unam - ents

Universidad Nacional Autónoma de México Escuela Nacional de Trabajo Social

Estadística Aplicada a la Investigación Social I

Ing. José Luis Sandoval Dávila

Área: Metodología y
Práctica de Trabajo Social

Semestre: 3

Créditos: 5

Carácter: Obligatoria

CONTENIDO

PRESENTACIÓN

La Escuela Nacional de Trabajo Social inició sus estudios de *Licenciatura en Sistema Universidad Abierta,* en el año escolar 2003, con el Plan de Estudios aprobado por el H. Consejo Universitario el 10 de julio de 1996. Fue reestructurado en el año 2002 con aprobación del Consejo Académico del Área de las Ciencias Sociales, en su sesión del 26 de noviembre de 2002.

En el Sistema Universidad Abierta, la relación entre asesores, estudiantes y material didáctico es fundamental. En este sentido, en la escuela se puso especial atención para lograr mayor calidad en los materiales.

De ésta manera, el material que ahora te presentamos debe constituirse en una herramienta fundamental para tu aprendizaje independiente, cada uno de los componentes que lo integran guardan una congruencia con el fin de que el estudiante pueda alcanzar los objetivos académicos de la asignatura.

El material pretende desarrollar al máximo los contenidos académicos, temas y subtemas que son considerados en el programa de estudio de la asignatura. Esto no pretende soslayar el papel y responsabilidad preponderante del estudiante, que debe profundizar en la búsqueda de conocimientos en todas aquellas fuentes que tenga a su alcance hasta hacer realidad los objetivos y el perfil de egreso propuesto.

Este material es perfectible, por ello, con el apoyo de las experiencias de los estudiantes y otros profesores, serán revisados y actualizados por el asesor de manera permanente, cuyos aportes sin duda, contribuirán para su mejora y enriquecimiento.

Te damos la más cordial bienvenida y te deseamos toda clase de éxitos en los estudios que inicias en esta, tu escuela, la **Escuela Nacional de Trabajo Social** de la **Universidad Nacional Autónoma de México.**

INTRODUCCIÓN

Este material presenta los fundamentos más relevantes de la estadística aplicada a la investigación social y proporciona al estudiante del Sistema Universidad Abierta, de manera amena, sencilla y práctica, las herramientas estadísticas y diversos criterios para el análisis cuantitativo y cualitativo de los datos involucrados en los fenómenos sociales, actuales o de interés, que contribuyan a generar la información necesaria para la efectiva toma de decisiones en el contexto de las ciencias sociales.

El material ilustra la secuencia cronológica que el estudiante deberá seguir para describir grupos de datos, considerando sustancialmente los aspectos gráfico y numérico. El procedimiento conducirá al alumno por las etapas básicas del proceso para el tratamiento estadístico de datos en la investigación social.

Con este material se pretende que el estudiante de ciencias sociales cambie el paradigma de que la estadística, como rama de las matemáticas, requiere de firmes conocimientos de esta ciencia; nada más alejado de la realidad, ya que la estadística requiere esencialmente de sentido común e involucramiento en los fenómenos estudiados. Por supuesto que trataremos con números y fórmulas, pero paradójicamente no dependemos de su dominio. Se insiste en los conceptos, ya que se considera fundamental que en toda área de desarrollo, las definiciones marcan la diferencia entre la eficiencia, la eficacia y la efectividad del trabajo.

Por tal virtud, se procura evitar cálculos y exceso de ejercicios, éste no es un cuaderno de trabajo, ya que para ello existen muchas fuentes; lo más importante es que la práctica que deberá realizarse dependerá de las condiciones y circunstancias que se presenten al momento de su realización, apegándose a la situación global del mercado social.

La estadística en la ciencias sociales se distingue notablemente de aquella con aplicaciones a diversos ámbitos del conocimiento tales como: administración, medicina, actuaría, producción, ingeniería; en virtud de que estas últimas,

tomadas como referencia, basan o sustentan sus métodos en el tratamiento de variables numéricas, las cuales presentan un marco de referencia natural, a través de sus unidades de medida.

Debido a que las ciencias sociales se enfrentan a variables de orden cualitativo; para su tratamiento y análisis, la estadística social considera sustancialmente su naturaleza la cual permite dimensionar a las variables no numéricas en función de sus atributos, ya sean del orden clasificatorio o jerárquico, sin soslayar la importancia que las variables numéricas tienen en los fenómenos sociales. Sin embargo, no debe perderse de vista que todo fenómeno analizado a través del tratamiento estadístico de sus datos, sea cual fuere su naturaleza, deberá ser dimensionado ya sea por medio de sus valores o de sus atributos, correspondiendo esa función a las escalas de medición, herramienta fundamental de la estadística social.

Esto subraya la importancia de la estadística social en cualquier área del conocimiento, ya que todo fenómeno o materia de estudio en que se vea involucrada la estadística no podrá prescindir de la presencia de variables del orden cualitativo.

Una vez identificada la naturaleza de las variables, se determinará el tipo de tratamiento a que habrá de ser sometida cada una de ellas, con el propósito de describir, en primer término, la naturaleza del fenómeno o tema en estudio. Se menciona lo anterior debido a la necesidad de utilizar y aplicar las herramientas adecuadas que, como se señaló anteriormente, ofrece la estadística social, la cual se diferencia de cualquier otra aplicación por la presencia de variables de orden principalmente cualitativo.

Para su descripción, las variables requieren de ser sometidas al proceso de organización de datos, lo que simplificará el análisis. La descripción se sustenta principalmente en los aspectos gráfico y numérico, destacando en ellos la forma, tendencia y dimensión del fenómeno en estudio.

En esta asignatura la descripción de variables se centra en su tratamiento numérico, requiriéndose conocer sus niveles de concentración, dispersión, tendencia y proyección. Esto se soporta en métodos para el cálculo de

promedios, medidas de dispersión, medidas de posición, medidas de distribución y factores de relación entre variables, lo que permitirá explicar el comportamiento de las variables en función de otras como el tiempo y aspectos cualitativos como el comportamiento organizacional y niveles de satisfacción de la población ante los servicios públicos, por citar ejemplos.

Sin embargo, debe considerarse que los cursos de estadística en el nivel licenciatura no son simples sesiones para el cálculo numérico, manejo de calculadora o sistemas computacionales, ya que éstos son solamente los medios o herramientas que se utilizan para la obtención de resultados y a partir de ellos inicia realmente la intervención del profesional en el uso de la estadística, con enfoques firmes hacia el análisis de resultados, generación de información y toma de decisiones.

Con mucho interés se espera la opinión, comentarios y sugerencias del lector, para que este material sea actualizado y simplifique el acceso de los estudiantes de ciencias sociales, al mundo maravilloso de la estadística como estrategia para competir.

Para complementar el aprendizaje y entendimiento de los principios de la estadística, que por naturaleza es muy amplia, se recomienda, en primer término el estudio de este material y consultar otras fuentes, con lo que finalmente el lector se formará el mejor de los juicios para el uso y aplicaciones de la estadística en las ciencias sociales.

OBJETIVO GENERAL

Al finalizar el curso, el estudiante conocerá las herramientas de la estadística descriptiva y la efectividad de su utilización como apoyo a la investigación social.

PERFIL DE EGRESO

Al finalizar el curso el alumno tendrá la habilidad para:

- Utilizar el proceso para la organización y descripción de un conjunto de datos.
- Identificar la escala de medición de las variables y aplicar las técnicas estadísticas que le corresponda a cada una de ellas.
- Emplear las medidas de tendencia central que mejor representen a un conjunto de datos.
- Analizar las características de un conjunto de datos a partir de la forma y tendencia de su distribución.
- Analizar la relación entre variables y el grado de asociación entre ellas.

Asimismo, adquirirá las aptitudes que le permitirán utilizar con efectividad los recursos estadísticos para la generación de información y la toma de decisiones en el contexto social principalmente.

TEMARIO GENERAL

1. CONCEPTOS.
   - 1.1 Conceptos de Estadística.
   - 1.2 Objetivo de la Estadística.
   - 1.3 Clasificación de la Estadística.
   - 1.4 Variables.

2. NIVELES DE MEDICIÓN.
   - 2.1 Niveles o escalas de medición.
   - 2.2 Enfoque numérico a variables cualitativas.

3. ORGANIZACIÓN DE DATOS.
   - 3.1 Etapas del proceso estadístico de una investigación.
   - 3.2 Fuentes de recolección de datos.
   - 3.3 Aspectos más importantes para describir un conjunto de datos.
   - 3.4 Organización de datos.
     - 3.4.1 Ordenación de datos.
     - 3.4.2 Clasificación de datos.
       - 3.4.2.1 Intervalo Cerrado.
       - 3.4.2.2 Intervalo abierto.
       - 3.4.2.3 Rango
     - 3.4.3 Distribución de datos.
       - 3.4.3.1 Frecuencia Absoluta.
       - 3.4.3.2 Frecuencia Relativa.
       - 3.4.3.3 Frecuencia Porcentual.
       - 3.4.3.4 Frecuencia Acumulada.
   - 3.5 Marca de clase.
   - 3.6 Representaciones gráficas de la organización de datos.
     - 3.6.1 Histograma.
     - 3.6.2 Polígono de Frecuencias.
     - 3.6.3 Diagrama de Barras.
     - 3.6.4 Gráfica Sectorial.
     - 3.6.5 Ojiva.

4. MEDIDAS DE TENDENCIA CENTRAL O PROMEDIOS.
    - 4.1 Conceptos.
    - 4.2 Media.
    - 4.3 Media Ponderada.
    - 4.4 Mediana.
    - 4.5 Moda.
    - 4.6 Relación que guardan la Media, Mediana, y Moda con las variables por nivel de medición.
        - 4.6.1 Escalas nominales.
        - 4.6.2 Escalas Ordinales.
        - 4.6.3 Escalas de Intervalos y de proporción.
    - 4.7 Algunos ejemplos de la relación entre niveles de medición y medidas de tendencia central.

5. MEDIDAS DE POSICIÓN.
    - 5.1 Cuartiles.
    - 5.2 Deciles.
    - 5.3 Percentiles.

6. MEDIDAS DE DISPERSIÓN.
    - 6.1 Rango.
    - 6.2 Desviación Estandar.
    - 6.3 Coeficiente de Variación.

7. MEDIDAS DE DISTRIBUCIÓN.
    - 7.1 Sesgo.
        - 7.1.1 Sesgo Negativo.
        - 7.1.2 Sesgo Positivo.
        - 7.1.3 Distribución Sesgada.
    - 7.2 Curtosis.

8. ANÁLISIS DE REGRESIÓN Y CORRELACIÓN.

8.1 Análisis de regresión.

8.1.1 Gráfico.

8.1.2 Semipromedios.

8.1.3 Mínimos Cuadrados.

8.1.4 Ecuaciones normales para el ajuste por el método de mínimos cuadrados.

8.2 Análisis de Correlación.

MAPA CONCEPTUAL

ESTADÍSTICA DESCRIPTIVA

Variables por nivel de medición :

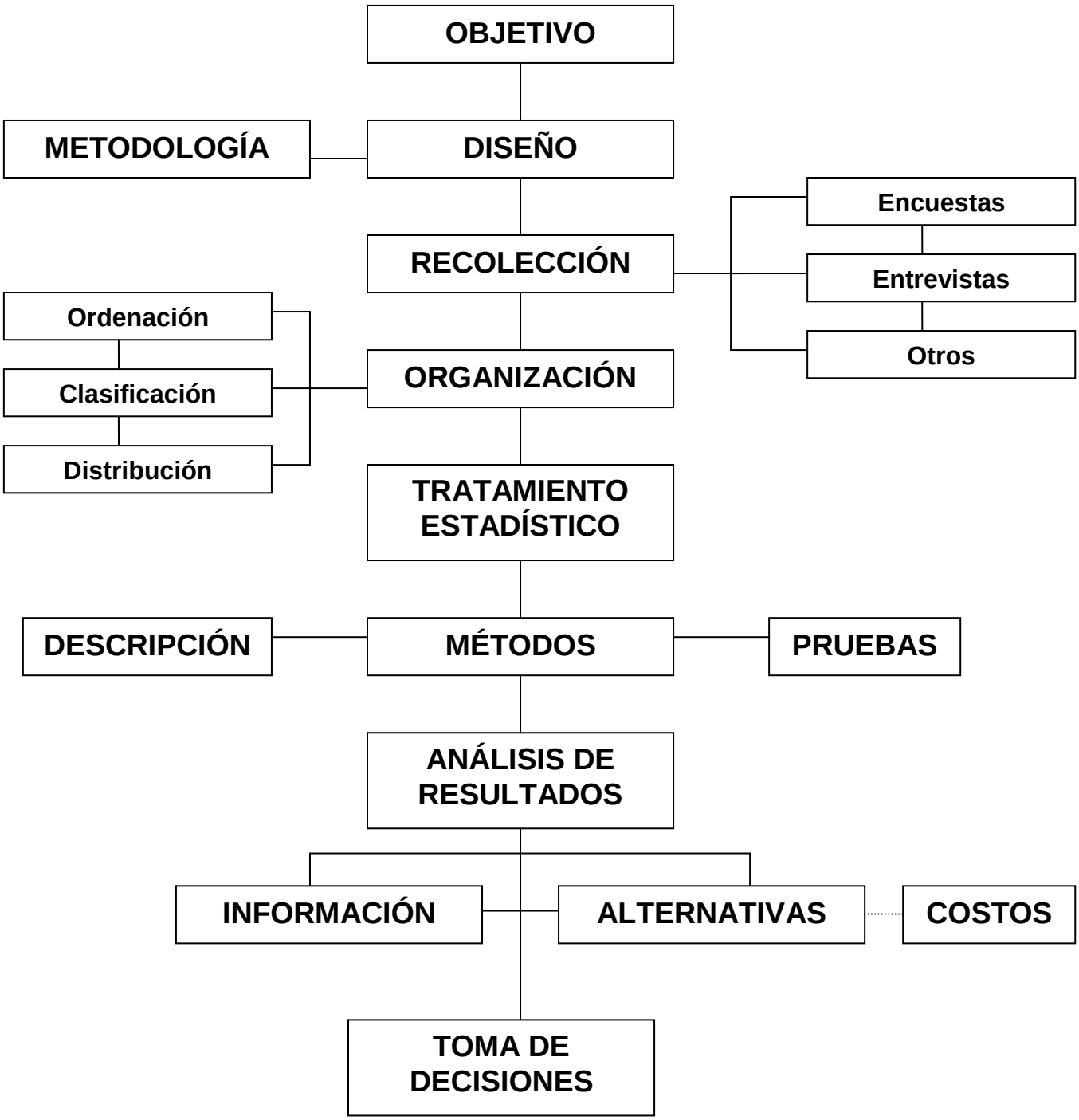
Nominal
Ordinal
Intervalar } Escalar
Racional

	Organización de datos			**Medidas descriptivas**	
Tablas: Distribución de frecuencias Ordenación Clasificación Distribución		**Gráficas :** Histogramas Diagramas de barras Polígonos de frecuencias Ojivas Gráficas de sector Diagramas de dispersión	**Tendencia central :** Media Mediana Moda	**Posición :** Cuartiles Deciles Percentiles	**Dispersión :** Rango Desviación Estándar Varianza Coeficiente de variación **Distribución :** Sesgo Curtosis

Análisis de Regresión y Correlación

Regresión : Diagrama de dispersión Simple y múltiple Lineal y no lineal Positiva y negativa **Ecuación de regresión :** Métodos de regresión : Gráfico Semipromedios Mínimos cuadrados	**Correlación :** Coeficiente de correlación de Pearson Coeficiente de correlación de Spearman Correlación nula Correlación perfecta

Esquema del proceso estadístico en la Investigación Social



UNIDAD 1

CONCEPTOS.

