos y se observa que las variables pueden expresarse numéricamente, por ejemplo, pesos, alturas, ingresos, etc., y además existe la necesidad de agruparlos para su manejo estadístico, entonces podemos usar la **serie progresiva de intervalos** de clase o simplemente serie de intervalos. Con este tipo de serie se puede distinguir los límites de los intervalos que forman las categorías. Esto facilita un manejo aritmético más amplio que en cualquier otro tipo de serie mencionada.

Una vez recolectados los datos se vuelve imprescindible para su tratamiento, la debida organización de ellos. Este proceso de organización de los datos entraña los siguientes pasos fundamentales:

1º. Ordenación. Consiste en colocar los datos de menor a mayor.

2º. Clasificación. Se colocan los datos mostrando la frecuencia con que se presentan, y

de esta manera distinguimos el número de valores diferentes que existen en la población o muestra. Este paso favorece también al conteo de las frecuencias para la agrupación en intervalos.

3º. Agrupación por intervalos.

a) Se determina el rango (dato mayor menos el dato menor) de la distribución.

b) Se elige el número de intervalos adecuado al número de datos. No existe un acuerdo general entre los diferentes autores para determinar el número de intervalos. Sin embargo es útil conocer las siguientes propuestas. Una sugerencia muy conocida es la de determinar el número del mínimo y el máximo de intervalos: no menos de 5 ni más de 20 (SPIEGEL; 1978: 28). Otro autor (MILLS; 1969: 50) ofrece una alternativa que se orienta a encontrar primero la amplitud del intervalo y es presentada como el procedimiento de Sturges:

“Dada una serie de N observaciones, de la que se conoce el recorrido (es decir, la diferencia entre la menor y la mayor de las observaciones) podemos conseguir un adecuado intervalo de clase i, mediante la fórmula

$$i = \frac{\text{Recorrido}}{1 + 3{,}322 \log. N}$$

la cual indica la amplitud del intervalo y, por consecuencia, determina el número de intervalos”. En esta fórmula i = C, así como recorrido es igual a rango.

Por otra parte tenemos la propuesta de Stevenson, quien recomienda entre un mínimo de 5 y un máximo de 15; a la vez enuncia una regla empírica consistente en calcular la raíz cuadrada de N y ajustarla, si es necesario a los límites de 5 y 15. Por ejemplo, para 400 observaciones, su raíz cuadrada es 20, entonces se debe usar 15 intervalos. Para N: 40, su raíz cuadrada es 6.32 que se deberá redondear a 6 ó 7 (STEVENSON; 1981: 40).

Por último queda una tercera posición que señala al investigador la posibilidad de decidir el número de intervalos basándose en su conjunto de datos y sus objetivos específicos, factores que pueden variar considerablemente de una investigación a otra

(LEVIN; 1979: 23 -24).

c) Una vez determinado el número de intervalos (N.I.), se divide el rango entre ese número y se encuentra la amplitud de cada clase. Es preferible usar una amplitud igual en todos los intervalos de cada tabla de frecuencias.

d) Para construir el sistema de intervalos, se le suma al dato menor la amplitud “C” y queda formado el primer intervalo. A este resultado se le suma de nuevo "C" y así delimitamos el segundo intervalo y así sucesivamente.

Ejemplo: Puntuaciones de una muestra registradas en la aplicación de una prueba de coeficiente intelectual (C.I.):

95	97	102	85	115	89	115	100	90	92
110	121	103	110	100	86	102	110	95	84
80	118	81	98	92	110	83	105	91	107
86	78	90	95	113	80	92	94	89	114
105	90	105	80	82	86	95	79	105	116

1º. Ordenación.

78	81	86	90	92	95	100	105	110	115
79	82	86	90	92	95	102	105	110	115
80	83	86	90	94	97	102	105	110	116
80	84	89	91	95	98	103	107	113	118
80	85	89	92	95	100	105	110	114	121

2º. Clasificación.

x	f	x	f	x	f	x	f	x	f	x	f
78	1	83	1	90	3	97	1	105	4	115	2
79	1	84	1	91	1	98	1	107	1	116	1
80	3	85	1	92	3	100	2	110	4	118	1
81	1	86	3	94	1	102	2	113	1	121	1
82	1	89	2	95	4	103	1	114	1		

3º. Agrupación.

a) Rango: Dato mayor - dato menor

121 - 78 = 43

Rango: 43

b) Número de intervalos: 5

$$\text{Amplitud: } C = \frac{\text{Rango}}{\text{N.I.}}$$

$$C = \frac{43}{5} = 8.6$$ Se redondea a 9 para facilitar la creación de los límites.

4º. Formación de intervalos.

78 + 9 = 87
87 + 9 = 96
96 + 9 = 105
105 + 9 = 114
114 + 9 = 123

Así tenemos:

Categorías	Límite superior real
78 - 87	86.9
87 - 96	95.9
96 - 105	104.9
105 - 114	113.9
114 - 123	122.9

5º. Tabla de frecuencia.

Después del último paso anterior se integra la tabla de frecuencia.

Categorías (puntuaciones)	f	fr	fa	fra	x
78 - 87	13	.26	13	.26	82.5
87 - 96	14	.28	27	.54	91.5
96 - 105	7	.14	34	.68	100.5
105 - 114	10	.20	44	.88	109.5
114 - 123	6	.12	50	1.00	118.5
	N = 50	1.00			

En la tabla se observan 5 columnas:

f = frecuencia de clase. Nos indica el número de casos que quedan comprendidos en cada uno de los intervalos.

fr = frecuencia relativa. Es el porcentaje que representa la frecuencia de clase de cada intervalo respecto al número total de frecuencias (N). Se obtiene dividiendo la frecuencia de clase entre N, y para expresarlo en porcentaje se multiplica por 100.

fa = frecuencia acumulada. Nos expresa la acumulación de frecuencias de clase, en forma progresiva, de cada intervalo. Se obtiene sumando acumulativamente las frecuencias de clase de cada nivel.

fra = frecuencia relativa acumulada. Se obtiene con el mismo procedimiento que la (fa), solamente, que expresado en términos de porcentaje.

x = marca de clase. Es el punto medio de cada intervalo y representa a todos los valores comprendidos en él. Se obtiene con la fórmula siguiente:

Ejemplo: $x = \frac{\text{Límite inferior + Límite superior}}{2}$

$x = \frac{87 + 96}{2}$

x = 91.5 (marca de clase del segundo intervalo)

La distancia que hay entre una marca de clase y otra es igual a la amplitud del intervalo.

NOTAS:

1. Cuando se divide el rango entre el número de intervalos y el resultado es un número cuya fracción es menor de 0.5 es recomendable redondear siempre al número inmediato superior, porque de esta manera el límite superior del último intervalo abarca seguramente al mayor dato de la muestra o población.

Por ejemplo: si dividimos el rango obtenido en la muestra mencionada, entre 8 tenemos:

$$C = \frac{43}{8} = 5.38$$

Si redondeamos a 6 Si redondeamos a 5

Intervalos	Intervalos
76 - 84	78 – 83
84 - 90	83 – 88
90 -96	88 – 93
96 - 102	93 – 98
102 - 108	98 – 103
108 - 114	103 – 108
114 - 120	108 – 113
120 - 126	113 – 118

Se puede observar que 118 deja afuera el dato 121 porque se pierde 38 centésimas de unidad en cada intervalo formado.

2. Existe otra manera muy común de presentar un sistema de intervalos:

Límites simbólicos	Límites reales
78 – 86	77.5 – 86.5
87 – 95	86.5 – 95.5
96 – 104	95.5 - 104.5
105 – 113	104.5 - 113.5
114 – 122	113.5 - 122.5

Cuando hay datos con valores como 95.5 ó 113.5 se presenta un problema de tabulación, ya que no pueden estar en dos intervalos si no se definen los límites entre 95.5 y 95.6 ó 113.5 y 113.6. Al definir los límites se observa que si el dato tiene centésimas superiores a 0.55, entonces pertenece al siguiente intervalo.

Al sistema de límites de la nota 1 se le llama "**abierto del límite superior**" y al de la nota 2, "**abierto de ambos lados**".

3. Cuando se trabaja con una variable continua, el sistema de intervalos más conveniente es el de los "límites superiores abiertos" porque se pueden agrupar los datos sin ningún problema, aunque se trate de números con fracciones cercanas a los linderos.

Por ejemplo: en X rama de la producción, 24 fábricas tienen la siguiente producción en toneladas de mercancía:

Toneladas	Fábricas
15 - 20	7
20 - 25	4
25 - 30	8
30 - 35	3
35 - 40	2
	24

Ahora bien, si entre los datos encontramos fábricas con una producción de 29.91, 20, 30.51 toneladas, no tendríamos duda alguna sobre la categoría que les corresponde, si recordamos los límites reales de cada intervalo del sistema de límites superiores abiertos. En consecuencia, el dato 29.91 corresponde a la categoría 25 - 30, y los datos, 30 y 30.51 a la categoría 30 - 35.

En cambio, si se trata de variables discontinuas, el sistema de los límites abiertos resulta más apropiado, porque sabemos que no se presentan valores fraccionarios. Además, las categorías resultan más claras y explícitas. Por ejemplo:

Número de personas por casa	f
0 - 2	6
3 - 5	8
6 - 8	10
9 - 11	9
12 - 14	10
	43

4. La distribución de frecuencias es una colección organizada de datos que tiene como finalidad mostrar de manera resumida la distribución de los valores de una variable a lo largo de las distintas categorías o intervalos que abarca la variable misma. Por esta razón presentamos los criterios más comunes observados en la mayoría de los autores con respecto a la presentación de distribución de frecuencias:

I. Es mejor presentar los datos agrupados que sin agrupar.
II. Los intervalos deben ser del mismo tamaño, por principio de regularidad y evitar los intervalos abiertos.
III. La amplitud de los intervalos debe ser en números enteros o redondeados; deben evitarse los decimales en el tamaño de C y los límites de las categorías (BLALOCK; 1978: 54).
IV. Debe prevalecer el sentido común en cuanto al número de intervalos; lo que se puede interpretar en la posibilidad de decidirse por un número de intervalos o una amplitud de intervalo en función de los objetivos de la investigación, independientemente de las fórmulas propuestas.
V. Intervalos especiales:
a) Los intervalos para los grupos de edad en las pirámides de población se forman con una amplitud de 5 años universalmente.
b) Si los intervalos registran calificaciones de escolaridad y su escala es de 0 a 100 conviene adaptar los intervalos a decenas y evitar que el límite superior de la última categoría sea mayor que 100 (RASCON; 1983: 83 - 86).

Ejercicio. Agrupa la siguiente muestra de 60 salarios mensuales de 1986 (en miles de pesos) en una serie progresiva de intervalos de las dos maneras siguientes:
Ensayando la fórmula de Sturges y utilizando 5 intervalos iguales.

40.3	55.8	80.5	63.8	77.2	50.2
35.6	55.4	63.7	92.4	65.9	58.1
30.4	87.4	97.3	95.7	39.7	88.4
28.9	70.7	90.4	66.2	64.4	99.0
90.2	60.2	25.7	35.5	80.7	79.9
50.5	72.4	85.2	43.3	85.0	80.6
68.1	69.5	59.0	91.0	42.2	22.3
70.7	71.5	50.0	37.6	37.7	64.2
60.4	76.0	78.4	26.5	53.8	15.6
65.6	95.8	55.7	68.3	46.5	85.2

3. GRÁFICAS.

Las series estadísticas pueden ser representadas gráficamente en tal forma que nos indiquen las tendencias o variaciones de una población o muestra. Los tipos de gráficas más comunes son: las de barras, las de líneas y las circulares.

3.1 Gráficas de barras.

Entre las gráficas de barras destacan el histograma: se construye levantando líneas verticales en los límites reales de los intervalos de clase, formándose series de rectángulos. La altura de las barras está determinada por la frecuencia de clase o relativa de cada categoría (Gráfica 1).

Si las series son nominales u ordinales, los rectángulos pueden estar separados, pero si quieren representar series cronológicas o de intervalos, las barras deben ser contiguas. Las gráficas de barras se usan de preferencia:

- cuando sea muy pequeño el número de las categorías (2 a 5) contenidas en la serie estadística.
- cuando interese observar que porcentaje representa la frecuencia de cada categoría respecto al total.
- cuando la variable estudiada no exija un tipo especial de representación gráfica, ya que la de barras es muy fácil de comprender.

3.2 Gráficas de líneas.

Dentro de la categoría de gráficas lineales, el polígono de frecuencias es el más usado. Se obtiene uniendo los puntos que forman las intersecciones de las líneas que indican las frecuencias de clase o frecuencias relativas, de cada categoría (Gráfica 2). Las gráficas de línea se usan:

- cuando se necesita observar la tendencia de los cambios continuos de cada categoría;
- para representar datos clasificados en tipos de series cronológicas y de intervalos.
- cuando se necesita comparar con claridad los cambios de dos o más distribuciones de frecuencias.

3.3 Gráficas circulares.

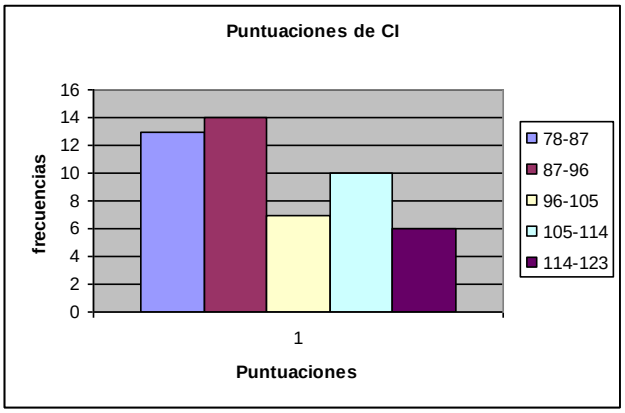
Es la que toma como base de representación el círculo y son mejor conocidas como las “gráficas de pastel” (Gráfica 3). Este tipo de gráficos resulta muy apropiado para observar la proporción (%) que tiene la frecuencia de cada categoría en el contexto de la distribución general. Su uso se ha generalizado porque permite comparar con facilidad los distintos resultados que provengan de una misma población. Por el contrario, su uso no es muy apropiado para representar series cronológicas ni series en donde existan muchas categorías con poca frecuencia.

Ejemplo: con la siguiente tabla de frecuencias mostraremos los tres tipos de gráficas:

Puntuaciones	f	fr	fa	fra	x
78 - 87	13	.26	13	.26	82.5
87 - 96	14	.28	27	.54	91.5
96 - 105	7	.14	34	.68	100.5
105 - 114	10	.20	44	.88	109.5
114 - 123	6	.12	50	1.00	118.5
	N = 50				

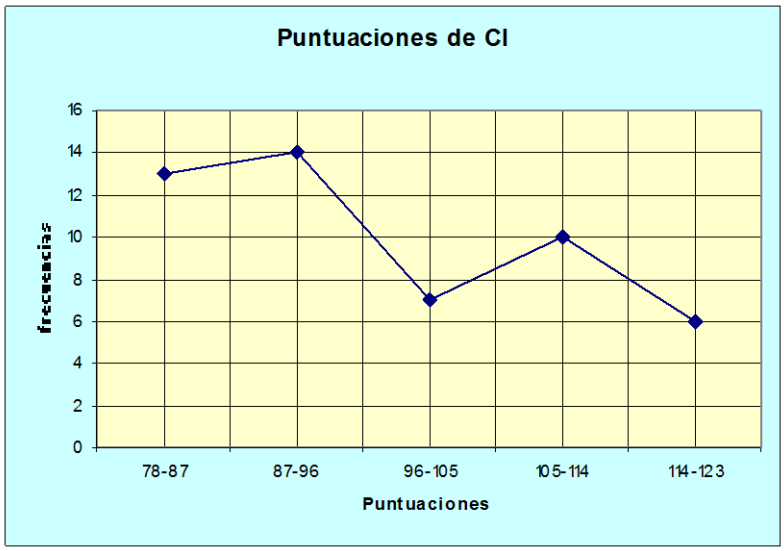
GRAFICA 1

Puntuaciones	f
78-87	13
87-96	14
96-105	7
105-114	10
114-123	6



GRÁFICA 2

Puntuaciones	f
78-87	13
87-96	14
96-105	7
105-114	10
114-123	6

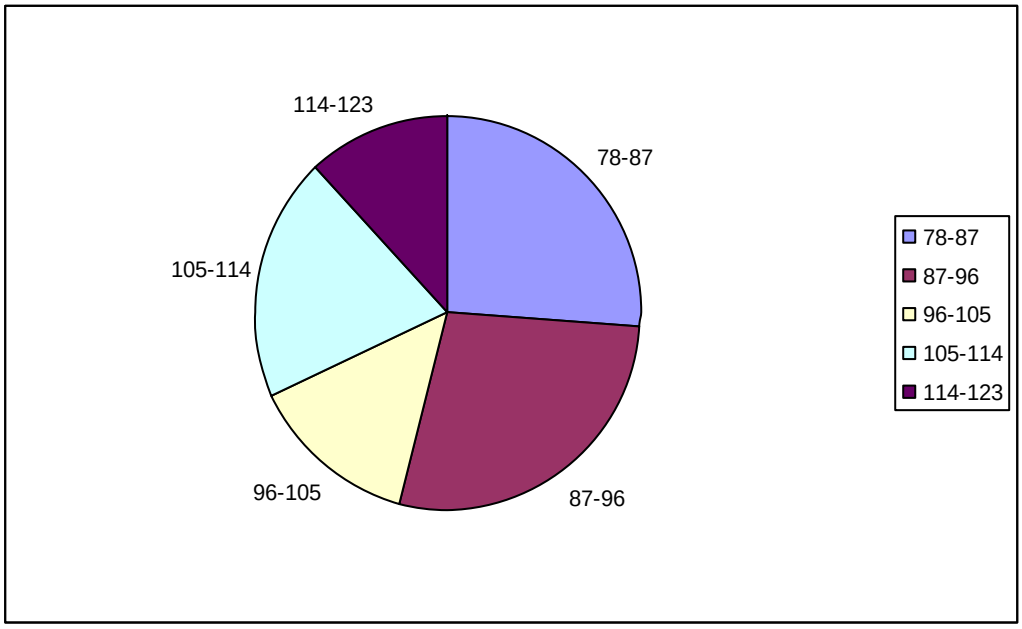


GRÁFICA 3

Puntuaciones	f
78-87	13
87-96	14
96-105	7
105-114	10
114-123	6

Tabla para la construcción de la Gráfica Circular

Intervalos	f	fr	frx360°	Redondeo	acumulado
78-87	13	0.26	93.6	94	94
87-96	14	0.28	100.8	101	195
96-105	7	0.14	50.4	50	245
105-114	10	0.2	72	72	317
114-123	6	0.12	43.2	43	360
	50	1	360	360	