INTRODUCCIÓN.

La estadística aplicada se fundamenta en diversos conceptos que permiten al estudiante la identificación de los recursos que deberán ser empleados durante el tratamiento de datos. Estos conceptos contribuirán a que el proceso de inducción al curso se simplifique, en virtud de que se conocerán algunos enfoques de la estadística, su objetivo, su clasificación y las variables que son tratadas mediante sus técnicas, métodos y procedimientos; así como el tipo de datos sujetos de análisis.

OBJETIVO.

El alumno conocerá los conceptos básicos de estadística, para su identificación en el tratamiento de datos de fenómenos sociales.

TEMARIO.

1. CONCEPTOS.
    - 1.1 Conceptos de Estadística.
    - 1.2 Objetivo de la Estadística.
    - 1.3 Clasificación de la Estadística.
    - 1.4 Variables.

1. CONCEPTOS.

1.1 CONCEPTOS DE ESTADÍSTICA.

- Conjunto de normas, técnicas, métodos y procedimientos, utilizados en la investigación social, mediante la recopilación de datos y el análisis de resultados.

- En casos particulares, se ocupa de recoger, clasificar, representar y resumir los datos de muestras, y de hacer inferencias acerca de las poblaciones de las cuales proceden.

- Conocimiento y estudio de los métodos para la obtención, organización, presentación y descripción de información numérica.

1.2 OBJETIVO DE LA ESTADÍSTICA.

- Proporcionar las técnicas, métodos y procedimientos requeridos para analizar conjuntos de datos y así simplificar la descripción e inferencia de sus resultados

- Obtener conclusiones de una población, a partir de las observaciones y análisis realizados a una muestra.

- Realizar investigaciones por medio de la recolección de datos para la generación de resultados que posteriormente a su análisis, contribuirán a la toma de decisiones.

1.3 CLASIFICACIÓN DE LA ESTADÍSTICA.

La estadística se clasifica en dos ramas:
La estadística DESCRIPTIVA y la estadística INFERENCIAL.

- La estadística descriptiva se encarga de analizar la forma y dimensión de un grupo específico de datos.

- La estadística inferencial se encarga de obtener las características e información de una población a partir de una muestra.

- La estadística inferencial estudia los datos de una muestra, para generalizar las características de la población de la cual provienen.

- La estadística inferencial plantea, resuelve el problema de establecer previsiones y conclusiones generales, relativas a una población mediante leyes de la probabilidad y haciendo uso de métodos inductivos.

EJEMPLO: Cuatro familias de la colonia A y tres de la colonia B se seleccionan para determinar la oportunidad con la que realizan el pago de los servicios públicos, esto es, qué tan a tiempo los pagan. La oportunidad de las familias de la colonia A es de 18, 19, 23, 24 días posteriores a la fecha límite. La oportunidad de las familias de la colonia B es de 20, 20, 25 días en el mismo contexto.

a) La oportunidad del pago de las cuatro familias de la colonia A es menor que la de las tres de la colonia B.

b) Probablemente, la oportunidad promedio del pago de todas las familias de la colonia A sea de 21 días posteriores a la fecha límite.

c) Si el importe de los servicios de la colonia A es el mismo que los de la colonia B, ¿es recomendable otorgar facilidades a la colonia B?

¿Qué rama de la estadística se está refiriendo en cada inciso?

1.4 VARIABLES.

Existen dos tipos de variables: las numéricas y las no numéricas, también identificadas como cuantitativas y cualitativas.

Las variables numéricas están clasificadas a su vez en discretas y continuas.

Variables discretas: son aquellas que pueden tomar valores determinados dentro de un intervalo dado. Se utilizan principalmente para el conteo, en ellas no es posible encontrar valores intermedios entre dos valores inmediatos.

EJEMPLOS:

El número de familias en un conjunto habitacional.

La cantidad de aparatos de televisión por vivienda.

El número de personas solicitantes de empleo.

En estos ejemplos se aprecia la imposibilidad de poder tener fracciones de familia, de aparatos de televisión o de personas demandantes de empleo.

Variables continuas: pueden tomar cualquier valor dentro de un intervalo dado, utilizándose principalmente para medir.

EJEMPLO:

La estatura de las mujeres de una determinada ciudad.

El tiempo de tardanza entre uno y otro autobús de servicio público.

En estos casos la variable puede tomar cualquier valor dentro de un intervalo de estaturas o tiempo. Es posible que la estatura de una mujer sea de un metro

con 63 centímetros u otra fracción. Un autobús puede demorar 10 minutos con 15 segundos y 7 décimas, y si se tuviera un instrumento que permitiera medir más detalladamente el tiempo entre una y otra unidad de transporte público, sería factible su medición exacta.

La estadística es una rama de las matemáticas sustancial en la formación académica y profesional de cada persona, ya que es una herramienta que tiene por objeto facilitar el estudio mediante una metodología de aprendizaje que va de lo simple a lo complejo. Contribuye a la toma de decisiones a través de recoger, organizar y procesar datos para obtener información en la que se sustenta la toma de decisiones.

En el contexto de la estadística social, a estas variables numéricas les puede corresponder, según su naturaleza, la escala de medición de intervalo o de razón, conceptos que serán abordados en la siguiente unidad.

Es menester señalar que a las variables discretas les corresponden datos discretos y a las variables continuas, datos continuos; sin embargo, por razones prácticas, circunstancialmente las variables discretas adoptan datos continuos y las continuas toman datos discretos.

EJEMPLO: las estadísticas indican que las familias mexicanas tienen 3.5 hijos y que el ingreso de la población es de $60 pesos diarios.

De la primera parte del ejemplo se deduce que la variable “Número de hijos” es discreta, por lo tanto sus datos deberán ser discretos; sin embargo, por razones estadísticas se presenta como dato continuo, cuyo significado indica que las familias tienen de 3 a 4 hijos. En la segunda parte del ejemplo tenemos una variable continua y el dato que recibe es discreto. Por razones prácticas se redondea el ingreso y se expresa como un valor entero.

UNIDAD 2

NIVELES DE MEDICIÓN.

INTRODUCCIÓN.

La diferencia entre la estadística social y sus diversas aplicaciones en otras disciplinas se encuentra en el tipo de variables que inciden en los fenómenos sociales, donde principalmente la ocurrencia es de aquellas del orden cualitativo, de atributo o categóricas, ofreciéndose por lo tanto una alternativa para su tratamiento a través de mediciones no necesariamente numéricas.

El tratamiento de ese tipo de variables no numéricas, reviste una particular importancia en ciencias sociales, debido a la existencia de métodos, técnicas y pruebas estadísticas propias a estas variables; lo que permite realizar un análisis más apropiado, clasificándolas como nominales, ordinales, intervalares y racionales; integrando a estas dos últimas en la categoría escalar.

OBJETIVO.

El alumno identificará la escala o nivel de medición correspondiente a las variables en estudio y su importancia en las ciencias sociales.

TEMARIO.

2. NIVELES DE MEDICIÓN.
   - 2.1 Niveles o escalas de medición.
   - 2.2 Enfoque numérico a variables cualitativas.

2. NIVELES DE MEDICIÓN.

2.1 NIVELES O ESCALAS DE MEDICIÓN.

Naturalmente, las variables poseen características que permiten dimensionarlas en función de sus valores o atributos, lo que se identifica como nivel o escala de medición.

Las variables cuantitativas se miden en escalas de intervalo o razón. Las llamaremos también escalares, por tal virtud recuérdense los conceptos de valor absoluto y relativo de un número.

Intervalo: En este caso todas las variables son numéricas. Los valores pueden ser clasificados en categorías que tienen un orden o jerarquía, la diferencia entre sus valores es significativa. La diferencia entre dos valores tiene una dimensión real y su origen (0) es relativo.

EJEMPLO: la temperatura en un centro de trabajo, que puede ser medida a través de la escala de grados Centígrados o Fahrenheit. El origen en esas escalas es diferente: cero grados centígrados es equivalente a 32 grados Fahrenheit lo que indica que ese punto cero es relativo, ya que al medir la temperatura de un mismo lugar se tienen dos referencias distintas, aunque equivalentes.

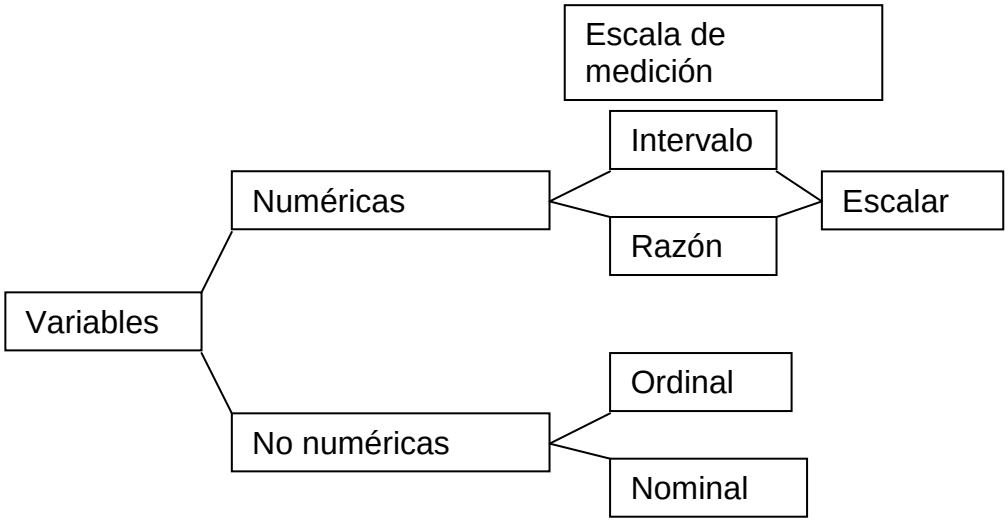
Razón: Con las mismas características de las variables de intervalo, en éstas su origen es absoluto. Ejemplo: el número de hijos de las mujeres de una institución gubernamental. En cualquier lugar y tiempo, tener cero hijos representa exactamente lo mismo, no hay equivalencias.

Nominal: Estas variables tienen la característica de ser no numéricas o cualitativas. Son únicamente clasificatorias y su organización depende exclusivamente del criterio del investigador. Ejemplo: estado civil, sexo, nacionalidad.

Ordinal: Estas variables pueden ser numéricas o no numéricas. Las numéricas únicamente indican o refieren orden, careciendo de significado las diferencias entre sus valores. Ejemplo: el número de cuenta de un alumno.

Las no numéricas con este nivel ordinal, son aquellas que sus categorías indican una relación de jerarquía entre ellas, esto es, evidencian un sentido de mayor o menor que.

EJEMPLO: estrato social, puesto en el trabajo. Aún sin poder especificar la diferencia real entre pertenecer a un estrato social o a otro, o determinar la diferencia entre dos categorías en el catálogo de puestos de una organización se sabe que existen diferencias jerárquicas entre las características de ambas variables.



Finalmente, debemos observar que las variables cualitativas tienen un nivel de escala nominal u ordinal, mientras que las variables numéricas tienen un nivel de escala de intervalo o de razón.

2.2 ENFOQUE NUMÉRICO A VARIABLES CUALITATIVAS.

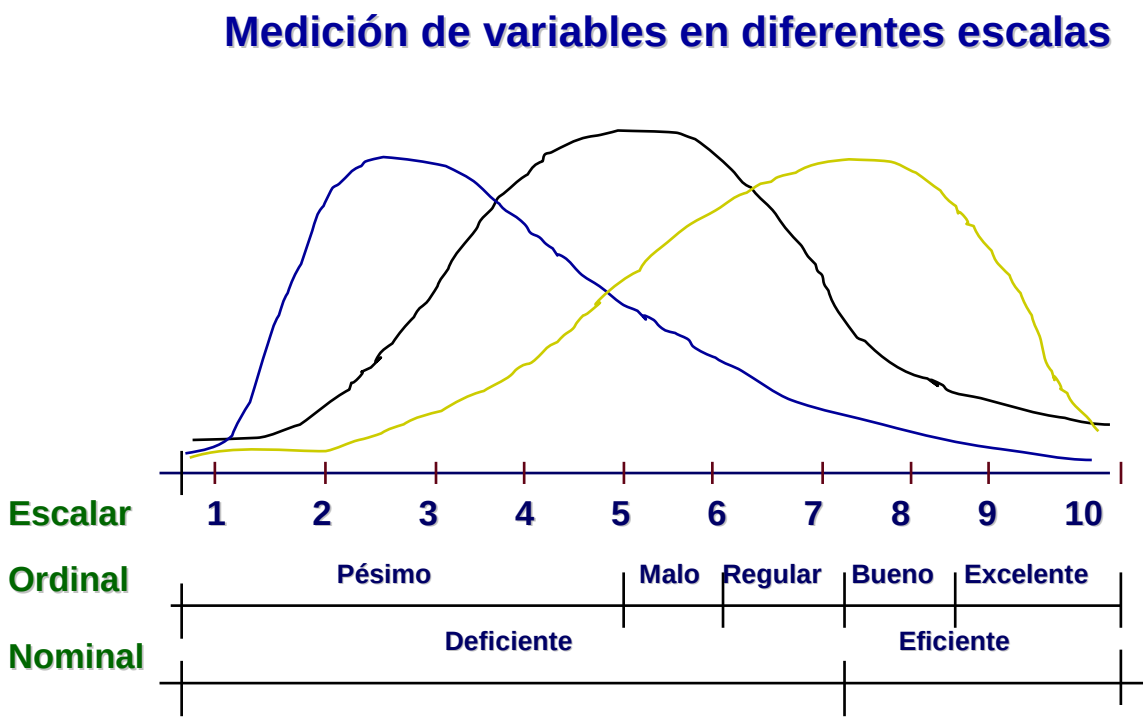
Una de las principales características en el tratamiento de datos en ciencias sociales, vista como limitaciones o complicaciones por diversos sectores del área, es que sus variables son principalmente del orden cualitativo, por lo que el único tratamiento que los clásicos le daban a las investigaciones era el conteo de cifras expresadas en porcentajes y sus representaciones gráficas. Esto sigue siendo una herramienta de información descriptiva muy importante y seguramente imprescindible para los profesionales de la investigación social; sin embargo, esos alcances implican limitaciones que impiden que el tratamiento de datos y su análisis en ciencias sociales tenga mayor flexibilidad, variedad y alcance.