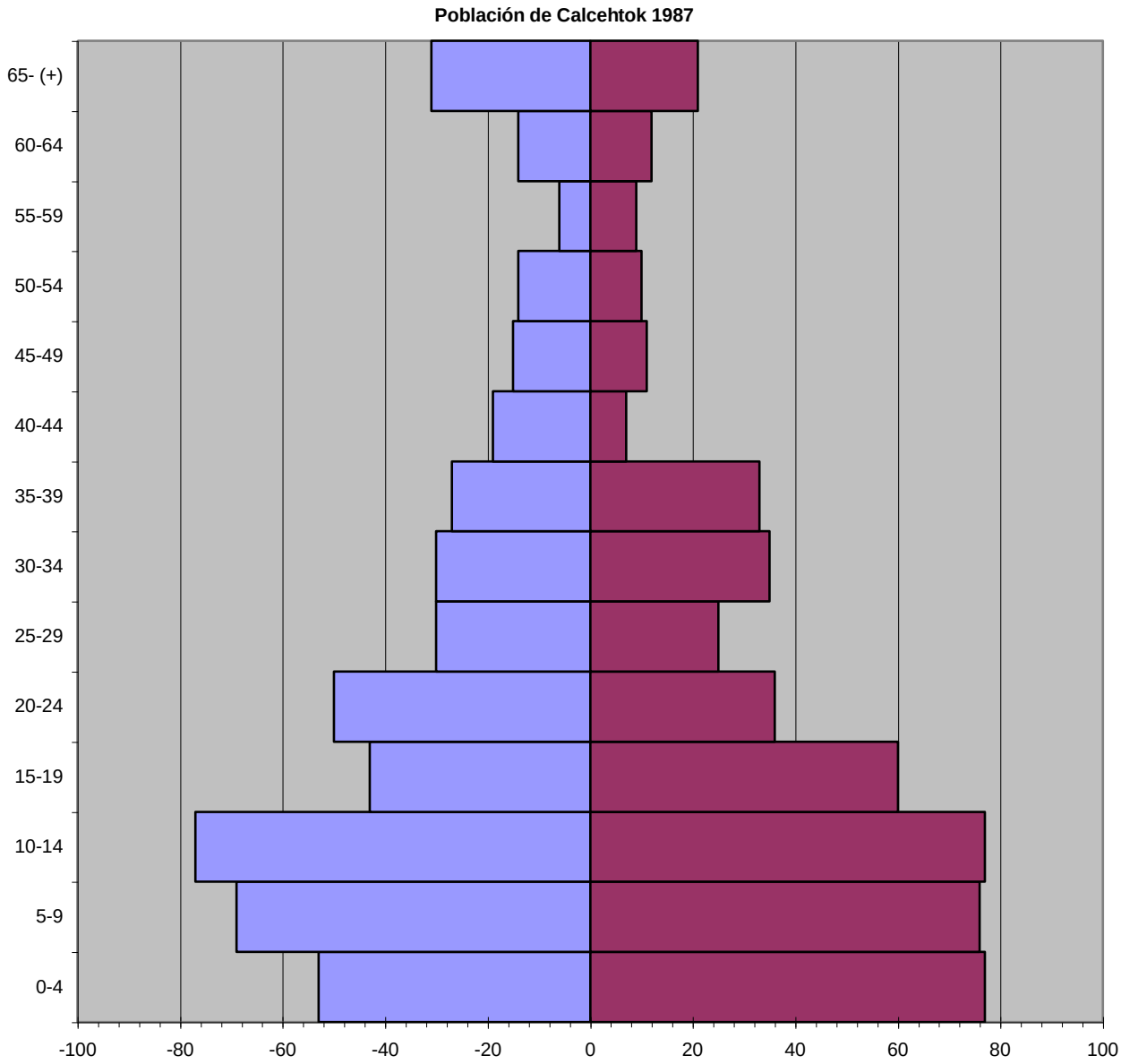
3.4 Pirámide de Población

Una de las aplicaciones más útiles en las ciencias sociales de las gráficas es la que representa la pirámide de población, la cual permite conocer la estructura por edad y sexo de una población específica y compararla con otras que representen parámetros generales. Básicamente la pirámide de población son dos gráficas de barras con un solo sistema de intervalos que señalan los grupos de edad con su frecuencia correspondiente para hombres y mujeres.

Población de Calcehtok por grupos de edad y sexo.

1987

Edad	hombres	Mujeres	total
0-4	53	77	130
5-9	69	76	145
10-14	77	77	154
15-19	43	60	103
20-24	50	36	86
25-29	30	25	55
30-34	30	35	65
35-39	27	33	60
40-44	19	7	26
45-49	15	11	26
50-54	14	10	24
55-59	6	9	15
60-64	14	12	26
65- (+)	31	21	52
Total	478	489	967


Población de Calcehtok 1987
65- (+)
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
-100
-80
-60
-40
-20
0
20
40
60
80
100

CAPITULO III

MEDIDAS DE CENTRALIZACIÓN

1. SU SIGNIFICADO.

En el capítulo anterior destacamos las ventajas de contar con series estadísticas que agrupan y describen los conjuntos de datos. Sin embargo, en muchas ocasiones estas series no son suficientes para responder más interrogantes sobre la misma población.

Más allá de las series estadísticas se necesita, para el análisis de tales poblaciones ciertas medidas que resuman las características más importantes de los conjuntos de datos. Por ejemplo: uno de los indicadores más significativos del grado de desarrollo de todos los países es el nivel educativo. Éste se aproxima midiendo el aprovechamiento escolar, especialmente de los niños. Si fuera necesario caracterizar a toda esta población, de poco serviría tener los millones de puntajes de los escolares a la mano agrupados en una serie estadística.

En tales situaciones se recurre a las medidas de centralización o tendencia central. Las medidas de centralización son parámetros que, dentro de una distribución de frecuencias generado por una población o muestra, tratan de representar al conjunto en una cifra típica.

La representación mediante un promedio de una distribución de datos se justifica por la tendencia de las grandes masas de datos de concentrarse alrededor de un valor central, del cual se desvían con más o menos regularidad todos los casos observados.

Ejemplo. ¿Cuál es el promedio de aprovechamiento escolar en los niños que estudian la educación primaria en México? El **promedio** de aprovechamiento de los niños que estudian la primaria en México es de **58.49 puntos** (Secretaría de Educación Pública en Mundo al Día, 22/10/2001).

Generalmente la necesidad de conocer esa cifra típica de las distribuciones se suscita cuando se tienen que comparar dos o más conjuntos de datos. Entonces los promedios facilitan esta tarea porque en lugar de comparar cientos o miles de datos

sólo se comparan dos o varias números ya que sólo habrá que tomar uno por cada población. En la medida que las distribuciones sean más semejantes en cuanto a tamaño y dispersión el resultado de la comparación tendrá un significado claro. Cuando las muestras o poblaciones comparadas entrañan mayores diferencias de tamaño y dispersión habrá que ser cuidadosos en las interpretaciones que se le dé a los resultados, pues como todos los instrumentos estadísticos, las medidas de centralización tienen sus limitaciones.

Si los valores individuales que forman parte de una distribución varían diametralmente y no muestran tendencia alguna hacia la centralización, entonces ningún promedio puede representarlos. Por ejemplo, si tomamos las magnitudes de población de tres lugares de México, 350, 15,000 y 3’000,000 y obtenemos la media aritmética de estas cantidades, 1’005,119.6, se observa que este valor no representa a ninguna de las tres poblaciones.

Los estudiosos de la Estadística han logrado muchas medidas de centralización, pero las principales son: la media aritmética, la moda y la mediana.

2. MEDIA ARITMÉTICA (X).

La media aritmética es la medida de centralización más conocida y la más utilizada como promedio en general. Su definición se enuncia de la siguiente manera: es la suma de todos los valores de un conjunto de datos dividida entre el número total de datos.

$$X = \frac{\sum x}{N}$$

Propiedades:

Primera. La media es el punto de equilibrio de una distribución determinada. Esta propiedad se puede comprobar sumando algebraicamente las distancias que los datos tienen respecto a la media; esta suma debe ser igual a cero. Esta propiedad está representada por la siguiente fórmula:

$$\sum (x - X) = 0$$

Segunda. La suma de valores de los elementos es igual a la media multiplicada por el número de elementos. Esta propiedad está representada por la siguiente fórmula:

$$\sum x = XN$$

Características.

Primera. El valor de la media aritmética está determinado por cada uno de los valores de la serie. Cuando la distribución de los datos se ajusta a la regularidad que se expresa con la curva normal o campana de Gauss, la media aritmética es el mejor promedio para representar a dicha distribución.

Pero si la distribución es asimétrica la medía aritmética puede dejar de ser la que mejor tipifique al conjunto de datos. En tanto más asimétrica, más inadecuada será.

Segunda. Dos distribuciones de frecuencias con diferente dispersión pueden acusar una misma media. Esta característica deberá ser tomada en cuenta cuando se usan medios aritméticos como puntos de comparación.

3. MODA (MO).

La moda es el dato u observación que se presenta un mayor número de veces. El cálculo de esta medida es muy fácil en datos sin agrupar, porque para hallarla, solamente es necesario mirar una clasificación de datos y describir cual aparece con mayor frecuencia.

Propiedades.

En el sentido matemático estricto, la moda no tiene propiedad alguna. En sentido general, se puede afirmar que las propiedades principales de la moda son:

Primera. La facilidad de sus cálculos cuando los datos no están agrupados y,

Segunda. La utilidad que tiene en los datos y series nominales. Por ejemplo al estudiar los artículos artesanales de dos lugares distintos pero de similar producción, al uso de la moda sería más adecuado a estas dos series nominales comparadas.

Características.

Primera. El valor de la moda no está afectado por la magnitud de los valores.

Segunda. En una distribución puede haber dos o más modas; lo que le resta significado a esta medida como medida de centralización.

Tercera. La utilidad de la moda es mayor a medida que aumenta la tendencia de aparición de un valor. Esto significa que en las distribuciones asimétricas la moda funciona mejor que la media aritmética

4. MEDIANA (MD). Es el valor del dato calculado que divide a una distribución dada en dos partes iguales de tal manera que el 50% del número total de casos, dispuestos en orden de su magnitud, se encuentre por debajo del dicho dato y el 50% restante se encuentre por encima.

a) Fórmula para calcular la mediana de una serie de datos no agrupados; se procede a ordenarlo primero de acuerdo con la magnitud del valor de cada dato ascendente o descendente. Después se selecciona el dato que quede en medio de todos. Si el número de datos es impar, siempre encontraremos el valor mediano a simple vista, por ejemplo, la mediana de 6, 7, 7, 8, 9, 11, 12. El valor mediano es 8. Ahora bien si el número de datos es par, se encontrará la mediana a través de un promedio aritmético de los dos datos centrales; por ejemplo: 25, 39, 28, 72, 56, 52, 35, 40, que ordenados quedan 25, 28, 35, 39, 40, 52, 56, 72. La mediana en este caso es igual a: 39 + 40

Características.

Primera. La mediana no está afectada por los valores extremos de una serie de datos. Por ejemplo, la mediana de 26, 28, 35, 37, 42 es la misma que la 26, 27, 35, 38, 60.

Segunda. La mediana puede calcularse aunque no se conozca todos los datos, con la condición de que se conozca su frecuencia y la situación general de todos los datos y se disponga de una información precisa acerca de aquellos que estén próximos al centro de la serie.

Tercera. Para calcular el valor de la mediana los datos deben ser ordenados en forma creciente o decreciente, atendiendo a la magnitud los valores de los datos.

5. ¿CUÁL USAR?

Frecuentemente los investigadores sociales se preguntan cuál es la medida de centralización óptima para sus datos. La respuesta a esta pregunta debe basarse con los siguientes criterios:

Las necesidades del investigador.

Las características de su muestra o población.

En cuanto al primer criterio podemos señalar que el investigador puede estar sujeto a normas establecidas de comparación. Por ejemplo si vamos a comparar el grado de escolaridad de los niños de una comunidad x con un parámetro estatal tenemos que observar la concordancia entre las técnicas utilizadas para medir esa variable. Si el parámetro se obtuvo con una media aritmética, el de la comunidad tendrá que ser obtenido con la misma herramienta estadística. El no observar esta precaución le restaría validez al resultado de esta comparación y la seriedad de la investigación ya que el significado de las medidas involucradas es distinto.

En cuanto al segundo criterio, las características de la muestra o población, debería ser la causa principal para utilizar una medida de centralización determinada. En términos generales, se sugiere lo siguiente:

Si una muestra está distribuida con regularidad, es decir, similar a la curva normal, cualquiera de las tres medidas pueden calificarla (X, Mo, MD) aunque es preferible la media aritmética porque su naturaleza matemática le permite un mayor número de manejos estadísticos con un significado muy claro y válido.

Si la muestra presenta mucha dispersión y ningún dato o intervalo tiende a destacar por su frecuencia quizá la mediana sea la medida justa.

Si la muestra dispersa, pero se puede advertir la presencia de algunos datos extremos y la concentración de los mismos en un determinado punto, entonces la moda puede presentar con más validez a ese grupo de datos.

6. MEDIA ARMÓNICA.

Es un tipo de promedio cuya aplicación sólo resulta adecuada en un campo restringido, pero que deberá emplearse para evitar errores al manejar determinado tipo de datos. Debe utilizarse al promediar tasas de tiempo y tiene sus ventajas precisas en el manejo de cierta clase de datos sobre precias.

Supongamos que un artículo determinado se cotiza a distinto precio en tres tiendas diferentes: a “cuatro unidades por dólar”, a “cinco unidades por dólar” y a “veinte unidades por dólar”. Se desea conocer el precio medio por unidad. El promedio aritmético de las cifras (4, 5 y 20) es 9.66; si lo aceptamos como el número medio vendido por dólar, el precio medio en centavos sería de 100 entre 9.66 = 10.35 centavos por unidad.

Las cotizaciones originales son equivalentes a precios unitarios por centavos de 25, 20 y 5. El promedio de estos precios es de 16.66 por unidad. La discrepancia entre 10.35 y 16.66 centavos se debe al uso incorrecto de la media aritmética para promediar cotizaciones en la forma de “tantos por dólar”. Tal media es, en realidad, un promedio ponderado en el que se da mayor peso a las cotizaciones que incluyen un mayor número de unidades del artículo.

Puede obtenerse el resultado correcto tomando la media armónica de las tres cotizaciones originales.

Definición.

“La media armónica de una serie de números es el recíproco de la media aritmética de los recíprocos de los números individuales”.

La media armónica de 4, 5 y 20 es de 6 unidades por dólar y el precio medio en centavos es 16.66 (100/6 = 16.66).

Ejemplos:

Un grupo de trabajadores construye los primeros 120 metros de una avenida con una productividad de 12 metros diarios; en cambio, los siguientes 120 metros, lo hace a razón de 18 metros por día. Se trata de saber cual es la productividad diaria durante todo el trabajo. Si se decidiera por la media aritmética sólo habría que sumar los dos valores de la productividad y dividir entre 2. El resultado sería 15 metros diarios.

Por otra parte, los primeros 120 metros requieren 10 días de trabajo y para los siguientes 120 metros sólo se tardan 6.67 días; es decir, toda la avenida la harían en 16.67 días. Si la productividad promedio es de 15 m/día en los 16.67 días se construirían (15X16.67) 250 metros. Resultado inconsistente y falso ya que la obra es sólo de 240 metros.

En este caso, se obtiene la media armónica, 14.40 m/día y se multiplica por 16.67, dando como resultado los 240 metros.

Características.

Primera. La media armónica se usa en la obtención de promedios en los casos donde los datos tienen unidades que implican dos tipos de variables. Ejemplos:

productividad: cantidad de trabajo / cantidad de tiempo, densidad de población: número de habitantes por kilómetro cuadrado.

Segunda. La media armónica no tiene las propiedades algebraicas de la media aritmética (MILLS; 1969:119).

¥ § £ & # € ± ¶ ₸ Ж

Ш ф ђ Ѳ Я ¥ д Њ

CAPÍTULO IV

MEDIDAS DE VARIACIÓN

1. INTRODUCCIÓN.

Cuando se hacen análisis de una o más distribuciones de frecuencias, se nota que los promedios, por sí mismos, no cubren todas las necesidades para la descripción y comparación de los datos.

En ocasiones, dos grupos de datos pueden tener promedios aproximados lo que nos haría creer que las distribuciones son semejantes. Sin embargo, los promedios no nos indican la variación interna de los valores; y para ciertos fines esto puede ser muy importante, además de que el conocer la magnitud de la dispersión de los datos enriquece el conocimiento de las muestras o poblaciones comparadas. Por ejemplo: dos países pueden mostrar un promedio de ingresos por persona muy semejante, pero al analizar la dispersión de la distribución de los ingresos, se encuentra que son diferentes. Mientras que en un país hay una distribución de la riqueza aceptable, en el otro hay una inequidad.

Para superar esta deficiencia en la descripción, se utilizan las medidas de variación. Éstas constituyen los instrumentos que la estadística usa para conocer la magnitud de la dispersión que los datos tienen.

La cualidad principal de las medidas de variación es mostrar en una cifra única las diferencias de concentración de los datos de una distribución de frecuencias. Como consecuencia inmediata podemos describir y comparar mejor a las muestras o poblaciones que estamos estudiando.

A lo largo de la historia de la Estadística se han desarrollado múltiples formas de medir la concentración o dispersión de los valores que componen los conjuntos de datos. En este Capítulo se tratarán cuatro medidas principales de variación: el rango, la desviación media, la desviación típica y el coeficiente de variación.

2. RANGO

El rango es la medida de variación que se obtiene con mayor facilidad. Es muy frecuente su uso en el lenguaje coloquial para dar una idea de la magnitud de la variación de un conjunto de datos, precisamente por la sencillez de su cálculo.

El rango se define como la diferencia entre el dato de mayor valor y el de menor valor; es decir, el rango se obtiene con una simple sustracción, siempre que se tengan los valores disponibles.

Para hallar el rango se usa esta simple fórmula

Rango: Dato mayor – dato menor

Características

Primera. Es la medida de variación más fácil de calcular, su significado es muy claro.

Segunda. El rango está determinado solamente por dos valores extremos, lo que le da instabilidad pues un cambio significativo en cualquiera de ellos puede afectar la magnitud de esta medida de variación.

Tercera. El rango no puede ser sometido a cálculos matemáticos más complejos.

3. DESVIACIÓN MEDIA

Antes de abordar la desviación media, es oportuno señalar que, entre las medidas de variación que estamos tratando, hay dos que se vinculan con la media aritmética (**X**): la desviación media (**Dm**) y la desviación típica (**s**). Su relación con la media aritmética se basa en el hecho de tomar a esta medida como su punto de referencia para obtener la dispersión de los datos individuales. Sin embargo, en algunas ocasiones puede usarse la mediana para ejercer la misma función; en este caso ya no le llamaríamos desviación media sino desviación mediana (YAMANE; 1977: 38).

El aspecto principal de las dos medidas mencionadas, es que toman en cuenta todos los datos para encontrar su valor.

La desviación media se define como el **promedio aritmético de las distancias de los datos con respecto a la media aritmética**. Es decir, a cada dato se le resta la magnitud de la media aritmética y luego se suman todos los datos tomando su valor absoluto
(| **x** |) esto es, ignorando su signo resultante. La sumatoria obtenida se divide entre el número de datos y el resultado es la desviación media.

$$DM = \frac{|\sum x - X|}{N}$$

La razón por la que se ignoran los signos resultantes de cada resta está explicada por la primera propiedad de la media aritmética: “La media es el punto de equilibrio de una distribución determinada. Esta propiedad se puede comprobar sumando algebraicamente las distancias que los datos tienen respecto a la media; esta suma debe ser igual a cero”. Entonces, si no se eliminaran los signos de los casos negativos, no podría encontrarse la medida de variación.

4. DESVIACIÓN TÍPICA

La segunda medida de dispersión relacionada con la media aritmética es la desviación típica. Se define como la raíz cuadrada de la media aritmética de los cuadrados de las diferencias de los datos respecto a su media aritmética.

$$S = \sqrt{\frac{\sum (x - X)2}{N}}$$

De la anterior definición se desprende que la diferencia fundamental entre la (**Dm**) y la (**s**) y es que ésta si toma en cuenta los signos que tienen las desviaciones de los datos con respecto a la media aritmética y los elimina mediante el **manejo algebraico** que

indica su fórmula. En cambio, la desviación media los elimina justificadamente pero en forma arbitraria

En el ámbito de la investigación se usa preferentemente la desviación típica, muy conocida también como la desviación estándar. Su estado de perfeccionamiento matemático permite incluirla en muchas fórmulas para cálculos más complejos como pruebas de significación y de hipótesis

5. COEFICIENTE DE VARIACIÓN

El coeficiente de variación es una medida de variación expresada en porcentaje.