“Si en ciencias sociales las variables fueran numéricas, otra cosa sería”, afirmaba en mis inicios como catedrático universitario, sin saber que sólo faltaba un poco de visión e involucramiento con las técnicas, necesidades y recursos de la investigación social, y que siempre hemos tenido al alcance de nuestra mano. Por tal virtud, deseo ofrecer al lector una idea para que las variables cualitativas sean tratadas, en gran medida, como variables numéricas.

Muy sencillo. En la investigación es frecuente observar la tendencia y preferencia de estudiantes, profesores, investigadores y profesionales en los estudios de mercado y de opinión, así como en encuestas diversas, por respuestas jerárquicas en las preguntas de un cuestionario, una de ellas es la escala de Likert, entre otras. Estas tienen la limitante, por ejemplo, de que cuando ya fue categorizada la respuesta y contestada por el entrevistado, no hay forma de darle otro tratamiento estadístico que no sea el ordinal o el clasificatorio, y en algunos casos la realización de algunas pruebas para contrastar opiniones y respuestas, lo que limita o restringe los alcances que se pueden tener si las variables y sus datos pudiesen ser tratados con una escala numérica que permitiera darle una dimensión cuantitativa de mayor impacto y, si la investigación lo requiere o el investigador lo necesita, también darle el tratamiento clasificatorio y jerárquico que actualmente reciben.

¿Cómo lograrlo? Diseñemos los cuestionarios o preguntas solicitando como respuestas valores en una escala del 1 al 10, donde 1 significa lo peor, lo más malo, lo pésimo, lo incalificable; y 10 significa lo mejor, lo excelente, lo máximo; o sea, los extremos de la opinión. Esto indicaría que si una respuesta recibe la calificación de 8, 7 ó 5, el investigador dará la interpretación según su marco de referencia en relación con la exigencia del estudio o necesidades de la investigación.

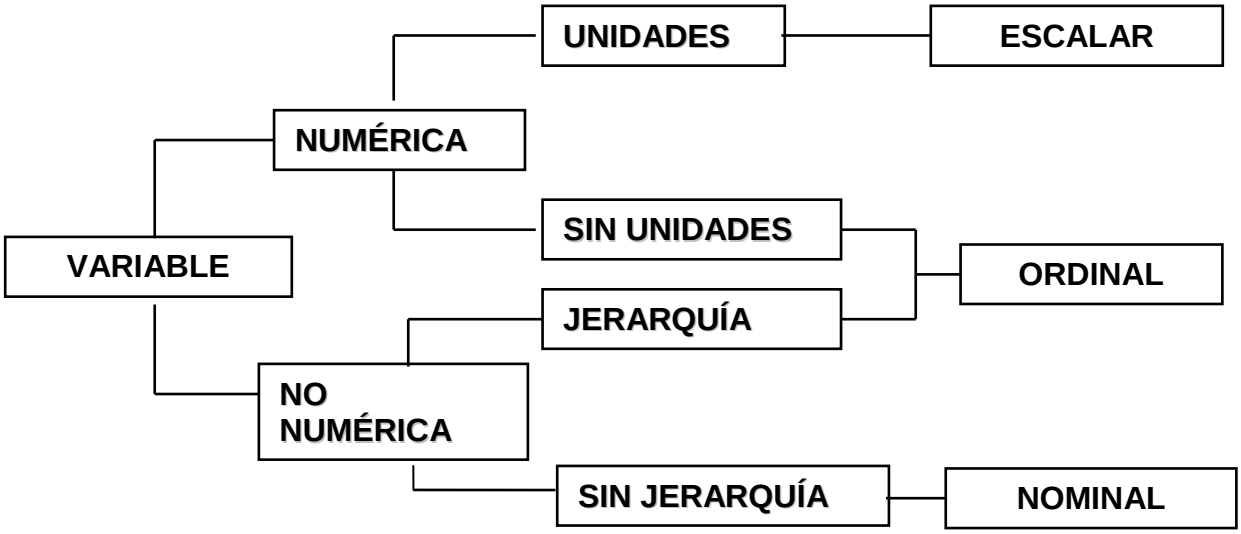
La gran ventaja que ofrece esta recomendación es que los datos y variables, concretamente el fenómeno en estudio, podrá ser medido en cualquier nivel y con los alcances que sean necesarios, ya que podrán obtenerse medidas como promedios que indiquen la concentración de los datos, su dispersión, su tendencia, sus niveles de concentración y distribución, su relación con otras variables, pudiendo con ello describirlos, compararlos y explicar su comportamiento en relación con diversas variables inmersas en el contexto de la investigación y más.



En la figura anterior se observan algunas equivalencias que podrían hacerse a los datos numéricos y el tratamiento tanto nominal clasificatorio como ordinal jerárquico, lo que flexibilizaría el análisis de un fenómeno social dándole mayor trascendencia a la información que se obtendría al respecto.

Ahora bien, si se desea continuar con el tratamiento clásico de datos en estadística social, el siguiente esquema facilitará la identificación del tipo de variable por nivel de medición y así seleccionar los métodos y técnicas estadísticas que les correspondan.

Identificación de variables por nivel de medición



UNIDAD 3

ORGANIZACIÓN DE DATOS.

INTRODUCCIÓN.

Para la descripción de un conjunto de datos se dispone de dos aspectos, gráfico y numérico, lo que permite identificar el comportamiento de las variables en estudio, su forma, su tendencia; algunos de sus rasgos más sobresalientes y complementariamente su dimensión. Esto se simplifica cuando los datos son sometidos a un proceso de organización, destacando en éste la ordenación de los datos, su clasificación ya sea por intervalos o por atributos y su distribución absoluta, porcentual o relativa; según las necesidades de la investigación.

Asimismo, la representación gráfica ofrece una alternativa sustancial para la descripción, empleando para ello diversas formas que son utilizadas para un mayor impacto y complemento de la información descrita al respecto.

OBJETIVO.

Al finalizar la unidad, el alumno tendrá la habilidad para organizar los diferentes tipos de variables y datos en estudio, simplificando con ello el proceso de análisis y presentación de resultados.

TEMARIO.

3. ORGANIZACIÓN DE DATOS.
    - 3.1 Etapas del proceso estadístico de una investigación.
    - 3.2 Fuentes de recolección de datos.
    - 3.3 Aspectos más importantes para describir un conjunto de datos.
    - 3.4 Organización de datos.
        - 3.4.1 Ordenación de datos.
        - 3.4.2 Clasificación de datos.
            - 3.4.2.1 Intervalo Cerrado.
            - 3.4.2.2 Intervalo abierto.
            - 3.4.2.3 Rango
        - 3.4.3 Distribución de datos.
            - 3.4.3.1 Frecuencia Absoluta.
            - 3.4.3.2 Frecuencia relativa.
            - 3.4.3.3 Frecuencia Porcentual.
            - 3.4.3.4 Frecuencia Acumulada.
    - 3.5 Marca de clase.
    - 3.6 Representaciones gráficas de la organización de datos.
        - 3.6.1 Histograma.
        - 3.6.2 Polígono de Frecuencias.
        - 3.6.3 Diagrama de Barras.
        - 3.6.4 Gráfica Sectorial.
        - 3.6.5 Ojiva.

3. ORGANIZACIÓN DE DATOS.

3.1 ETAPAS DEL PROCESO ESTADÍSTICO DE UNA INVESTIGACIÓN.

Para realizar una investigación sustentada en las herramientas estadísticas, deberá seguirse un proceso sencillo, que al mismo tiempo facilitará el análisis de los resultados obtenidos. El proceso se integra con las siguientes etapas:

- PLANEACIÓN.
- OBJETIVO DE LA INVESTIGACIÓN.
- RECOLECCIÓN DE DATOS. Mediante el empleo de encuestas, entrevistas o diversas fuentes de observación.
- ORGANIZACIÓN DE DATOS. Siguiendo las etapas de ordenación, clasificación y distribución.
- PROCESAMIENTO DE DATOS. Con lo que habrán de producirse tablas, gráficas y diversas medidas numéricas, para realizar el análisis correspondiente, que en un primer plano, se limita a la descripción.
- ANÁLISIS DE RESULTADOS. Empleando para ello diversos métodos y técnicas estadísticas.
- CONCLUSIONES. Lo que deriva en la presentación de la información y planteamiento de alternativas para la toma de decisiones.
- TOMA DE DECISIONES. Etapa en la cual se pondera la relatividad de cada factor, y del costo que implica la selección de alternativas: económicas, oportunidades, recursos, impacto, y trascendencia.

3.2 FUENTES DE RECOLECCION DE DATOS.

En las ciencias sociales debe considerarse que los datos observados serán determinantes en la calidad de información que habrá de obtenerse, razón que obliga a elegir la fuente más confiable para la recolección de datos, que entre

otras son: encuestas, entrevistas, grupos de enfoque, documentales y observación directa.

La elección de la fuente óptima de observación o recolección de datos marcará la diferencia entre la confiabilidad de la investigación y la calidad de los resultados obtenidos, por lo que se deberá tener en cuenta las características de la población objeto de estudio, así como las técnicas de levantamiento de datos, sin soslayar el objetivo de la investigación y el tratamiento estadístico que habrá de dárseles.

3.3 ASPECTOS MÁS IMPORTANTES PARA DESCRIBIR UN CONJUNTO DE DATOS.

En este apartado se consideran dos aspectos sustanciales para describir un conjunto de datos; uno de ellos, el gráfico, mediante el cual se observa la forma de la distribución, algunos de los rasgos más sobresalientes y la tendencia de la variable o fenómeno, entre otros, por lo que deberá elegirse el tipo de representación gráfica más adecuada al tipo de datos observados.

El segundo aspecto es el numérico, el cual permite observar la dimensión de los datos en estudio por medio de porcentajes, índices y diversas medidas numéricas de la estadística como las de tendencia central, dispersión, posición, distribución y correlación; considerando también sustancial la proyección y pronóstico de ocurrencia de los fenómenos en estudio y su tendencia prospectiva.

La única justificación para recopilar datos es que estos se van a utilizar para un propósito específico. Aquí surge la importancia de que sean seleccionados efectivamente, pues son la parte fundamental para el análisis del fenómeno en estudio a través del uso de técnicas, métodos, pruebas y procedimientos estadísticos. De los datos pueden obtenerse determinados indicadores que casi siempre pueden tratarse de manera numérica, según se comentó en la unidad anterior.

3.4 ORGANIZACIÓN DE DATOS.

DISTRIBUCIÓN DE FRECUENCIAS

Toda vez que los datos de una investigación han sido recopilados, es necesario y recomendable que estos sean sometidos a un proceso de organización, el cual consiste en tres etapas: ordenación, clasificación y distribución.

3.4.1 Ordenación de Datos.

En esta primera etapa los datos puedes pueden ser ordenados creciente o decrecientemente, decisión que toma el investigador. Recordemos que en ciencias sociales las variables principalmente son cualitativas, por lo que esta etapa deberá obviarse, si es el caso. Sin embargo, si las variables son numéricas como también se recomienda, se sugiere que se observen el mayor y el menor de los datos y se resten entre sí para obtener el rango, medida que ayudará en las etapas complementarias del proceso

EJEMPLO: si se tiene los siguientes datos, correspondientes a la evaluación del desempeño de un grupo de empleados: 67, 87, 56, 89, 95, 87, 75, 69, 93; obsérvese que el menor es 56 y el mayor 95, en este sentido podría afirmarse que ya los hemos ordenados crecientemente. Al restarlos se obtiene 39, lo que identificamos como el Rango.

Con el siguiente caso se ilustrará y explicará todo el proceso para la organización de los datos.

Caso en estudio:

Los datos de la siguiente tabla corresponden a los salarios de un grupo de 60 personas que, por diversas razones, fueron desplazados de su fuente de trabajo y han iniciado un proceso de demanda en contra de sus empresas. La relación laboral de todos ellos con sus centros de trabajo fluctúa entre 5 y 8 años. El “Salario inicial” indica las percepciones que cada uno de ellos recibió durante los primeros tres meses de ese periodo y el “Salario final” la de los

últimos seis meses de relación con su empresa, ambas expresadas en miles de pesos.

Salario inicial	Salario final	Salario inicial	Salario final	Salario inicial	Salario final
27	57	14	42	12	22
18	40	11	26	14	30
12	21	15	38	16	34
13	21	12	27	15	35
21	45	11	24	15	45
13	32	9	16	13	25
18	36	9	21	15	27
9	21	12	31	13	26
12	27	27	60	14	28
13	24	14	32	15	30
16	30	15	42	9	22
12	28	11	31	21	48
14	27	13	29	21	45
16	35	15	31	20	41
13	27	15	36	18	54
15	40	9	19	10	26
14	46	11	23	16	33
22	56	13	30	14	21
14	28	9	15	10	30
13	27	15	28	13	28

Para ejemplificar el proceso, se dará tratamiento a la variable “Salario inicial”, dejando al lector la tarea de hacerlo con la variable “Salario final”.

Como se indicó, el primer paso es ordenar los datos, por lo que al observar la variable “Salario inicial” se tiene que el menor de ellos es 9 y el mayor 27, esto significa que el rango de salarios recibidos en los primeros 3 meses de la relación laboral, va de 9 mil a 27 mil pesos. Al restarlos se obtiene un rango de 18 mil pesos. Con esto se ha realizado la primera etapa.