Cuando existe la necesidad de comparar la variación de los valores en dos conjuntos de datos, el investigador puede encontrarse dos tipos de problemas:

Primero. Que los datos comparados tengan valores muy diferenciados en magnitud. Por ejemplo, si necesitamos comparar los datos de la captura pesquera en Yucatán con los datos del mismo rubro pero a nivel nacional encontraremos que la desviación estándar de la captura a nivel nacional es necesariamente mayor si que esto implique realmente una mayor variabilidad de los datos.

Variación de captura pesquera.

	Yucatán	Nacional
X Media	10 toneladas	2000 toneladas
S Des. Típica	2.8 toneladas	320 toneladas

Segundo. Si el conjunto de datos comparados tienen unidades distintas entre si por pertenecer a otro tipo de especie, no va ser posible una comparación racional. Por ejemplo, si queremos comparar la variación de la captura pesquera con la del número de pescadores que salen cada día al mar, no solo tendremos la dificultad de las distintas escalas de magnitud sino que también la diferencia de unidades de cada variable.

	Captura	Pescadores
X Media	10 toneladas	500 hombres
S Des. Típica.	2.8 toneladas	28 hombres

Para este tipo de comparaciones se recurre a una medida de variación relativa. Esto significa que se transforman las unidades originales en cifras que indican porcentajes. Esto se logra dividiendo la desviación típica entre la media aritmética para que nos exprese que porcentaje es la medida de dispersión con respecto a su medida de centralización. Así tenemos que el coeficiente de variación es la proporción entre la desviación típica y la media aritmética, expresada en términos de porcentaje

$$C\,V = \frac{s}{X} \times 100$$

Si aplicamos esta fórmula a los ejemplos mencionados anteriormente tenemos los resultados siguientes:

Captura pesquera

Yucatán	Nacional
CV = (2.8/10)100 = 28 %	CV = (320/2000)100 = 16 %

La captura pesquera local varía más que la nacional

Captura pesquera y fuerza de trabajo

Toneladas	Hombres
CV = (2.8/10)100 = 28 %	CV = (30/500)100 = 6 %

La captura en toneladas varía más que el número de hombres que salen al mar

CAPÍTULO V

CORRELACIÓN

Con frecuencia, el investigador desea conocer cual es la relación que existe entre dos variables, mas que medir cada una de ellas en forma independiente. Para estos casos se usa el coeficiente de correlación de Pearson (r). Esta medida existe como un índice estadístico por la necesidad de **medir el grado de relación de dos variables** cuantificables ejemplos de éstas serían: temperatura, longitud, salario, coeficiente intelectual, productividad, etc.

La escala para medir la correlación fluctúa entre (-1) y 1, es decir, son los valores límite para el coeficiente (r). Se interpreta que los resultados cuyo valor final que se ubiquen entre 0 y 1 son variables cuya relación es positiva y los que se ubiquen entre 0 y –1 exhibirán una relación negativa. Sin embargo, la fuerza de la correlación entre dos variables necesita ser explicada de manera más específica ya que no todos los valores del coeficiente, positivos o negativos, son necesariamente significativos.

Por esta razón es conveniente señalar de una vez que los valores (-0.5) y (0.5) agregan elementos de referencia para una mejor interpretación de esta medida.

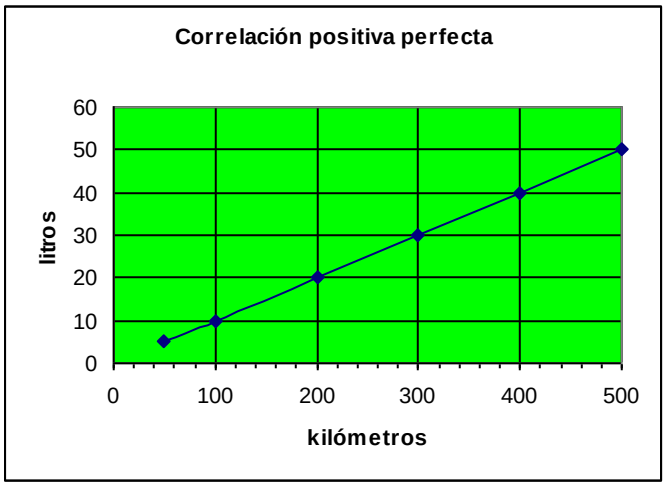
-1_____-0.5_____0_____0.5_____1

Dentro del espacio a que corresponden las **correlaciones positivas** se presentan tres momentos importantes para la caracterización de los valores de (r):

Correlación = 1: es una **correlación positiva perfecta** entre dos variables e indica una influencia directa y proporcional de la variable independiente (X) sobre la dependiente (Y).

Ejemplos:

1.- De acuerdo con la distancia en kilómetros (X) que recorra un vehículo automotor será la cantidad en litros (Y) que consuma.

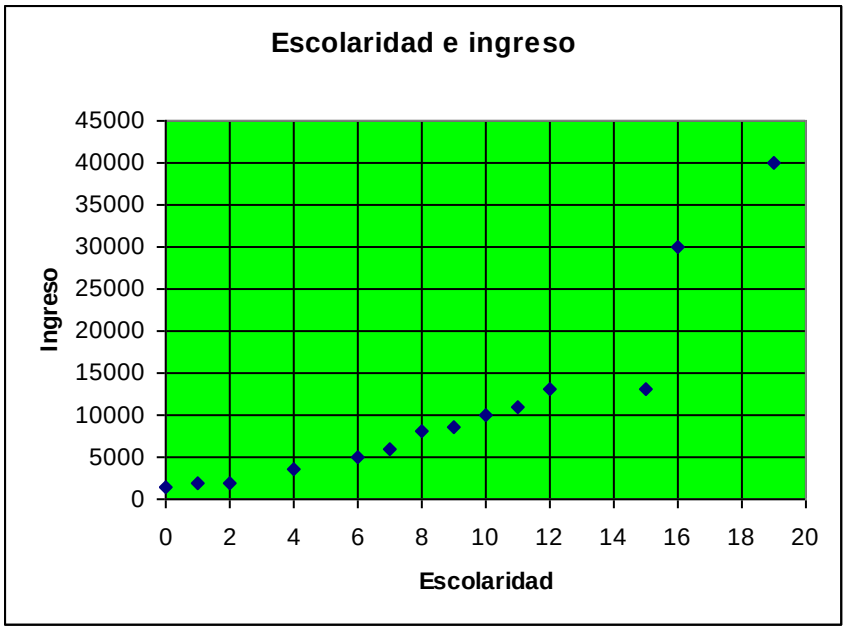


2.- El impuesto gravado (Y) a los contribuyentes estará determinado por los ingresos (X) que obtengan.

Correlación = menor de 1 y mayor de 0.5: se trata de una correlación positiva entre dos variables que debe interpretarse como una influencia directa, no proporcional sino a muestra de tendencia de la variable independiente (X) sobre la variable dependiente (Y).

Ejemplos:

1.- Los individuos que tiene una mayor escolaridad (X) registrada en número de años terminados tendrán mayor posibilidad de obtener mejores ingresos (Y) por su actividad laboral.



2.- El ingreso económico (X) que logran los padres de familia pueden estar muy relacionados con el aprovechamiento escolar (Y) que sus hijos registren en su educación.

En los dos ejemplos anteriores se puede observar que la variable X parece influir sobre la variable Y pero a manera de tendencia y no la determina proporcionalmente como en el caso de las correlaciones positivas perfectas.

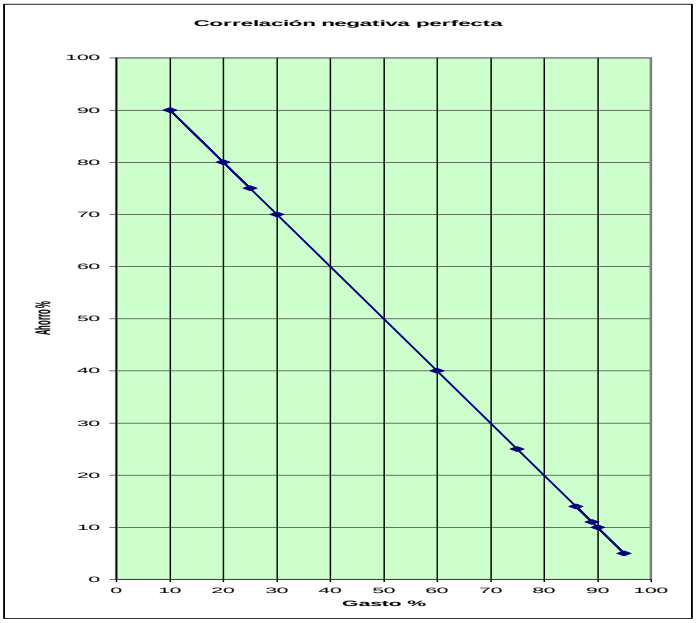
Correlación = menor de 0.5 y mayor que 0: en estos casos debemos rechazar el planteamiento que hace la hipótesis pues el valor de la relación es bajo y debe entenderse que la variable independiente (X) no influye a la variable dependiente (Y). Peor aún si el valor resultante es cercano al cero.

Para las **correlaciones negativas** estos tres momentos se presentaría de manera equivalente, pero en sentido opuesto. Veamos:

Correlación = -1: se le denomina **correlación negativa perfecta** y denota la influencia inversa y proporcional de la variable independiente (X) sobre la variable dependiente (Y).
Ejemplos:
1.- La velocidad (X) que adquiera un cuerpo al recorrer una distancia es determinante para la cantidad de tiempo (Y) que invierta en hacerlo.

2.- Si dividimos todo el ingreso económico de un conjunto de personas en gasto y ahorro se puede observar que a un porcentaje alto de gasto (X) se asociará un porcentaje bajo de ahorro (Y).

Sujeto	Gasto %	Ahorro %
A	90	10
B	95	5
C	86	14
D	89	11
E	75	25
F	60	40
G	30	70
H	20	80
I	10	90
J	25	75

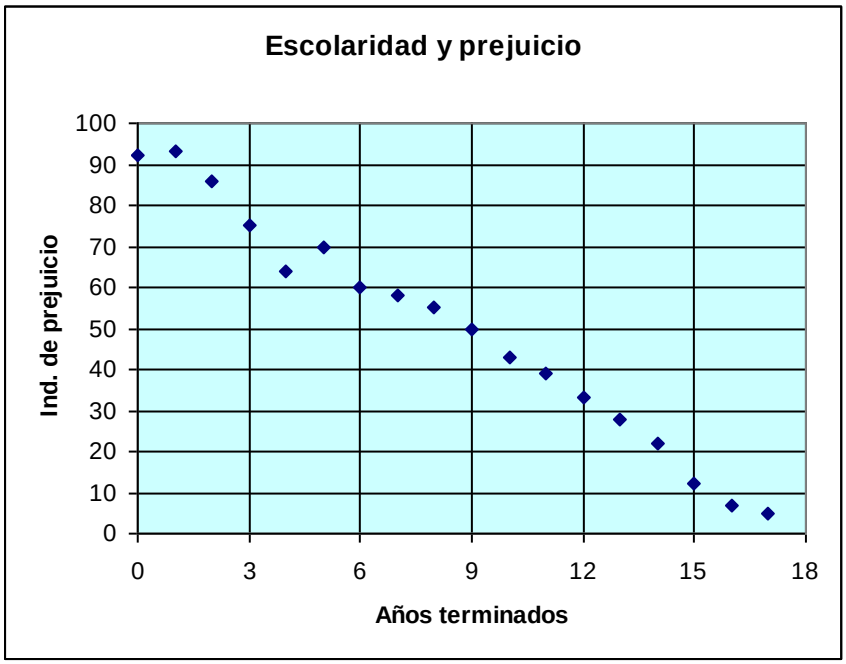


Correlación = mayor de –1 y menor de - 0.5: se considera como una correlación negativa fuerte que debe ser interpretada como una influencia de efecto inverso, no proporcional sino a manera de tendencia de la variable independiente (X) sobre la dependiente (Y).

Ejemplos:

1.- Cuando aumenta el índice de precios de una ciudad (X) los habitantes tienden a disminuir su consumo cotidiano de bienes y servicios (Y).

2.- Las personas que tienen una mayor escolaridad (X), registrada en años terminados, tienden a disminuir sus prejuicios (índices) sobre el control de natalidad (Y).



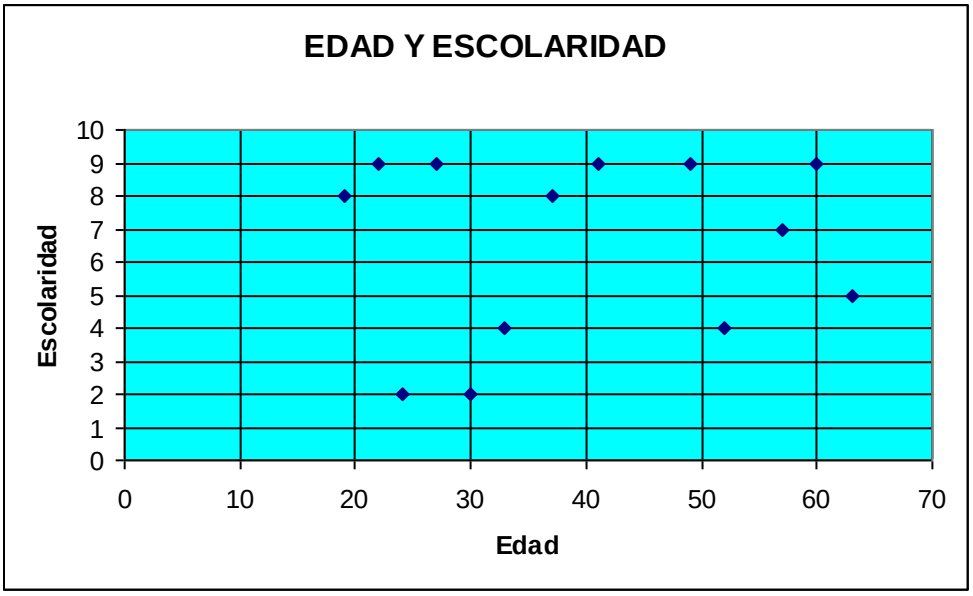
3.- En las zonas donde se han difundido más las campañas de vacunación (X) se presentarán menos casos de incidencia (Y) de la enfermedad que se trate.

Las correlaciones negativas fuertes pueden interpretarse como una influencia significativa de (X) sobre (Y) pero de no de manera inversamente proporcional.

Correlación = mayor que –0.5 y menor que 0: indica una **relación débil** o una falta de relación entre las variables señaladas en la hipótesis por lo que ésta debe rechazarse pues la variable independiente (X) no está afectando consistentemente a la variable dependiente (Y). Esta situación se acentúa si el valor del coeficiente de correlación es muy cercano al cero.

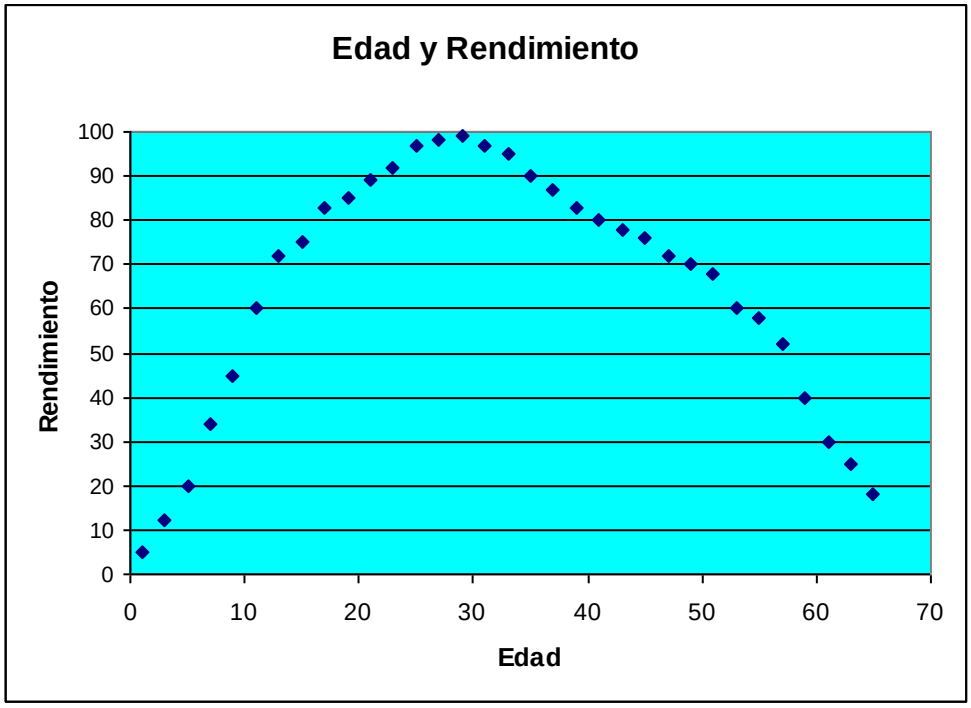
Ejemplo: Se toma una muestra entre un grupo de adultos mayores de 18 años para ver si existe una relación entre su edad y su escolaridad. La hipótesis es que no hay correlación porque en este rango de edad la gente ya dejó de estudiar.

Sujetos	Edad	A. de E.
A	19	8
B	22	9
C	24	2
D	27	9
E	30	2
F	33	4
G	37	8
H	41	9
I	49	9
J	52	4
K	57	7
L	60	9
M	63	5



Advertencia: es necesario tener precaución al interpretar el valor 0 obtenido partir del Coeficiente de Correlación de Pearson pues éste está basado en la fórmula de la línea recta y por lo tanto podría no detectar la relación entre variables que está relacionadas con otro tipo de líneas que no sean la recta.

Por ejemplo: las variables “edad” y “rendimiento físico” están íntimamente ligadas y es perfectamente conocido el hecho que los infantes tendrán un bajo rendimiento físico debido al poco desarrollo de sus facultades motrices y musculares. A partir de la juventud y madurez los seres humanos logran alcanzar el auge de esas facultades y su rendimiento alcanza el punto máximo. Después de la madurez, hombres y mujeres entran en un período de disminución de capacidades y aquellas facultades que impulsaron el rendimiento físico el cual se va debilitando. Al llegar de plano a la senectud, cuando la variable “edad” tiene sus valores máximos, el “rendimiento físico” retorna a los valores mínimos en virtud de que los sujetos ya carecen de fuerzas. Este comportamiento no puede ser medido por la fórmula de la línea recta. Tendría que usarse una fórmula correspondiente a la tendencia de la relación entre esas variables.



En otros casos, la correlación estará limitada por otros factores que tienen su explicación fuera de lo cuantitativo. Un autor sugiere como ejemplo una relación curvilínea entre el tamaño de la familia “número de hijos”, y el estatus socioeconómico. En su ilustración señala que las personas de bajo y alto estatus socioeconómico tendrán un mayor número de hijos en tanto que las familias con estatus socioeconómico medio tendrán un menor número de hijos (LEVIN; 1979: 202-203)

En este caso, el coeficiente de correlación sería débil o cero, ya que la relación curvilínea sería incompatible con la fórmula de la línea recta. Sin embargo, no se puede dejar de admitir la causalidad en la hipótesis que sugiere el autor. Esta situación se debe a que la forma de la relación está determinada por otra clase de línea que no se comporta como señala la fórmula de la línea recta.

Algunas variables pueden tener una relación consistente hasta cierto límite, después del cual se nulifica esa relación por la presencia de una constante. Veamos un caso como ejemplo: un conjunto de obreros tendrá un incremento en cierta habilidad, manifestada en número de piezas terminadas (Y), conforme reciba el adiestramiento respectivo, medido en semanas (X). Pero llegará el momento en que los valores de la

variable (Y) ya no podrán aumentar por mucho que se incremente (X) dado que existen límites físicos que determinarán el número posible de piezas que puede producir un ser humano. Entonces (Y) tendrá un valor constante independientemente que (X) aumente o no.