3.4.2 Clasificación de Datos.

Una vez ordenados los datos pasamos a la **segunda etapa:** clasificación. Si los datos son no numéricos, la categorización es natural, pero si los datos son numéricos, como en este ejemplo, entonces la clasificación deberá hacerse mediante intervalos, llamados intervalos de clase; para ello habrá que responder dos preguntas: ¿cuántos intervalos se necesitan? y ¿cuál es su amplitud? La amplitud es la diferencia entre el valor mayor y el menor de cada intervalo, también llamados límite superior e inferior respectivamente, algo similar al rango mencionado.

Existen dos tipos de intervalos:

3.4.2.1 Intervalo Cerrado.

En el que se conoce tanto su límite inferior como su límite superior.

Ejemplo: Personas de 15 a 20 años de antigüedad y con estatura de 1.60 a 1.75 metros.

3.4.2.2 Intervalo Abierto.

En el que uno de sus límites se desconoce o se excluye.

Ejemplo: Personas menores de 35 años y estatura mayor de 1.60 metros.

3.4.2.3 Rango.

Es la diferencia entre los límites de un intervalo.

El rango se puede subdividir en otros rangos, intervalos de clase, según el procedimiento y criterios utilizados para la organización de datos o distribución de frecuencias; así, cada intervalo de clase representa un rango.

Para determinar el número de intervalos de clase, debe considerarse alguno de los siguientes criterios:

- **Primer criterio:** Que las necesidades de la investigación indiquen la cantidad de subgrupos que deberán integrarse para el análisis.

- **Segundo criterio:** Utilizar alguna herramienta estadística para calcular el número de intervalos en función del total de datos del conjunto. Una herramienta recomendada en este caso es la fórmula de Sturges:

$$K = 1 + 3.322 \log_{10} N \qquad \text{donde N = número de datos.}$$

Si se utiliza esta fórmula con los 60 datos del ejemplo, el logaritmo en base 10 de ese total es igual a 1.78 que al ser sustituido en esa expresión aritmética reporta el siguiente resultado:

$$K = 1 + 3.322 (1.78) = 6.91$$

Este resultado se redondea al entero superior lo que indica que tendríamos que clasificar los 60 datos del ejercicio en 7 intervalos de clase.

- **Tercer criterio:** Que el investigador, usted, determine cuántos intervalos considera necesarios para realizar su análisis, sin perder de vista que muchos intervalos provocan dispersión en los datos y pocos evidencian carencia de forma, que finalmente no ofrecen información suficiente.

Continuando con el ejemplo y utilizando el primero de los criterios mencionados, supóngase que la investigación requiere de 6 intervalos para clasificar a los 60 demandantes, con lo que se responde la primera pregunta planteada anteriormente. Para contestar la segunda pregunta, dividiremos el rango entre el número de intervalos indicado, dando como resultado la amplitud de cada uno de ellos:

$$\text{Amplitud de los intervalos} = \text{Rango} / \text{Número de intervalos}$$

Amplitud de los intervalos = 18 / 6 = 3

Esto indica que cada uno de los 6 intervalos tendrá una amplitud de 3 unidades, recordando que los datos están expresados en miles de pesos, esto es, 3 mil pesos de amplitud en cada uno de ellos.

Por lo tanto, los intervalos de clase se formarán sumando 3 a partir del valor más pequeño del conjunto; recuérdese que el menor es 9 en este caso, por lo tanto los intervalos quedarán como sigue:

Intervalos de clase		
9	-	12
12	-	15
15	-	18
18	-	21
21	-	24
24	-	27

Esta clasificación es llamada continua, debido a que el límite superior de cada intervalo tiene el mismo valor que el límite inferior del intervalo siguiente. Este valor es llamado frontera o límite real de clase.

Una clasificación discreta es aquella en la que el límite superior de un intervalo, en una tabla de organización de datos, es diferente al límite inferior del siguiente, como se aprecia en la tabla:

Intervalos de clase		
9	-	12
13	-	15
16	-	18
19	-	21
22	-	24
25	-	27

3.4.3 Distribución de Datos.

La tercera etapa consiste en distribuir los datos, según sea el caso: numéricos o no numéricos; en cada uno de los intervalos o categorías. Los diferentes tipos de distribución son los siguientes:

3.4.3.1 Frecuencia Absoluta.

Consiste en determinar cuántos elementos del conjunto pertenecen a cada intervalo o categoría. La suma de las frecuencias absolutas debe ser igual al número total de datos en estudio.

Para determinar las frecuencias absolutas, también llamadas frecuencias de clase, se cuentan los elementos del conjunto que se encuentran dentro de los límites de cada intervalo.

3.4.3.2 Frecuencia Relativa.

Consiste en determinar la proporción de elementos en cada intervalo o categoría. La suma de las frecuencias relativas es siempre la unidad.

La frecuencia relativa o proporcional se obtiene al dividir la frecuencia absoluta de cada intervalo entre el numero total de datos o elementos del conjunto.

Frecuencia Relativa FR = Frecuencia absoluta / total de frecuencias

Donde F = frecuencia absoluta de un intervalo y N= es el total de elementos del conjunto.

3.4.3.3 Frecuencia Porcentual.

Muy utilizada en ciencias sociales: consiste en calcular el porcentaje que, del total de elementos del conjunto, pertenece a cada intervalo o categoría.

Al organizar los datos, los porcentajes pueden obtenerse de las proporciones multiplicando por 100. La palabra porcentaje significa de cada cien, esto es: una distribución porcentual se puede calcularse a partir de la distribución relativa, multiplicando cada una de ellas por 100.

Porcentaje = FR X 100 Frecuencia relativa multiplicada por 100.

El porcentaje también se calcula dividiendo la frecuencia absoluta del intervalo entre el total de elementos del conjunto y el resultado multiplicarlo por 100.

$$\text{Porcentaje} = (F / N) \times 100$$

F es la frecuencia de clase y N es el total de datos del conjunto.

Recordemos que el proceso de organización de datos es generalmente nombrado distribución de frecuencias, cuyo concepto nos remite a identificarla como un agrupamiento de datos en categorías mutuamente excluyentes

indicando el número de observaciones, proporciones o porcentajes de cada intervalo o categoría.

Continuando con el ejemplo de las 60 personas desplazadas de su fuente de trabajo, realizaremos la etapa de distribución, en la que incluiremos tanto frecuencias absolutas como porcentuales y relativas. Los cálculos de las frecuencias relativas y porcentuales se realizaron según se explicó anteriormente.

Intervalos de clase	Frecuencia absoluta	Frecuencia_relativa	Frecuencia_porcentual
9 - 12	18	0.30	30
12 - 15	28	0.46	46
15 - 18	7	0.12	12
18 - 21	4	0.07	7
21 - 24	1	0.02	2
24 - 27	2	0.03	3
Totales	60	1.0	100

Es menester precisar que a pesar de que la clasificación se realice de manera continua, como en este caso, o discreta, finalmente la distribución de los datos del conjunto será discreta, lo que significa que cada uno de ellos solamente estará en uno y en un sólo intervalo. Por lo tanto, en la siguiente tabla se observa la forma en que se realizaron realmente la segunda y tercera etapa del proceso de organización de los datos del ejemplo:

Intervalos de clase Con clasificación continua	Intervalos de clase con clasificación discreta	Frecuencia absoluta	Frecuencia relativa	Frecuencia porcentual
9 - 12	9 - 12	18	0.30	30
12 - 15	13 - 15	28	0.46	46
15 - 18	16 - 18	7	0.12	12
18 - 21	19 - 21	4	0.07	7
21 - 24	22 - 24	1	0.02	2
24 - 27	25 - 27	2	0.03	3
Totales		60	1.0	100

Para aclarar el contenido de la tabla anterior, la primera columna indica que la clasificación de los datos se hizo continua; sin embargo, la distribución se realizó de manera discreta, según la segunda columna, lo que indica, por ejemplo, que si uno de los datos a distribuir es igual a 12, se ha incluido sólo en el primer intervalo. Si uno de ellos es igual a 15, éste se distribuyó en el

segundo intervalo y así sucesivamente, lo que hace que cada intervalo sea totalmente excluyente e independiente a los demás.

3.4.3.4 Frecuencia Acumulada.

Para cierto tipo de análisis o necesidades de información, resulta menester realizar otro forma de distribución que indique cómo se van concentrando los datos hasta determinados valores o límites de la tabla de organización de datos, esta distribución es llamada acumulada y puede incluir a cualquiera de las frecuencias: absoluta, relativa o porcentual; sugiriendo se calcule sólo la que sea necesaria para los fines de la investigación.

Entiéndase como frecuencia acumulada de un dato a la suma de la frecuencia de este dato con las frecuencias de todas las anteriores.

Siguiendo nuestro ejercicio, en la tabla de abajo se calculó la frecuencia acumulada porcentual o frecuencia porcentual acumulada, explicando después de ella, alguna de sus interpretaciones.

Intervalos de clase Con clasificación continua			Frecuencia absoluta	Frecuencia porcentual	Frecuencia porcentual acumulada
9	-	12	18	30	30
12	-	15	28	46	76
15	-	18	7	12	88
18	-	21	4	7	95
21	-	24	1	2	97
24	-	27	2	3	100
Totales			60	100	

Como se refirió anteriormente, por necesidades de la investigación se requiere de cierta información; en este caso supongamos que se demanda saber cuál es el comportamiento de los salarios hasta cierto límite de ingresos, por lo que la columna de la frecuencia porcentual acumulada que se observa en la tabla de arriba, nos ayudará a obtener esa información:

Sólo el 30% de los demandantes tuvo ingresos máximos de 12 mil pesos durante los tres primeros meses de servicio en su empresa. Asimismo, se aprecia en la columna que se indica de la tabla, que el 76% tuvo ingresos máximos de 15 mil pesos durante el mismo periodo. Al comparar estos dos resultados, se concluye que casi la mitad (46%) de esas personas tuvieron ingresos mensuales entre 4 y 5 mil pesos, en los primeros 3 meses de trabajo, lo que permite considerar el enfoque que cada una de las partes involucradas en el caso daría a esa información.

Las etapas referidas en este apartado corresponden al proceso de organización de datos o distribución de frecuencias.

3.5 MARCA DE CLASE.

Una vez que los datos han sido organizados, están listos para su descripción gráfica o numérica, lo que representa la primera etapa de información en la investigación. Dado que los datos se han agrupado en intervalos de clase, se ha perdido cierto control en la dimensión de cada uno de ellos ya que al encontrarse concentrados en un rango de valores, su manejo implicará diferencias de magnitud en lo que es llamado “Error por agrupamiento”. Sin embargo, este error es considerado y en múltiples ocasiones minimizado, al obtenerse un valor que representará durante todo el tratamiento que se le dé al conjunto. Este valor es llamado **Marca de Clase** y se calcula sumando los límites inferior y superior de cada intervalo, dividiendo entre dos el resultado de la suma:

$$\text{Marca de clase} = (\text{Límite inferior} + \text{Límite superior}) / 2$$

En otras palabras, la marca de clase es el punto medio de cada intervalo de clase.

La marca de clase es el valor más característico y representa a todos los datos que puedan estar integrados en éste.

En la tabla de nuestro ejemplo, se han calculado, de la manera indicada, las marcas de clase de cada uno de los intervalos de la distribución, mismas que serán utilizadas para cálculos posteriores.

Intervalos de clase Con clasificación continua			Marca de Clase X	Frecuencia absoluta
9	-	12	10.5	18
12	-	15	13.5	28
15	-	18	16.5	7
18	-	21	19.5	4
21	-	24	22.5	1
24	-	27	25.5	2
Totales				60

3.6 REPRESENTACIONES GRÁFICAS DE LA ORGANIZACIÓN DE DATOS.

Como ha sido referido, para describir un conjunto de datos, el investigador podrá valerse de los aspectos gráfico y numérico; el gráfico le indicará la forma y el numérico la dimensión de los datos en observación. Con la tabla de organización de datos ha obtenido los elementos numéricos básicos para su descripción desde el punto de vista porcentual, por lo menos; con los que puede generar información que oriente el sentido de la investigación, en la búsqueda de explicación en el comportamiento de las variables en estudio.

Para la descripción gráfica, podrá disponer de una amplia galería representada en este material por Histogramas, Polígonos de Frecuencias, Diagramas de Barras, Gráficas Sectoriales o de Pastel y Ojivas, entre muchas más.

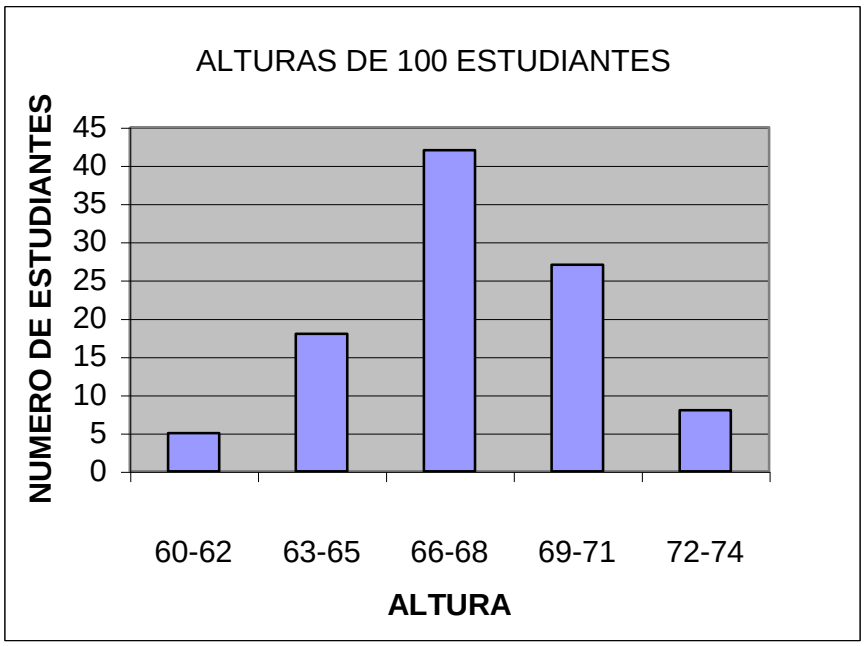
3.6.1 Histograma.

Es la representación gráfica de variables numéricas organizadas en tablas de frecuencias. Consiste en una serie de rectángulos que tienen sus bases sobre el eje horizontal con longitud igual al tamaño de los intervalos de clase y altura correspondiente a la frecuencia: absoluta, porcentual o relativa, según sea el interés o necesidades del investigador.

De otra manera, es la representación gráfica de la tabla de frecuencias; éste muestra datos cuantitativos. Los intervalos de clase, que pueden ser o no ser iguales, están marcados sobre el eje horizontal. Las frecuencias son marcadas sobre el eje vertical. Se construye por medio de rectángulos unidos cuyos anchos son los de los intervalos de clase que ellos representan, cuyas alturas representan a las frecuencias.

En el siguiente ejemplo se muestran las alturas en pulgadas de 100 estudiantes de la UNAM:

ALTURA EN PULGADAS	NUMERO DE ESTUDIANTES
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
Total	100



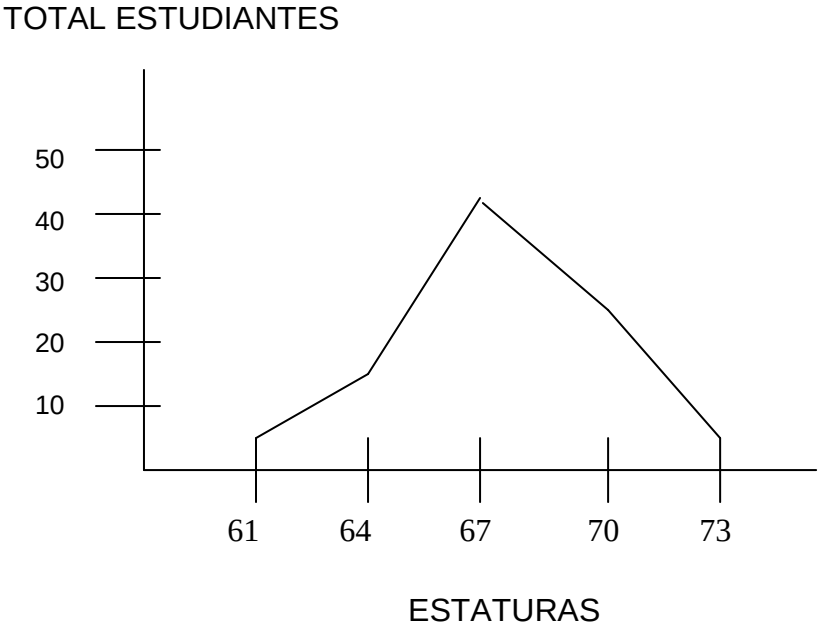
La organización de los datos de los 100 estudiantes fue realizada con una clasificación discreta, por lo que en el histograma se observa que entre una barra y otra no existen las fronteras o límites reales mencionados en la etapa de clasificación del apartado correspondiente a la organización de datos.

3.6.2 Polígono de Frecuencias.

Gráfico que une los puntos obtenidos entre las marcas de clase y la frecuencia correspondiente en una distribución. En el ejemplo de los 100 estudiantes, esta gráfica partirá de la siguiente tabla:

ALTURA EN PULGADAS	MARCA DE CLASE	NUMERO DE ESTUDIANTES
60-62	61	5
63-65	64	18
66-68	67	42
69-71	70	27
72-74	73	8
TOTAL		100

Por lo que su polígono de frecuencias quedará de la siguiente forma:



En los paquetes de computación como SPSS y Excel, esta gráfica es llamada de áreas.

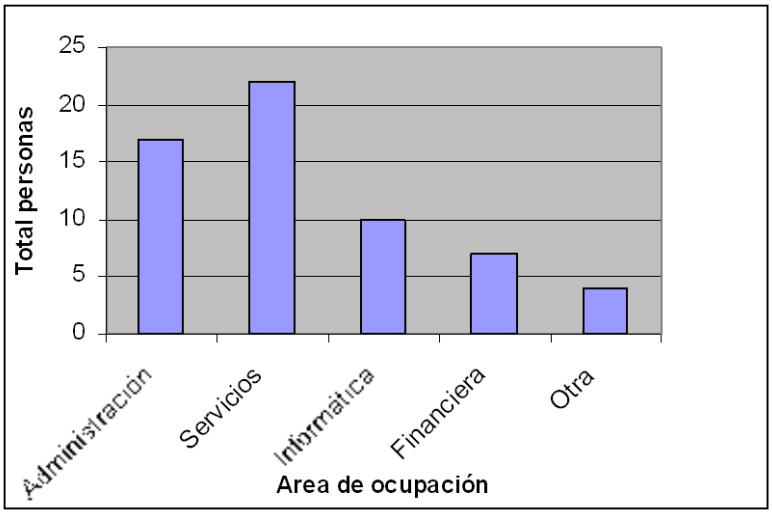
3.6.3 Diagrama de Barras.

Esta gráfica es empleada para representar variables no numéricas o categóricas, recuerde las variables con nivel de medición nominal como sexo, religión, nacionalidad. Es la forma más sencilla y económica de representación de datos; consiste en un conjunto de barras separadas por un espacio. Cada barra tiene en su base la categoría que representa, hombre y mujer en el caso de la variable sexo, y como altura la frecuencia respectiva: absoluta, porcentual o relativa.

En la tabla siguiente se presenta la distribución de las áreas en que prestaban sus servicios los 60 exempleados de nuestro ejemplo:

ÁREA DE OCUPACIÓN	NÚMERO DE PERSONAS
Administración	17
Servicios	22
Informática	10
Financiera	7
Otra	4

Su representación gráfica a partir de un diagrama de barras tiene la siguiente forma:



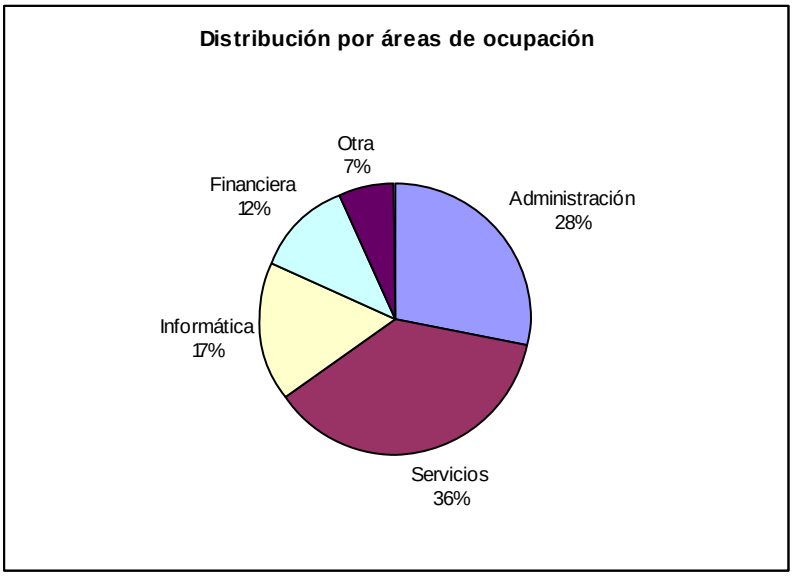
3.6.4 Gráfica Sectorial.

Esta gráfica es utilizada para representar principalmente variables no numéricas y es también llamada gráfica de pastel, ya que la distribución de frecuencias, aparenta la repartición de rebanadas de un delicioso pastel. En este caso lo que se distribuye son los 360 grados de una circunferencia con la representación proporcional que le corresponde según la frecuencia respectiva.

En la siguiente tabla se muestra la distribución de los 360 grados, con cifras redondeadas, de una circunferencia, según el porcentaje que le corresponde a cada categoría de la variable del ejemplo de las áreas en que prestaban sus servicios los 60 exempleados de nuestro ejercicio:

ÁREA DE OCUPACIÓN	NÚMERO DE PERSONAS	PORCENTAJE	GRADOS QUE LE CORRESPONDEN
Administración	17	28	101
Servicios	22	36	130
Informática	10	17	61
Financiera	7	12	43
Otra	4	7	25
Total	60	100	360

Con los datos de esta tabla elaboraremos la gráfica sectorial, misma que tendrá la siguiente forma:



¿Cómo se calculó la distribución de los 360 grados respecto a los porcentajes?

Tomaremos como ejemplo la categoría de Administración, a la cual le corresponde el 28% del total de personas. Para ello se utiliza la siguiente expresión aritmética:

$$\text{Total de grados} = (\text{porcentaje})(360)$$

$$\text{Total de grados} = (28\%)(360) = 100.8$$

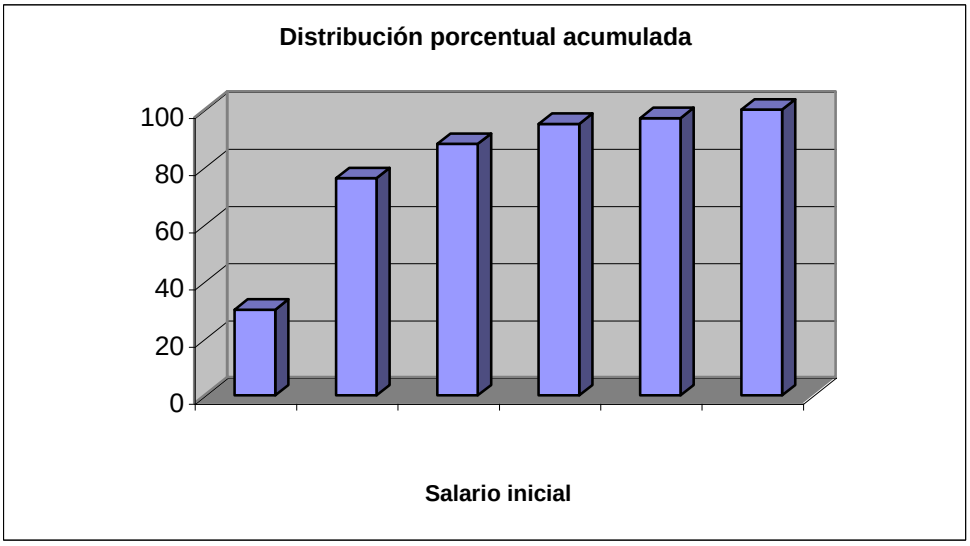
Redondeando el resultado para trabajar con datos discretos, tenemos que el total que le corresponde a esa categoría de ocupación es 101 grados, conforme se aprecia en la tabla de arriba.

3.6.5 Ojiva.

Ésta es una gráfica que representa a la distribución de frecuencias acumuladas, sean absolutas, porcentuales o relativas. Es aplicable a variables numéricas y variables jerárquicas u ordinales; reiterando la sugerencia de que, en lo posible, las variables en ciencias sociales sean utilizadas como numéricas, lo que simplificará todo proceso, con la facilidad de que pueden ser convertidas a cualquier escala de medición, según lo sugerido en la segunda unidad de este trabajo.

Dado que la Ojiva representa la distribución de frecuencias acumuladas, es una gráfica ascendente o descendente, según el orden que se le dé a la organización de los datos. En nuestro ejemplo el orden es ascendente, por lo que la Ojiva será creciente.

INTERVALOS DE CLASE	FRECUENCIA ABSOLUTA	FRECUENCIA PORCENTUAL	FRECUENCIA PORCENTUAL ACUMULADA
9 - 12	18	30	30
12 - 15	28	46	76
15 - 18	7	12	88
18 - 21	4	7	95
21 - 24	1	2	97
24 - 27	2	3	100
Totales	60	100	


Distribución porcentual acumulada
100
80
60
40
20
0
Salario inicial

UNIDAD 4

MEDIDAS DE TENDENCIA CENTRAL O PROMEDIOS.

INTRODUCCIÓN.

Analizar un conjunto de datos inmersos en un fenómeno social, implica describir sus niveles de representación general, con sus consideraciones, así como los valores o características de mayor concentración o de interés para el investigador. Esto es observado a partir de la necesidad numérica para dimensionar el impacto de la variable, utilizando para ello medidas como la media, mediana, moda y su relación con las escalas de medición.

OBJETIVO

El alumno conocerá las medidas de tendencia central aplicables en los fenómenos sociales, y utilizará la más adecuada al tipo de variable en estudio, considerando las ventajas y desventajas que ofrece cada una de ellas.

TEMARIO

4. MEDIDAS DE TENDENCIA CENTRAL.
    - 4.1 Conceptos.
    - 4.2 Media.
    - 4.3 Media Ponderada.
    - 4.4 Mediana.
    - 4.5 Moda.
    - 4.6 Relación que guardan la Media, Mediana, y Moda con las variables por nivel de medición.
        - 4.6.1 Escalas nominales.
        - 4.6.2 Escalas Ordinales.
        - 4.6.3 Escalas de Intervalos y de proporción.
    - 4.7 Algunos ejemplos de la relación entre niveles de medición y medidas de tendencia central.

4. MEDIDAS DE TENDENCIA CENTRAL O PROMEDIOS.

4.1 CONCEPTOS.

Estas medidas son valores o características sobre las cuales tienden a concentrarse la mayor parte de los elementos de un conjunto. Estas son representadas principalmente por la media, mediana y moda. Aquí algunos conceptos:

a) Según **Ya-Lun Chou**: *"Las medidas de tendencia central se llaman promedios. Un promedio es un valor típico en el sentido de que se emplea a veces para representar todos los valores individuales de una serie o de una variable".*

b) De acuerdo con **Herbert Arkin**: *"Un promedio es un valor típico con el que se intente resumir o describir una masa de datos. También sirve como una base para medir o evaluar valores extremos o poco usuales. El promedio es una medida de localización del punto de tendencia central".*

c) **Frederick E. Croxton** afirma que: *"Se usa la expresión medidas de dispersión o promedios para identificar aquellos valores que pueden calcularse con el fin de caracterizar la distribución de las frecuencias".*

d) ***Samira García Durán****: "Los promedios son medidas de tendencia central. Son valores que sirven para representar alguna característica de un determinado grupo de datos".*

4.2 MEDIA.

La media es el promedio de mayor uso; sin embargo, es aplicable únicamente a variables numéricas. Se calcula sumando los datos y dividiendo el resultado entre el total de ellos. Tiene la ventaja de que para su cálculo toma en cuenta a todos los elementos del conjunto, pero con la gran desventaja de que, en la medida de que el rango se vaya

haciendo mayor, la media va perdiendo fuerza o representatividad del conjunto.

Ventajas:

La media es el promedio utilizado más frecuentemente y es sencillo de entender. Su cálculo es simple. Todos los elementos del conjunto participan en la obtención de este promedio.

Desventajas:

Resulta afectada por el alejamiento de los valores extremos del conjunto de datos, rango. Es alterada también según el desplazamiento de los datos del conjunto, esto es: si los datos tienden hacia el extremo inferior o superior del conjunto, entonces la media se dirigirá en ese sentido, lo cual representa una desventaja significativa en relación con los otros promedios o medidas de tendencia central.