Por esta razón, cuando obtengamos como resultado un coeficiente de correlación igual a cero (r =0), debemos decir que estamos ante una **ausencia relativa** (y no absoluta) de correlación por que podría ser que las variables tratadas tengan otra forma de ser correlacionadas.

Aun cuando existe consenso entre lo que se considera una relación fuerte, ya sea positiva o negativa, es bueno revisar otras opiniones acerca de los puntos o límites desde los cuales ya se puede considerar una correlación fuerte o débil. Los autores siguientes ilustran esta variedad:

Por ejemplo, Levin (1979: 203) señala lo siguiente: “En cambio volvemos nuestra atención hacia los coeficientes de correlación, que expresan numéricamente tanto la fuerza como la dirección de la correlación lineal en línea recta. Tales coeficientes de correlación se encuentran generalmente entre - 1.00 y + 1.00 como sigue:

- 1.00 Correlación negativa perfecta.
- 0.95 Correlación negativa fuerte.
- 0.50 Correlación negativa moderada.
- 0.10 Correlación negativa débil.
  0.00 Ninguna correlación.
  0.10 Correlación positiva débil.
  0.50 Correlación positiva moderada.
  0.95 Correlación positiva fuerte.
  1.00 Correlación positiva perfecta”.

Otros autores lo plantean de manera distinta: “En general, los valores de correlación Pearson que fluctúan entre o.80 y 1.00 ó – 0.80 y – 1.00 son fuertes; los que van de 0.40 a 0.60 o de – 0.40 a – 0.60 se consideran moderados y los que van de 0.00 a 0.20 o de 0.00 a – 0.20 se consideran débiles” (WEINBERG Y GOLDBERG; s/f: 101).

El Coeficiente de Correlación de Pearson y la CAUSALIDAD entre variables.

Quizá parezca obvio señalarlo pero dado que muchos libros de estadística lo tratan como establecidos. Consideramos necesario puntualizar que el investigador debe definir y argumentar la relación de **causalidad** antes de usar el coeficiente de correlación.

La definición debe mostrarnos la forma en que la variable independiente (X) influye sobre la variable dependiente (Y) en tanto que en la argumentación se debe exponer las razones de esa influencia de acuerdo con los datos o información complementaria. Por ejemplo:

En la relación de variables “población total” (P.T.) y “población económicamente activa” (P.E.A.) la **definición** sería: “a medida que aumenta la P.T. también aumentarán los valores de la P.E.A”. Al respecto el investigador tendrá que **argumentar o sustentar** como es que los factores sociales tales como la fase de desarrollo económico, el tipo de economía, etcétera, se manifiestan en la relación de la Población Total (X) y la Población Económicamente Activa (Y).

En esta fase analítica de la causalidad podríamos afirmar la determinación que tiene X sobre Y en el sentido que la población total influye sobre la población económicamente activa; pero esta relación no es exclusiva, mecánica o ni estrictamente unidireccional.

Difícilmente encontraremos una relación de influencia en la que una variable independiente (X) actúe con exclusividad sobre la variable dependiente (Y). La ley de la concatenación universal de los fenómenos permite comprender la presencia de otras **variables asociadas** o **concurrentes** en la relación X y Y.

En el caso de la P.E.A. una variable **asociada** podría ser la migración; como ejemplo de variable **concurrente**, tendríamos la natalidad**.**

Se denomina variable **asociada**, en este caso a la migración, porque contribuye a la relación de las variables señaladas más no lo hace de una manera determinante sino que se incorpora al proceso con un peso menor al de la P.T.

La variable natalidad también participa en la mencionada relación, pero como una consecuencia más que como un factor causal. Es más un efecto que una causa, pero está presente y por lo tanto concurre o es **concurrente** en la relación.

A su vez la magnitud de la P.T. está afectada por una gran variedad de factores cuya manifestación concreta sería el número de habitantes de una sociedad.

La **interrelación** entre las variables.

Conviene prevenir que la misma **interrelación** de variables podría suscitar que la causalidad **cambie de dirección** en ciertos momentos sin que por ello deje de ser cierta relación predominante.

Podría acontecer, por ejemplo, que la concentración de P.E.A. en una ciudad sea la causante del incremento en la P.T. y que esto suceda en un momento de auge, no debe perderse de vista el efecto contrario a lo largo de un lapso mayor de observación.

Siempre para ejemplificar la interrelación, se expone el caso de la relación entre “los ingresos económicos” y “los años de escolaridad”. En un momento dado se puede afirmar que las personas que tienen mejores ingresos económicos alcanzarán más años de escolaridad, pero en otra fase de la relación se podrá comprobar que a mayor número de años de estudios se tendrán mayores ingresos económicos.

Los científicos sociales deberán conocer y manejar toda la información disponible con respecto a las variables para evitar caer en errores cuando use el coeficiente de correlación.

CAPÍTULO VI

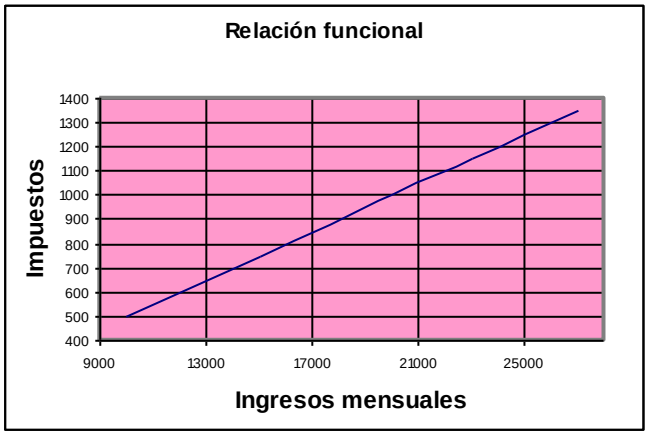
REGRESIÓN

Si al analizar la relación entre dos variables encontramos que su **coeficiente de correlación (r) es alto y que la causalidad** entre ambas ha sido ya establecida por medio de la información adicional cualitativa, entonces podemos utilizar los mecanismos del análisis de regresión para predecir los valores de una variable, a partir de los valores de la otra. Sin esas **dos condiciones** no tendrá mucho caso hacer alguna predicción salvo que por alguna razón sea importante realizar el ejercicio matemático. La noción básica para realizar el análisis de regresión simple, es la fórmula de la línea recta: **Y = a + bx**

Entre las variables existentes podemos encontrar dos tipos de relaciones: las funcionales y las estadísticas (NETER, WASSERMAN Y WHITMORE; 1978: 552)

Las **funcionales** son aquellas que se dan de una manera exacta y se expresan con una fórmula matemática, por ejemplo, (Y = a + bx). Por ejemplo, la relación entre los ingresos económicos de una persona y el monto de sus impuestos están clara y estrechamente relacionados. En estos casos, es decir, en las relaciones funcionales, el coeficiente de relación resulta ser igual a 1 ó -1.

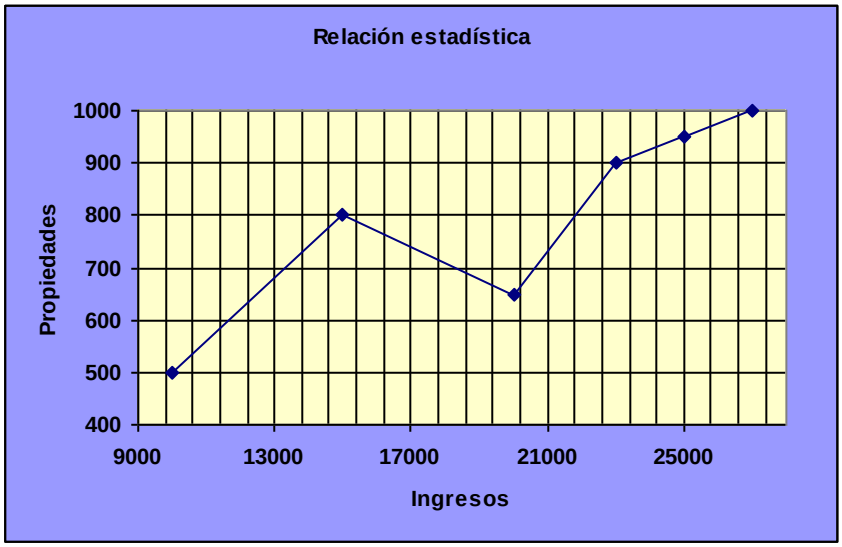
Ingresos mensuales	Impuestos
10000	500
15000	750
20000	1000
23000	1150
25000	1250
27000	1350



Las relaciones **estadísticas** nos muestran que existe una tendencia, que puede ser fuerte o débil, y que consiste en que los valores de "X" le corresponden aproximadamente a ciertos valores de "Y". A medida que el coeficiente de correlación se aproxima a 1 ó –1, la correspondencia entre dichos valores será más exacta, ya sea directa o inversamente proporcional, respectivamente; esto es, se acerca más al tipo de relaciones funcionales.

Por ejemplo, puede existir una relación entre el ingreso mensual de una serie de personas y el valor de sus propiedades respectivas, de tal manera que, a medida que una persona obtenga más ingresos, posee también mayor valor en propiedades. Esto no implica que la relación sea exacta, sino más bien indica una tendencia general en la correspondencia de los valores de la variable. La tendencia es medida por el coeficiente de correlación y su valor fluctúa en este caso entre 0.0 y 1.0, por ser una relación positiva o directa.

Ingresos mensuales	Valor en propiedad c. de miles
10000	500
15000	800
20000	650
23000	900
25000	950
27000	1000



Como ya se ha señalado, si el coeficiente de correlación entre las variables es significativo se puede utilizar los mecanismos del análisis de regresión para predecir los valores de una variable, a partir de los valores de la otra. Este análisis puede ser simple, múltiple y parcial. El primero únicamente hace referencia a dos variables, una independiente y la otra dependiente. Los análisis múltiple y parcial trabajan con tres o más variables, pero con la característica de que una variable solamente es la dependiente y las otras dos o más son variables independientes.
(SHAO; 1978: 616).