Por lo tanto, la media puede quedar fuertemente distorsionada por valores extremos, y por ello no ser un valor representativo del conjunto de datos.

La media no puede calcularse en las distribuciones que contienen intervalos abiertos, es decir, cuando se desconoce alguno de sus límites.

Particularmente, la **media aritmética** o promedio de una cantidad determinada de datos numéricos, es igual a la suma de las magnitudes de cada uno dividida entre el número de ellos.

Así, dados los números $a_1, a_2, \ldots, a_n$, la media aritmética será igual a:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i = (a_1 + \cdots + a_n)/n$$

Por ejemplo, la media aritmética de los siguientes datos 18, 25, 32, 35 y 15 es igual a la suma de todos esos valores, divididos entre cinco ya que éste es el número de elementos de ese conjunto.

$$(18 + 25 + 32 + 35 + 15) / 5 = 25$$

Para un conjunto de datos agrupados deberá multiplicarse cada dato por su frecuencia de clase y la suma de esos productos será dividida entre el total de datos:

$$\textbf{MEDIA} = (\Sigma F X) / \Sigma F$$

Donde: Σsumatoria

F = Frecuencia absoluta de cada intervalo

X = Marca de clase de cada intervalo

Para ejemplificar el cálculo de la media aritmética para un conjunto de datos agrupados, continuaremos con el ejemplo de las 60 personas desplazadas de su empleo. Para ello, en la siguiente tabla se calculó la marca de clase y se obtuvo el producto de cada una de ellas por la frecuencia respectiva. La suma de esos productos se dividieron entre el total de frecuencias, obteniendo así la media aritmética, promedio, del conjunto:

INTERVALOS DE CLASE	MARCA DE CLASE X	FRECUENCIA ABSOLUTA F	FX
9 - 12	10.5	18	189
12 - 15	13.5	28	378
15 - 18	16.5	7	115.5
18 - 21	19.5	4	78
21 - 24	22.5	1	22.5
24 - 27	25.5	2	51
Totales		60	834

Media aritmética = 834 / 60 = **13.9**

Lo que significa, a cifras cerradas, que el ingreso promedio de los empleados, durante los 3 primeros meses de ejercicio laboral, es de 14 mil pesos, por lo que debe considerarse qué tan significativo resulta ese ingreso medio en el proceso que han iniciado esas personas en contra de sus empresas y, por supuesto, cuál es la percepción de los patrones al respecto.

4.3 MEDIA PONDERADA.

Cuando un conjunto de datos (x_1, x_2,...,x_n) son relativamente semejantes, para promediar, ciertos factores (w_1, w_2,...,w_n) dependen de la importancia o peso específico de cada uno de los valores. En estos casos se recomienda calcular la llamada media aritmética ponderada, la cual considera el peso relativo que tiene cada uno de esos factores.

Supóngase que con la finalidad de tener más elementos para determinar la cantidad de dinero que habrá de recibir cada uno de esos demandantes, se desea encontrar el promedio ponderado de las cinco calificaciones que se consideraron para evaluar su desempeño en los últimos 3 meses de trabajo. Para ello, la segunda calificación vale el doble de la primera, la tercera el triple de la primera, la cuarta vale cuatro veces la primera y la quinta cinco veces.

Si una de esas personas demandantes de justicia laboral fue evaluada con los siguientes puntajes: 8.5, 7.3, 8.3, 6.4 y 9.2 ¿Cuál es el promedio de su evaluación?

Solución X= calificación W = importancia o peso relativo de la calificación

$X_1 = 8.5 ; W_1 = 1$

$X_2 = 7.3 ; W_2 = 2$

$X_3 = 8.3 ; W_3 = 3$

$X_4 = 6.4 ; W_4 = 4$

$X_5 = 9.2 ; W_5 = 5$

(8.5*1+7.3*2+8.3*3+6.4*4+9.2*5)/(1+2+3+4+5) = 119.6/15 = 7.97

7.97 es el promedio ponderado de las calificaciones de este empleado.

4.4 MEDIANA.

Esta medida es aplicable a variables numéricas y a variables ordinales. Es el valor que divide exactamente al conjunto de datos en dos partes iguales. Tiene la ventaja de que no es afectada por el rango, y la desventaja de que para su cálculo u observación, sólo toma en cuenta al valor o valores que están en el centro del conjunto o distribución.

Ventajas:

- La mediana para un conjunto de datos no agrupados, se obtiene fácilmente, bastará con ordenar los datos y dividir entre dos el total de ellos, localizando con esto el centro del conjunto, lugar donde se encuentra la mediana.
- Su valor no es afectado por los extremos del conjunto. Ocasionalmente es un valor más representativo de un grupo de datos que otros promedios, debido a que el rango no le afecta.
- Para conjuntos de datos agrupados por intervalos, la mediana puede calcularse aún cuando éstos sean abiertos.
- No se afecta por los valores de los extremos del conjunto.

Desventajas:

- No es tan aplicable como la media.
- Toma sólo los valores del centro del conjunto.

Una expresión simple para el cálculo de la mediana para un conjunto de datos agrupados es la siguiente:

$$M = L_1 + (((n+1)/2 – S) / F_M) \ C$$

Donde:

M = Mediana a obtener.

L_1 = Límite real inferior o valor frontera inferior del intervalo donde se encuentra la mediana (clase mediana).

n = Total de datos.

S = Frecuencia acumulada hasta el intervalo anterior al de la clase mediana.

F_M = Frecuencia de la clase mediana.

C = Amplitud del intervalo.

4.5 MODA.

La más simple y débil de las medidas de tendencia central. Es aplicable a todo tipo de variables y es aquel valor o característica que se presenta con mayor frecuencia en un conjunto. Lo mismo que la mediana, su ventaja es no verse afectada por el rango, en caso de datos numéricos, y con la desventaja de que sólo observa o considera a aquellos valores o características que se presentan el mayor número de veces.

Un conjunto de datos o una distribución de frecuencias puede tener más de un valor, intervalo o categoría que se presente con la mayor frecuencia, por lo que

las distribuciones con sólo un valor o categoría con esa característica son llamadas unimodales; no obstante puede haber algún conjunto en donde sean dos las categorías con la mayor frecuencia, siendo llamadas bimodales. En caso de que sean más de dos las categorías en esas condiciones, la distribución es llamada multimodal.

Ventajas:

- No es afectada por el rango. Es el valor de mayor frecuencia y por ello el promedio más descriptivo, según diversos investigadores de la materia, aunque como se señaló, es un promedio débil.
- Es sencillo aproximarla mediante un examen de los datos, cuando son reducidos.
- Si sólo hay unos pocos elementos no es necesario ordenarlos para determinar la moda.

Desventajas:

- Sólo considera a los datos que se presentan con mayor frecuencia.
- Se puede aproximar la moda, sólo cuando se dispone una cantidad limitada de datos.
- Su significación es limitada cuando no se dispone de un gran número de valores.
- Si no se repite ningún valor, la moda no existe.

Una expresión para el cálculo de la moda para un conjunto de datos agrupados es la siguiente:

$$Moda = Li + \left\langle \frac{\Delta 1}{\Delta 1 + \Delta 2} \right\rangle \mathbf{c}$$

Donde:

$\boldsymbol{L_i}$ **=** límite inferior de la clase modal, aquel intervalo en el que se encuentra la moda.

$\Delta 1 =$ diferencia entre la frecuencia de la clase modal y la frecuencia de la clase anterior a ella.

$\Delta 2 =$ diferencia entre la frecuencia de la clase modal y la frecuencia de la clase siguiente.

4.6 RELACIÓN QUE GUARDAN LA MEDIA, MEDIANA, Y MODA CON LAS VARIABLES POR NIVEL DE MEDICIÓN.

4.6.1 Escalas nominales.

Es la operación básica y más sencilla en toda ciencia es la de la clasificación. Separa elementos desde el punto de vista de determinadas características, decidiendo cuales son más semejantes y más distintos. Agrupa por categorías que sean lo mas homogéneas posible en comparación con la diferencia entre las categorías.

4.6.2 Escalas ordinales.

Se presente en un nivel superior al que empleamos para obtener la escala nominal, con esta escala podemos agrupara a individuos en categorías separadas y ordenarlos con respecto a otras. La escala ordinal es asimétrica en el sentido de que algunas relaciones especiales pueden ser verdad entre *A* y *B* y no serlo.

4.6.3 Escala de intervalos y de proporción.

Requiere el establecimiento de algún tipo de unidad física de medición que puede considerarse por todos como una norma común y sea repetible, esto es, que pueda aplicarse indefinidamente con los mismos resultados.

La moda se utiliza principalmente con variables nominales y es la única medida de tendencia central que se puede usar con variables nominales.

4.7 ALGUNOS EJEMPLOS DE LA RELACIÓN ENTRE NIVELES DE MEDICIÓN Y MEDIDAS DE TENDENCIA CENTRAL.

La mediana se utiliza principalmente con variables ordinales y junto con la moda son las únicas dos medidas de tendencia central que se puede usar con variables ordinales.

Sin embargo, cuando la variable es ordinal, no es apropiado promediar los dos valores medios. Simplemente se dice que la mediana se encuentra entre esos dos valores.

EJEMPLOS:

1. En un cuestionario que utiliza la escala Likert, las respuestas a una pregunta fueron “nunca, nunca, de vez en cuando, a menudo, muy frecuentemente”.

mediana = de vez en cuando

Cuando las observaciones han sido tabuladas en una tabla de distribución de frecuencias, la mediana corresponde a la categoría en la que se encuentra la frecuencia acumulativa del 50% de las observaciones.

UNIDAD 5

MEDIDAS DE POSICIÓN.

INTRODUCCIÓN.

Debido a la naturaleza de los conjuntos de datos, por diversas razones es necesario encontrar otras medidas que permitan realizar análisis más detallados de su comportamiento y niveles de concentración alrededor de ciertas características de las variables observadas.

En ciencias sociales es necesario observar determinadas características y valores de interés en un conjunto de datos, debido a que los promedios únicamente observan la tendencia al centro de la distribución, lo que impide profundizar en el análisis detallado de los datos.

Las medidas de posición resuelven ese problema y representan una alternativa interesante y efectiva para detallar los niveles y características de concentración de los datos. Estas medidas, representadas por los cuartiles, deciles y percentiles, dividen la distribución en 4, 10 y 100 partes iguales respectivamente, acumulando en cada una de ellas el 25%, 10% y 1%, en el mismo orden, del total de los elementos del conjunto.

Por otra parte, en ciencias sociales se observar ciertos valores o características de interés, sobre los cuales se concentran los datos observados; estos valores se identifican como medidas de posición entre las cuales destacan los cuartiles, deciles y percentiles.

OBJETIVO.

Al finalizar la unidad el alumno identificará las medidas de posición y su efectividad para detallar los niveles y características de los datos analizados.

TEMARIO.

5. MEDIDAS DE POSICIÓN.

5.1 CUARTILES.

Son 3 puntos o valores que dividen al conjunto en 4 partes iguales, concentrando en cada una de ellas el 25% del total de datos.

Q_1= Valor de la variable que acumula a su izquierda el 25% de la distribución.

Q_2= Valor de la variable que acumula a su izquierda el 50% de la distribución.

Q_3= Valor de la variable que acumula a su izquierda el 75% de la distribución.

5.2 DECILES.

Son 9 puntos o valores de la distribución que dividen al conjunto en 10 partes iguales, concentrando en cada una de ellas el 10% del total de datos.

D_1 = Valor de la variable que acumula a su izquierda al 10% de la distribución.

D_2 = Valor de la variable que acumula a su izquierda al 20% de la distribución.

D_3 = Valor de la variable que acumula a su izquierda al 30% de la distribución.

Y así sucesivamente hasta el D_9 que acumula a su izquierda al 90% de la distribución.

5.3 PERCENTILES.

Son 99 puntos o valores de la distribución que dividen al conjunto en 100 partes iguales, concentrando en cada una de ellas el 1% del total de datos y se representan como P_1, P_2, …, P_{99}

Para el cálculo de la medida de posición que particularmente se desea obtener, utilícese la fórmula de la mediana y adáptela según lo buscado:

La fórmula de la mediana es: $M = L_1 + (((n+1)/2 – S) / F_M) C$

Para encontrar cualquier posición en la distribución, se modifica **(n+1) / 2** en la fórmula de la mediana por la medida deseada, recuerde que M es mediana. Por ejemplo, si se requiere encontrar el primer cuartil, haga lo siguiente:

M será sustituida por Q_1.

(n + 1) / 2 se reemplazará, entienda la lógica del cambio, por **n / 4** debido a que se busca la primera cuarta parte del conjunto y ésta se encuentra en n/4. Si se deseara buscar el tercer cuartil Q_3, sería sustituida por **3n / 4** ya que el tercer cuartil localiza las tres cuartas partes de acumulación de los datos (75%).

Si se desea o necesita obtener cualquier medida de posición, siga la lógica de los dos párrafos anteriores.

UNIDAD 6

MEDIDAS DE DISPERSIÓN.

INTRODUCCIÓN.

Al analizar un conjunto de datos, resulta a menudo conveniente expresar numéricamente la variabilidad que existe entre ellos. Para llevar a cabo esta descripción, se usan varias estadísticas que usan relaciones internas entre los datos. Comúnmente estas relaciones tienen que ver con diferencias de los datos o funciones de ellos respecto de algunas estadísticas de posición. Dependiendo del tipo de diferencia usada, se obtienen distintas expresiones que entregan visiones parciales de la forma en que los datos varían. Estas visiones parciales se complementan para entregar un cuadro más completo de la dispersión observada entre los datos.

Las medidas de tendencia central son de un gran valor representativo para una masa de observaciones. Pero el valor de esas medidas dependerá de cuan variable sea la masa de información. Por eso se establecen medidas que tratan de explicar la dispersión de los datos y son: la desviación estándar y el coeficiente de variación, entre otras. Una medida de dispersión conveniente deberá tomar en consideración todos los datos de la serie considerando cada dato por su distancia al centro de la distribución.

Es posible tener dos conjuntos de datos que tengan el mismo promedio, pero que sean muy diferentes. Por ejemplo, es posible que dos trabajadores puedan obtener el mismo promedio en su desempeño laboral y sus actuaciones hayan sido totalmente diferentes. Uno de ellos pudo haber mantenido un desempeño constante durante el periodo observado; el otro por su parte pudo haber tenido desempeño muy variado.

OBJETIVO.

El alumno conocerá y aplicará las medidas de dispersión aplicables a la investigación social.

TEMARIO

6. MEDIDAS DE DISPERSIÓN.
   - 6.1 Rango.
   - 6.2 Desviación Estandar.
   - 6.3 Coeficiente de Variación.

6. MEDIDAS DE DISPERSIÓN.

6.1 RANGO.

El rango es la diferencia entre el dato mayor y el dato menor dentro de un grupo de datos, puede representarse así:

$$\text{Rango} = \text{Dato mayor} - \text{Dato menor}$$

Debido a que solo toma en cuenta estos dos datos, el rango es afectado directamente por el tamaño de la muestra, entre más grande sea la muestra más grande es el rango y pierde representatividad.

6.2 DESVIACIÓN ESTANDAR.

Esta medida de dispersión es aplicable a variables numéricas y tiene por finalidad observar el grado de alejamiento de los datos respecto a la media del conjunto. Tiene dos propiedades que la robustecen y son: se encuentra en los puntos de inflexión de la distribución y en el intervalo que forma con la media se concentran alrededor del 68.26% de los elementos observados.

Las medidas de dispersión ayudan a conocer la variabilidad que contienen los datos y con eso se interpreta mejor la tendencia que tienen los resultados. Estas medidas son favorables ya que, entre mas se trabajen los resultados, conoceremos más a la población estudiada. Las medidas también ayuda a conocer las características de la población y, aunque conlleven a una serie de desventajas, el aplicar varias de ellas permite tener mas ventaja en cuanto a la interpretación de los resultados.

La desviación estándar es una medida de la dispersión de un conjunto de puntajes alrededor de la media. Para obtener la desviación estándar se empieza por restar la media de cada uno de los puntajes, con lo cual se llega a una nueva serie de valores denominados puntajes de desviación. Luego se elevan al cuadrado estos puntajes de desviación, se suman los cuadrados y se divide la suma por el número de valores que integran la serie, con el fin de obtener la desviación cuadrática media.

La expresión más común para el cálculo de la desviación estándar para un conjunto de datos agrupados es la siguiente:

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

6.3 COEFICIENTE DE VARIACIÓN.

El coeficiente de variación es una medida relativa de dispersión que nos permite hacer comparaciones de diferentes grupos con diferentes unidades de medida o diferentes magnitudes y obtener mejores conclusiones.

Permite asimismo determinar la homogeneidad o consistencia entre los grupos observados, contestando la siguiente pregunta: ¿cuál de los grupos es más homogéneo o presenta menos dispersión o variación relativa respecto a los demás?

Para su cálculo se utiliza la siguiente expresión:

$$CV = s / \overline{X} * 100$$

Ejemplo: en un programa de capacitación, un grupo de ejecutivos obtuvo como evaluación promedio 50.62 puntos con una desviación estándar de 10 puntos, ¿cuál es la variación relativa de los resultados obtenidos en ese programa de capacitación?

Media = x = 50.62

S = 10

CV = 10 /50.62 * 100 = 19.75%

En este ejemplo se puede decir que, la desviación estándar es un 19.75% del promedio, o que las estimaciones pueden variar un 19.75% con respecto a la media.

UNIDAD 7

MEDIDAS DE DISTRIBUCIÓN.

INTRODUCCIÓN.

Indican el grado de asimetría y forma de un conjunto de datos o distribución de frecuencias. En ciencias sociales estas medidas revisten particular importancia debido a que desde el punto de vista gráfico puede observarse la tendencia y forma de la distribución, pudiendo con ello realizar con mayor oportunidad el análisis de los datos en estudio.

Estas medidas están representadas por el sesgo (asimetría) y la curtosis (forma).

OBJETIVO.

Al concluir la unidad el alumno diferenciará los tipos de distribución, así como la tendencia positiva o negativa que presentan las variables en una investigación social.

TEMARIO

7. MEDIDAS DE DISTRIBUCIÓN.
   - 7.1 Sesgo.
     - 7.1.1 Sesgo Negativo.
     - 7.1.2 Sesgo positivo.
     - 7.1.3 Distribución Sesgada.
   - 7.2 Curtosis.

7. MEDIDAS DE DISTRIBUCIÓN.

7.1 SESGO

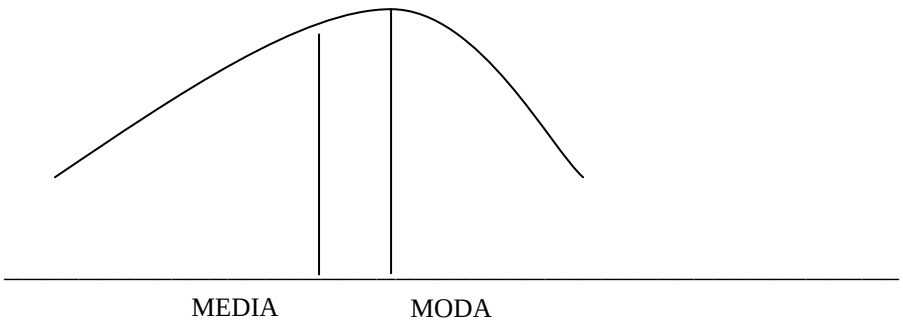
Mide el grado de asimetría de un conjunto de datos. Indica hacia dónde están tendiendo las unidades de observación y cuál es el carácter de esa tendencia, positiva o negativa, cuyo significado refiere las áreas de oportunidad que deben ser atendidas respecto al problema o fenómeno estudiado.

Compara la relación entre la media y la moda, lo que determina su carácter positivo o negativo.

Sesgo = media - moda

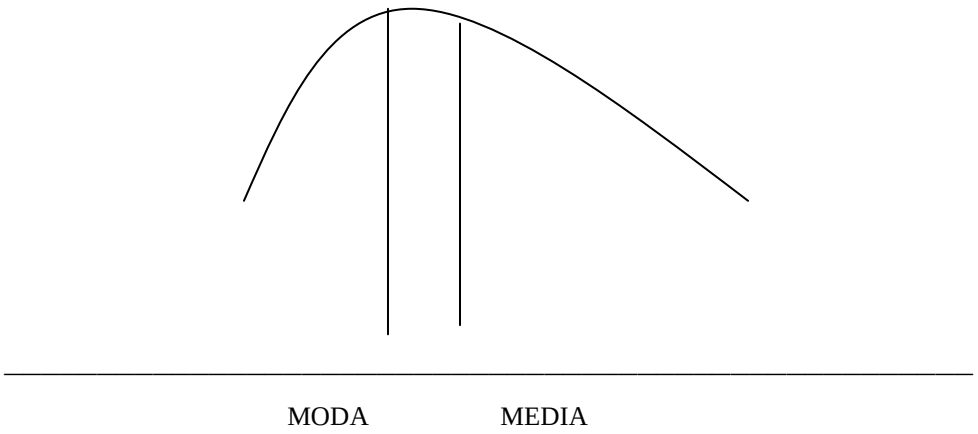
7.1.1 Sesgo Negativo.

El sesgo negativo es cuando la mayoría de los datos se agrupan o concentran a la derecha del conjunto, aquí aparece el valor de la moda a la derecha de la media. Esto indica que el conjunto de datos se van alargando hacia el lado derecho.



7.1.2 Sesgo Positivo.

En una distribución con sesgo positivo la moda se encuentra a la izquierda de la media, aquí la mayoría de los datos se agrupan hacia la izquierda. En esta distribución los datos se concentran hacia la derecha y se alarga hacia la derecha.



7.1.3 Distribución Insesgada.

Las distribuciones insesgadas, llamadas también de sesgo nulo, son aquellas que son simétricas, es decir, que el valor de la media y la moda son iguales. La curva en estas distribuciones no aparece alargada hacia ningún lado; en ciencias sociales es un estado ideal para sus fenómenos de estudio.

Generalmente en ciencias sociales, únicamente se requiere conocer la tendencia, sesgo, de una variable o fenómeno en estudio, soslayando su magnitud; sin embargo, en investigaciones trascendentes es menester conocer la dimensión de la tendencia, que para un conjunto de datos agrupados se calcula con la siguiente expresión aritmética:

$$SK = \frac{\Sigma F (X - MA)^3}{(S^3) (N - 1)}$$

Donde:

Σ = sumatoria
F = frecuencia absoluta o de clase
S = desviación estándar
X = marca de clase
MA = media para datos agrupados
N = número total de datos

7.2 CURTOSIS.

Curtosis es una palabra griega que indica pico y se refiere a la pendiente de una curva o en otras palabras, qué tan puntiaguda es una distribución, tomando como referencia su forma gráfica. Existen tres tipos diferentes de formas de curva en el contexto que se refiere: una es muy alargada hacia arriba o puntiaguda y se llama Leptocúrtica, significa pico alto. La segunda es relativamente plana y se denomina Platicúrtica. La tercera forma de curva es el patrón con el cual se compara la curtosis de otras curvas y poblaciones. Es una curva llamada normal la cual se le denomina mesocúrtica , la cual significa el estado ideal en los fenómenos sociales.

Como se refiere, la **curtosis** mide qué tan puntiaguda es una distribución respecto de una normal y se calcula, para un conjunto de datos agrupados, con la siguiente expresión:

$$k = \frac{\Sigma F (X - MA)^4}{(S^4) (N - 1)}$$

Σ = sumatoria
F = frecuencia absoluta o de clase
S = desviación estándar
X = marca de clase
MA = media para datos agrupados
N = número total de datos

El valor de comparación para determinar el tipo de distribución por medio de la curtosis es 3 (tres). Cuando el coeficiente toma este valor es igual a tres, se dice que la curva es normal o mesocúrtica. Si el valor del coeficiente es mayor que 3 refiere una curva puntiaguda, leptocúrtica; y si es menor que 3, entonces se concluye que los datos presentan una distribución aplanada, platicúrtica.

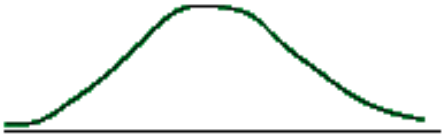
$K = 3$ (distribución mesocúrtica).

$K > 3$ (distribución leptocúrtica).

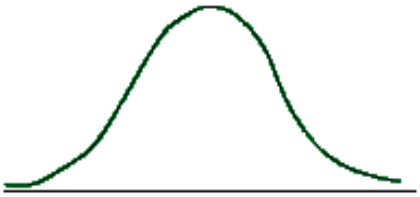
$K < 3$ (distribución platicúrtica).

Las siguientes figuras muestran gráficamente los tres tipos de curvas de acuerdo a las referencias anteriores:

Leptocúrtica

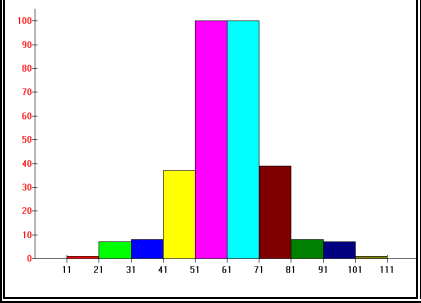
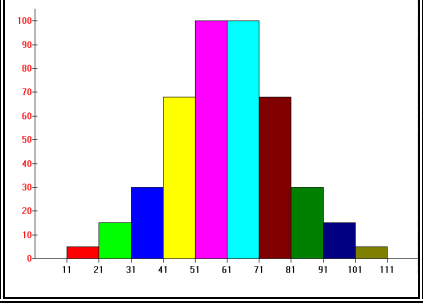
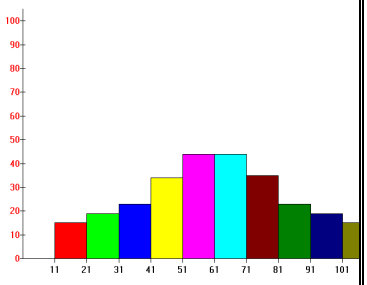


Platicúrtica



Mesocúrtica

Obsérvense las siguientes gráficas para confirmar la identificación de la curtosis sin pasar por el terreno del cálculo.

Curtosis positiva	**Curtosis nula**	**Curtosis negativa**
Leptocúrtica 	Mesocúrtica 	Platicúrtica 

UNIDAD 8

ANÁLISIS DE REGRESIÓN Y CORRELACIÓN.

INTRODUCCIÓN.

La planeación en ciencias sociales tiene un sustento confiable y firme cuando los pronósticos del comportamiento de un fenómeno social u organizacional se fundamentan en el análisis retrospectivo de las variables en estudio. Sucede lo mismo cuando el problema en estudio es explicado al analizar el grado en el que las variables se encuentran relacionadas, que dicho de otra manera, cuando se explica qué variables están influyendo en el fenómeno y sobre todo, la fuerza de esa influencia.

Con el análisis de regresión y correlación, el profesional, investigador o interesado en las ciencias sociales encontrará una herramienta sustancial y confiable para pronosticar al plazo que sea requerido, cuál será el comportamiento y valores esperados de un fenómeno social.

OBJETIVO.

Al finalizar la unidad el alumno establecerá la relación entre variables, así como el grado de asociación de las mismas y su explicación racional en los fenómenos estudiados.

TEMARIO.

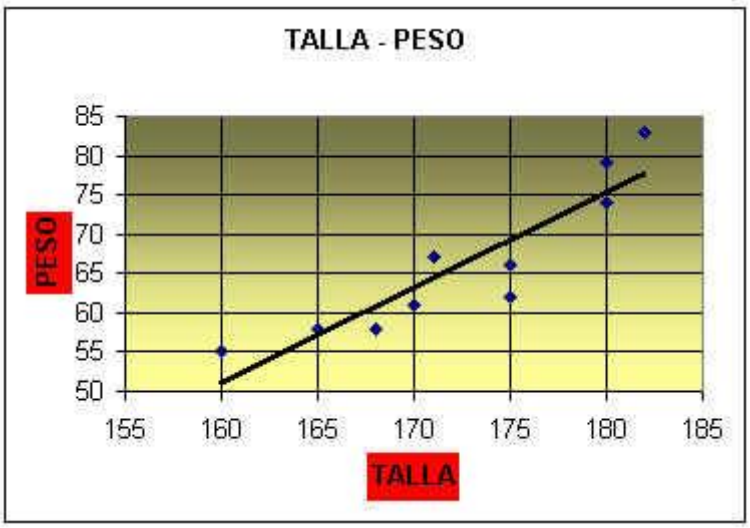
8. ANÁLISIS DE REGRESIÓN Y CORRELACIÓN.
   - 8.1 Análisis de regresión.
     - 8.1.1 Gráfico.
     - 8.1.2 Semipromedios.
     - 8.1.3 Mínimos Cuadrados.