Sin embargo, por ahora solo trataremos el análisis de regresión simple, en el cual, la tiene como directriz la fórmula de la línea recta es tomada como base y es la siguiente:

$$Y = a + bx$$

Donde ‘a’ y ‘b’ valores constantes y ‘X’ es el valor dado o conocido a partir del cual se obtiene ‘Y’.

Tomando el ejemplo en el que se relacionan las variables ingreso mensual y valor en propiedades se formula la siguiente pregunta ¿cuál será el monto de valor en propiedades para una persona que gana 30000, 35000 o 40000 pesos mensuales?

Aplicando el mecanismo de regresión se obtienen los siguientes resultados:

Ingresos mensuales	Valor en propiedad c. de miles
10000	500
15000	800
20000	650
23000	900
25000	950
27000	1000
30000?	1062
35000?	1193
40000?	1324

Una de las más frecuentes aplicaciones del análisis de regresión es el cálculo de valores de una variable Y cuando en la variable X está implicado el factor tiempo. Es decir, cuando los valores de X son días, meses y años (o cualquier unidad de tiempo con secuencia invariable. Este tipo de análisis pertenece al campo de estudio de las series cronológicas y sus tendencias. Sin embargo la necesidad de la predicción en el campo social, y aprovechando que ya se ha explicado el concepto y mecanismo de la regresión, podemos entonces tratarlo ahora.

La más importante recomendación para este uso del mecanismo de regresión es la siguiente: si las categorías de la variable cronológica se expresa en términos nominales (días, meses) entonces hay que sustituirlos por la secuencia de números naturales empezando por la unidad. Para hacer la predicción se utilizara el valor del número natural que corresponda en la secuencia. Ejemplo:

Inflación	N° Naturales	Índice
Enero	1	2.3
Febrero	2	2.8
Marzo	3	3.3
Abril	4	3.9
Mayo	5	4.4
Junio	6	5.1
julio?	7?	5.5

EJERCICIOS DE CORRELACIÓN Y REGRESIÓN

1. El índice de precios al consumidor (IPC) en una de los factores que más afectan al poder adquisitivo del salario mínimo (PASM). Para conocer esta relación encuentra el coeficiente de correlación y responde ¿cuál será el PASM dos zonas económicas cuyos IPC son de:

ZONA	IPC	PASM
1	240	210
2	260	190
3	280	170
4	250	180
5	275	175
6	300	150
7	220	200
8	230	210
r = -0.91		
	200?	225
	400?	84

2. Un psicólogo laboral afirma que el ingreso mensual (IM) de los ejecutivos es muy importante para elevar el coeficiente de motivación al logro (CML). Analiza los datos de la tabla de abajo y encuentra la medida de la relación y responde a las siguientes preguntas ¿Cuál será el coeficiente de motivación al logro de un ejecutivo que tenga...

a) ¿30 mil pesos mensuales?

b) ¿90 mil pesos mensuales?

Ejecutivos	IM(miles)	CML
A	35	55
B	38	60
C	85	90
D	44	60
F	62	75
G	58	70
H	36	50
I	70	83
J	75	89
r = 0.98		
	30?	52.22
	90?	98.56

3. En nuestro país cada día se registra una mayor participación de la Población Económicamente Activa Femenina (PEAF) en relación a la Población Económicamente Activa Total (PEAT) ¿cuál es, en realidad, la magnitud de esa relación y que PEAF se esperaría encontrar en municipios donde la PAET es de:

8,000	Hab.	Y= 0.93
30,000	Hab.	Y= 3.19
58,000	Hab.	Y= 6.07

Municipios	PEAT	PEAF
1	52.2	4.9
2	40. 3	4.7
3	37.8	4.5
4	32.5	3.7
5	25.7	2.9
6	24.1	2.8
7	19.3	1.7
8	15.6	1.8
9	14.5	1.5
10	10.1	1.1

r = 0.97

8?	1.188
30?	3.383
58?	6.176

4. Un investigador social quiere conocer la relación entre la escolaridad en años terminados (AT) y el grado de prejuicios religiosos (PR) a partir de una muestra de 40 personas.

Observa los datos recabados y estima el coeficiente de correlación entre las variables. A partir del mecanismo de regresión predice cual serían los grados de prejuicios religiosos que podría tener tres personas cuya escolaridad en años sea de 3, 8 y 19.

Caso	AT	PR	Caso	AT	PR
1	9	5	21	1	38
2	2	40	22	2	36
3	15	5	23	3	29
4	4	32	24	4	29
5	13	10	25	5	27
6	6	26	26	5	23
7	3	31	27	6	26
8	4	28	28	7	23
9	0	45	29	8	18
10	7	17	30	9	17
11	12	20	31	10	15
12	18	7	32	11	14
13	20	2	33	11	12
14	15	6	34	12	10
15	1	40	35	13	11
16	6	22	36	14	9
17	6	12	37	15	8
18	18	4	38	16	8
19	19	4	39	17	5
20	0	35	40	18	3

5. Un sociólogo afirma que existe una relación alta y positiva entre el número de habitantes de las ciudades y la cantidad de población económicamente activa. Para demostrarlo analiza la lista de datos correspondientes a las variables “número de habitantes” y “población económicamente activa”. Comprueba si el sociólogo tiene razón y estima cuál será la P.E.A. de una ciudad cuyo número de habitaciones sea de...

4,500 personas?

40,000 personas?

90,000 personas?

CIUDAD	***HABITANTES -----MILES-----***	***P.E.A***	***CIUDAD***	***HABITANTES -----MILES-----***	***P.E.A.***
1	75.3	26.3	***26***	15.7	3.5
2	24.4	9.5	***27***	71.1	22.9
3	7.5	4.6	***28***	74.2	25.0
4	18.2	9.3	***29***	49.2	17.2
5	22.4	4.8	***30***	41.7	15.6
6	12.5	2.4	***31***	32.2	12.9
7	31.7	10.1	***32***	37.4	15.1
8	42.1	14.7	***33***	25.5	8.9
9	33.2	12.6	***34***	22.8	8.0
10	60.2	23.1	***35***	13.8	4.8
11	70.1	27.5	***36***	16.1	5.6
12	72.3	22.3	***37***	9.2	4.2
13	69.4	22.2	***38***	7.9	3.8
14	8.4	1.9	***39***	73.5	28.7
15	26.2	9.2	***40***	71.6	22.1
16	31.4	12.0	***41***	66.7	21.3
17	19.6	8.9	***42***	64.2	21.5
18	45.3	18.9	***43***	58.3	20.4
19	42.5	11.9	***44***	54.2	20.0
20	10.6	2.7	***45***	47.1	17.5
21	53.6	17.8	***46***	43.6	16.3
22	60.8	21.3	***47***	38.2	12.4
23	57.4	21.1	***48***	32.8	10.5
24	36.4	14.7	***49***	27.5	9.1
25	17.2	9.0	***50***	21.3	7.5

BIBLIOGRAFÍA

AZORÍN POCH, Francisco

1972 *Curso de Muestreo y sus aplicaciones*, Ed. Aguilar. Madrid.

BLALOCK, Hubert

1978 *Estadística Social.* Fondo de Cultura Económica, México

COCHRAN, William G.

1976 *Técnicas Muestreo*, Ed. C.E.C.S.A. México

DUVERGER, Maurice

1978 *Métodos de las Ciencias Sociales*, Ed. Ariel. México

ELORZA PÉREZ TEJADA, Haroldo

2000 *Estadística para las ciencias sociales y del comportamiento.* Oxford University Press, México (MacStat 2.0).

EVIA CERVANTES, Carlos A.

1989 *Curso de Estadística*. Facultad de Ciencias Antropológicas. Mérida. Capítulo Segundo.

GARDUÑO ARGUETA, Jaime

1979 *Introducción al estudio del patrón de asentamiento del sitio de Cobá, Quintana Roo*. México. Tesis profesional de licenciatura de la Escuela Nacional de Antropología e Historia.

FLOUD, Roderick

1975 *Métodos cuantitativos para historiadores*. Alianza Universidad, Madrid. Cap. 4

LEVIN, Jack

1979 *Fundamentos de Estadística en la Investigación Social.* Harla, México.

LOHR, Sharon L.

2000 *Muestreo: diseño y análisis*. International Thomson. México.

MILLS, Frederick C.

1969 *Métodos Estadísticos*. Aguilar. Madrid.

MONTEMAYOR GARCÍA, Felipe
1973 *Fórmulas de Estadísticas para Investigadores*. INAH-SEP. México. Tomo I.
NETER, John, William Wasserman y G.A. Whitmore
1978 *Fundamentos de Estadística para Negocios y Economía*. México. Ed. C.E.C.S.A.
SECRETARÍA DE EDUCACIÓN PÚBLICA
2001 "Calidad educativa. De 'panzazo' aprueba Yucatán evaluación". En: *"El mundo al día"*. Mérida, Yucatán.
SHAO, Stephen
1978 *Estadísticas para economistas y administradores de empresas*. México. Herrero Hermanos, Sucs.
SPIEGEL, Murray
1978 *Estadística.* Libros McGraw - Hill, México, 1978
STEVENSON, William J.
1978 *Estadística para administración y economía.* Harla, México
RASCON, Octavio.
1983 *Introducción a la estadística descriptiva.* UNAM, México
TARRÉS, María Luisa
2001 *Observar, escuchar y comprender sobre la tradición cualitativa en la investigación social*. México. Facultad Latinoamericana de Ciencias Social y El Colegio de México.
WEINBERG, S.L. Y GOLDBERG, K.P.
S/f *Estadística básica para las ciencias sociales*, Ed. Interamericana. México
YAMANE, Taro
1977 *Estadística*. México. Editorial Harla.

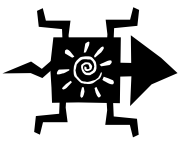


TABLA DE NÚMEROS ALEATORIOS (Montemayor; 1973: 399)

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