8. ANÁLISIS DE REGRESIÓN Y CORRELACIÓN.

8.1 ANÁLISIS DE REGRESIÓN.

El término regresión refiere la relación entre variables. Toda variable en el contexto universal mantiene relación con otras variables, algunas de ellas presentan mayor influencia o dependencia con la estudiada. Por ejemplo, el ingreso familiar mantiene relación con el número de miembros de la familia que trabajan, suponiendo que entre más sean las personas que aporten al ingreso familiar, mayor será éste. Existen familias en las que sólo uno de sus miembros trabaja y el ingreso familiar es mucho mayor que aquellas en que muchos de sus integrantes hacen aportaciones a la economía familiar.

Hacer un análisis de regresión es encontrar la mejor relación funcional entre las variables en estudio, obteniendo para ello la ecuación algebraica que las relaciona. Esta función puede ser lineal o no lineal, dependiendo de la forma de relación entre los datos, obsérvese la siguiente gráfica, llamada **Diagrama de Dispersión:**



En ella se observa que los datos siguen un comportamiento más o menos lineal, razón por la cual el tipo de regresión a realizar será lineal, representado por la recta cuya ecuación en forma general es la siguiente:

$$y = a + bx$$

Donde a representa **a** la ordenada al origen y **b** a la pendiente.

Por el número de variables que participan en la relación estudiada, la regresión puede ser simple, cuando son dos variables, y múltiple, cuando son más de dos las relacionadas. En la gráfica se aprecia un esquema de regresión lineal simple.

Para encontrar la ecuación de regresión lineal, a continuación se enuncian algunos de los métodos mayormente utilizados en las ciencias sociales:

8.1.1 Gráfico.

Depende de la composición de los datos, así como de la apreciación, sensibilidad y diseño del investigador. Es un método sencillo y rápido siempre y cuando el diagrama de dispersión esté construido con escalas iguales. Si no es así, este método no es recomendable.

El procedimiento de este método es el siguiente:

1. Elaborar el diagrama de dispersión, recordando que las escalas vertical y horizontal deben ser iguales.
2. Sobre los puntos graficados, trazar la línea recta que se considere tiene la mejor posición respecto a los datos.
3. Prolongar la recta hasta que cruce el eje vertical. Este punto por donde la recta corta a ese eje es el valor de la ordenada al origen **a**.
4. Medir el ángulo de inclinación de la recta graficada y obtener el valor de su pendiente, con esto se encuentra la pendiente **b**.
5. Los valores de **a** y **b** encontrados se sustituyen en la ecuación general

$$y = a + bx$$

Encontrando con ello la ecuación de la recta de regresión buscada.

8.1.2 Semipromedios.

Método analítico que consiste en encontrar dos puntos de la recta de regresión. El procedimiento de este método es el siguiente:

1. Ordenar los datos (X,Y) respecto a la variable independiente.
2. Dividir al total de datos en dos partes iguales. Ejemplo: si se tienen 8 datos se formarán dos grupos ordenados ascendentemente de 4 datos cada uno. Si la cantidad de datos es impar, sumar uno al total y dividir en dos partes iguales con los primeros n y los últimos n. Esto es, si se cuenta con 9 datos, al sumar uno se tienen 10, entonces un grupo se integrará con los 5 primeros y el segundo grupo, con los 5 últimos. El dato del centro formará parte de los dos bloques de datos.
3. Calcular la media aritmética de las abcisas (X) y ordenadas (Y) del primer grupo, obteniendo con ello un punto de la recta de regresión buscada.
4. Calcular la media aritmética de las abcisas (X) y ordenadas (Y) del segundo grupo, con lo que se obtiene un segundo punto de la recta buscada.
5. Estos dos puntos, de coordenadas A (X_1 , Y_1) y B (X_2 , Y_2), se sustituyen en la siguiente ecuación:

$$Y = ((Y_2 - Y_1) / (X_2 - X_1)) (X - X_1) + Y_1$$

6. Reducir esta ecuación a su forma más simple, obteniendo como resultado la ecuación de la recta de regresión buscada y que tendrá la siguiente forma:

$$\mathbf{y = a + bx}$$

8.1.3 Mínimos cuadrados.

Método analítico que parte de observar la desviación de los datos respecto a la recta de ajuste. Este es el método que ofrece mayor confiabilidad en un análisis de regresión, ya que se encarga de buscar la ecuación de la recta que tiene la mejor posición respecto a los datos observados.

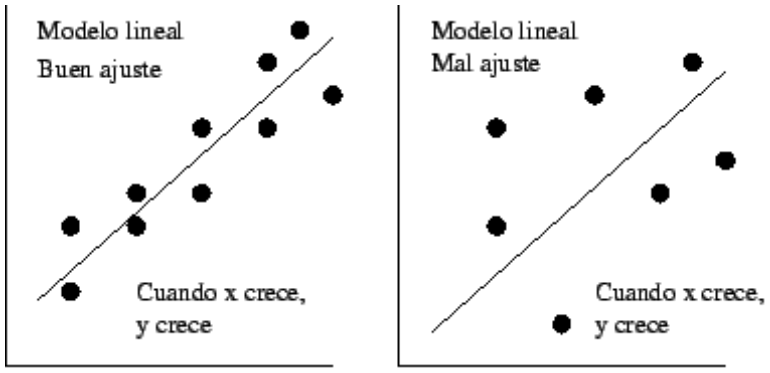
Para encontrar la ordenada al origen y la pendiente de la recta de regresión se utilizan las llamadas ecuaciones normales para el ajuste por el método de los mínimos cuadrados.

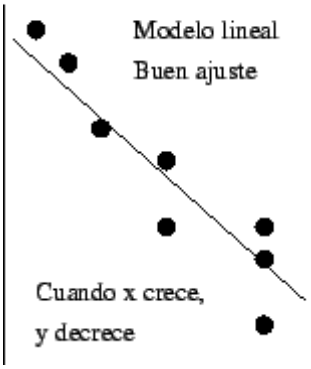
8.1.4 Ecuaciones normales para el ajuste por el método de mínimos cuadrados.

$$a = \frac{(\Sigma X^2)(\Sigma Y) - (\Sigma X)(\Sigma XY)}{N(\Sigma X^2) - (\Sigma X)^2}$$

$$b = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2}$$

Finalmente, un modelo de regresión indica el carácter de la relación entre las variables estudiadas, positivo y negativo, cuya fortaleza se obtiene realizando el análisis de correlación respectivo.





8.2 ANÁLISIS DE CORRELACIÓN.

Correlación refiere el grado de asociación entre dos o más variables dependiendo del modelo de referencia o interés requerido por la investigación. Indica también la fuerza de relación entre las variables en estudio, tratando de explicar qué tanta dependencia tienen entre sí.

En las ciencias sociales es frecuente analizar la asociación entre variables, en virtud de que los fenómenos que estudia tienen una explicación racional y ésta se observa a través de su grado de relación. Respecto al análisis de regresión, la correlación confiere un nivel inicial de confianza en las predicciones o pronósticos.

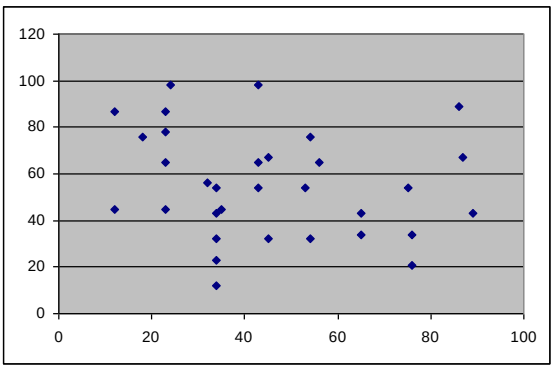
La correlación es medida mediante diversos factores, dependiendo del tipo de variables, numéricas o no numéricas. En este caso se abordará la correlación entre dos variables numéricas y para datos no agrupados, siendo el coeficiente de correlación de Pearson, la herramienta más adecuada al respecto.

El coeficiente de correlación de Pearson tiene un rango de variación de 0 (cero) a 1 (uno). Si el valor que toma es cero se dice que la correlación es nula. Si el coeficiente obtiene el valor de uno, refiere la llamada correlación perfecta. Esto indica que cuando la correlación es nula, las variables son independientes, por lo menos a partir de los datos analizados, razón que puede ser circunstancial.

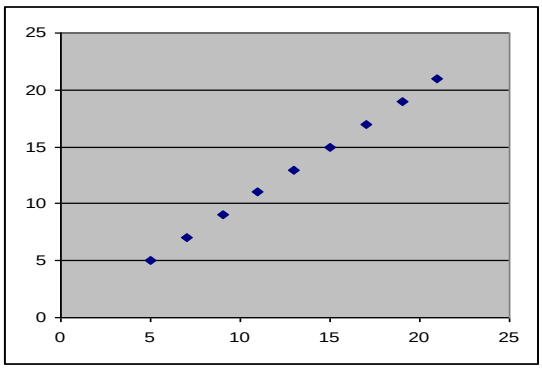
En el caso de la correlación perfecta se concluye que los datos están asociados o relacionados al 100%, situaciones extremas y prácticamente imposibles de presentarse en las ciencias sociales. Con esta premisa, el investigador social deberá estar conciente de que un fenómeno será confiablemente predecible, en la medida en que su nivel de correlación sea alto.

Un nivel de correlación es alto o aceptable, aunque esto no es una regla general, es aquel cuyo valor es de al menos 0.8 equivaliendo al 80% de relación. Existen investigaciones o estudios que consideran que una correlación del 80% es baja, concluyendo que el nivel de confianza al respecto, es casuístico y se fundamenta en el criterio del investigador.

El coeficiente de correlación será igual a cero si la correlación es nula. Las variables son independientes. →



El coeficiente de correlación será igual a uno si la correlación es perfecta. Las variables son dependientes. →



El coeficiente de correlación de Pearson se obtiene mediante la siguiente expresión:

$$= \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

Concluyendo, un coeficiente de correlación igual a cero refiere independencia entre las dos variables, y un coeficiente de correlación igual a 1 indica una dependencia total entre las dos variables, de tal manera que cuando una de ellas aumenta la otra también aumenta en la misma proporción, recordando que en ciencias sociales estos extremos son absolutamente inesperados.

PREGUNTAS FRECUENTES

¿En qué se distingue la estadística social de las demás aplicaciones de la asignatura?

En las ciencias sociales las variables son principalmente no numéricas y en las demás disciplinas son numéricas, lo que obliga a un tratamiento distinto.

¿Por qué sólo en la estadística social se utilizan las escalas de medición?

Por el tipo de variables, que pueden ser clasificatorias, jerarquizables y cuantificables; ya que en otras disciplinas las variables únicamente son medibles por su carácter numérico.

¿Cuál es la diferencia entre la estadística como asignatura normal y como taller?

En que en el taller se espera un producto al final del curso, en este caso un proyecto de investigación y en la asignatura normal no es un requisito indispensable, aunque también puede ser requerido por el profesor.

¿Qué importancia tiene la estadística en la formación de los trabajadores sociales?

Muy sencillo de apreciar: los trabajadores sociales de vanguardia, innovadores, visionarios o emprendedores tienen la característica de participar activamente en la toma de decisiones de las organizaciones, razón por la cual deben tener un contacto directo con la generación de información, y ésta principalmente surge del proceso de investigación, sustentado sustancialmente por el proceso estadístico.

¿Cómo se sustenta la toma de decisiones en la estadística?

La estadística tiene como misión auxiliar a la toma de decisiones, mediante la efectiva utilización de sus recursos. Toda información es generada a partir de la

observación, proceso y análisis de resultados, siendo este proceso la naturaleza del tratamiento de datos mediante el uso de herramientas estadísticas.

Si la toma de decisiones puede partir de la intuición, ¿por qué utilizar la estadística para ello?

Ciertamente, con habilidades no muy comunes, pueden tomarse decisiones por intuición, lo mismo que hacerlo mediante la información que produce el tratamiento estadístico; sin embargo, el sustento es sólido en este último caso, sin garantizar el éxito en las decisiones. Esto ocurre en ambas situaciones. Se recomienda por lo tanto que las decisiones sean tomadas con información, sin pasar por alto que con una dosis de inspiración puede traer mejores resultados.

¿Qué distingue a las medidas de tendencia central de las medidas de posición?

Las medidas de tendencia central, promedios, indican la concentración de los datos sobre un punto determinado, que regularmente es hacia el centro del conjunto, no siendo siempre los valores más representativos del mismo. Las medidas de posición se localizan en cualquier punto del conjunto o distribución, en función del interés del investigador.

¿Qué distingue a las medidas de dispersión de las medidas de distribución?

Las medidas de dispersión indican el alejamiento de los datos respecto al promedio del conjunto y las de distribución refieren la forma y tendencia de los datos.

¿Qué utilidad tienen los modelos de regresión en ciencias sociales?

Los modelos de regresión permiten realizar pronósticos del comportamiento de los fenómenos sociales y de esta manera poder realizar actividades de planeación y toma de decisiones.

BIBLIOGRAFÍA

Hubert M. Blalock , JR. *Estadística Social;* Fondo de cultura económica 1996

Gene V Glass , Julian C. Stanley *Métodos Estadísticos Aplicados a las Ciencias Sociales ;*Pretice Hall 1980

Frederick E. Croxton. Estadística General aplicada. México: Fondo de cultura económica, 1948.

Gabriel- R. Cubrí. Práctica de las encuestas estadísticas. España: Ariel, 1967.

Manuel López Cachero. Fundamentos y métodos de estadística. España: Pirámide, 1978.

Fernando Holguín Quiñones. Estadística descriptiva aplicada a las ciencias sociales. México: UNAM, 1979

Gómez Espadas, José L. *Teorías y problemas de estadística.* México: Mac Grawhill, 1987.

López Cachero, Manuel. *Fundamentos y métodos de estadística.* Madrid: Ediciones pirámide, 1978.

García, Alfonso [et al.]. *Estatística I: informática de sistemas.* Madrid: Universidad Nacional de Educación a distancia, 1994.

Christen, Howard B, Estadística paso a paso. –3ª ed.— Trillas, 1990 (reimp.2001).México.

Fundamentos de Estadística/ Justo Arnal-Antonio Olmedes,
Maqueta e il: Ediciones Daimon, Daniel Tamayo, Imp. GRAFOS.1981.