MÓDULO

ESTADÍSTICA DESCRIPTIVA

Mónica A. Santa Escobar

Versión Preliminar
Diciembre 30 de 2005

UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA – UNAD –
FACULTAD DE CIENCIAS BÁSICAS E INGENIERÍA
UNIDAD DE CIENCIAS BÁSICAS
Bogotá D.C., 2005

COMITÉ DIRECTIVO

Jaime Alberto Leal Afanador
Rector

Roberto Salazar Ramos
Vicerrector Académico

Sehifar Ballesteros Moreno
Vicerrector Administrativo y Financiero

Maribel Córdoba Guerrero
Secretaria General

Edgar Guillermo Rodríguez
Director de Planeación

MÓDULO
CURSO ESTADÍSTICA DESCRIPTIVA
PRIMERA EDICIÓN

ISBN
2005
Centro Nacional de Medios para el Aprendizaje

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Pág.

UNIDAD DIDÁCTICA 1
Conceptos preliminares

UNIDAD DIDÁCTICA 2
Medidas estadísticas

INTRODUCCIÓN

La Estadística es una disciplina que se aplica en muchos campos de la actividad del ser humano. Es muy frecuente encontrarse en las diferentes disciplinas del saber con incertidumbres como el pronosticar el crecimiento poblacional de un país, el crecimiento económico de una empresa o el crecimiento de producción y venta de un producto específico, el conocer la efectividad de diferentes abonos en el campo agrario, el determinar la tendencia de contaminación del agua o el aire, la clasificación de personal en una empresa para efectos de una buena y sana política laboral, etc.

Habitualmente, el propósito de la Estadística Aplicada es el de sacar conclusiones de una población en estudio, examinando solamente una parte de ella denominada muestra. Este proceso, llamado *Inferencia Estadística,* suele venir precedido de otro: la *Estadística Descriptiva*, en el que los datos son ordenados, resumidos y clasificados con objeto de tener una visión más precisa y conjunta de las observaciones, intentando descubrir de esta manera posibles relaciones entre los datos, viendo cuáles toman valores parecidos, cuáles difieren grandemente del resto, destacando hechos de posible interés, entre otros.

En todos los campos de la investigación se requiere a menudo el uso racional de los Métodos Estadísticos. Los procesos de planeación, control y toma de decisiones económicas, administrativas y financieras se basan en resultados obtenidos mediante el análisis estadístico de los fenómenos en ellos involucrados. El acelerado desarrollo de métodos, técnicas y tecnologías para el óptimo análisis de datos justifica que un profesional disponga de una sólida fundamentación conceptual para que realice apropiadamente su evaluación y aporte sustentaciones a su decisión. Las interpretaciones que generan los datos pudieran ser erróneas para aquellas personas que no cuentan con criterios válidos para captar la información. Es por ello que con este módulo se pretende que el estudiante se adentre a los conocimientos básicos de la *Estadística Descriptiva*.

Enfrentarse con datos de muy diversa índole es cosa de todos los días en cualquier práctica del ser humano. Sin embargo, dado la cantidad innumerable de estos, no siempre se comprende el real alcance de lo que dicen. Como parte de una base cultural necesaria para desempeñarse en el mundo de hoy, es requisito desarrollar una capacidad personal para extraer y describir información presente en un conjunto de datos. Y es precisamente allí donde resalta la importancia de la *Estadística Descriptiva* como primer paso en la determinación de decisiones e inferencias que pueden concluirse de la variada información que nos llega en forma de datos numéricos.

Con el presente módulo, se busca que el estudiante se encuentre en capacidad de interpretar, discriminar y relacionar los fundamentos básicos de la *Estadística Descriptiva*, a través del análisis de datos tomados de un fenómeno propio de su disciplina y que describa, examine y sintetice adecuadamente la información mediante métodos estadísticos sencillos.

El curso académico de *Estadística Descriptiva* hace parte de la formación básica disciplinar de los programas que oferta la Universidad Nacional Abierta y a Distancia —UNAD—. Consta de dos (2) créditos académicos, el sistema adoptado por la UNAD como estándar curricular en la formación universitaria, y es de tipo teórico, en tanto que busca la identificación y el reconocimiento de las problemáticas, perspectivas teóricas, conceptos, categorías, métodos y técnicas indispensables para la formación profesional.

Este texto contiene dos unidades didácticas[1], correlacionadas directamente con el número de créditos académicos asignados al curso académico. La primera de ellas, considera los Conceptos Básicos necesarios para el cumplimiento de los propósitos y objetivos del curso. En esta unidad se identifican algunos conceptos estadísticos como población, muestra, variable, dato, etc.; y se reconocen cada uno de los pasos a seguir para una correcta y acertada investigación estadística como son la planeación, la recolección de la información, su organización y su presentación gráfica. En la segunda unidad didáctica se reconocen algunas de las medidas estadísticas más comunes, tanto univariantes como bivariantes. Entre las primeras se contemplan las medidas de tendencia central, las medidas de dispersión y las de asimetría y apuntamiento y, como medidas estadísticas bivariantes, se trabaja la regresión lineal (simple, ponderada y múltiple), la correlación y los números índice. Como Anexo y complemento a esta segunda unidad, se incluyen algunos elementos básicos de la matemática: la sumatoria y productoria. Al final de cada tema, encontrará ejercicios de aplicación que buscan evaluar el grado de conocimiento adquirido, esta evaluación será retroalimentada en la información de retorno que encontrará al final de cada unidad didáctica.

Este texto busca aportar las herramientas teóricas y prácticas a los estudiantes para que logren, mediante análisis cuantitativos, la interpretación de diferentes fenómenos propios de su disciplina de formación y del entorno social, económico y político. Apunta al manejo estadístico de datos, dar las pautas en la recolección planeada de los mismos y proporcionar un conjunto de técnicas a

[1] Conjunto de conocimientos seleccionados, organizados y desarrollados a partir de palabras clave tomados como conceptos que los tipifican, en articulación con las intencionalidades formativas, destinadas a potenciar y hacer efectivo el aprendizaje mediante el desarrollo de operaciones, modificaciones y actualizaciones cognitivas y nuevas actuaciones o competencias por parte del estudiante. EL MATERIAL DIDÁCTICO. *Roberto J. Salazar Ramos.* UNAD, Bogotá D.C. 2004.

partir de las cuales se logra presentar, resumir e interpretar datos que pueden corresponder a una muestra o a un grupo total.

El módulo no pretende reemplazar las diferentes referencias bibliográficas clásicas de la Estadística, busca entregar los conceptos de un modo más didáctico, enfocado en el autoaprendizaje y en relación directa con la Guía de Actividades referenciada en el protocolo del presente curso. Al final de cada unidad, el estudiante encontrará las referencias bibliográficas básicas, pero no únicas, para que con ellas refuerce en conceptos y definiciones. Además, encontrará una serie de páginas web recomendadas que amplían los temas tratados. Se trata pues de un material didáctico de apoyo para el curso de *Estadística Descriptiva* de la UNAD, como parte de las diferentes y diversas herramientas didácticas en las que se apoya el aprendizaje autónomo.

Unidad Didáctica Uno

CONCEPTOS PRELIMINARES

Unidad Didáctica Uno
CONCEPTOS PRELIMINARES

1. Generalidades y Conceptos Básicos

1.1. Conceptualización de términos estadísticos

1.1.1. ¿Qué es la Estadística?
1.1.2. Conceptos básicos

2. Investigación Estadística

2.1. Planeación

2.1.1. Definición del objeto de investigación
2.1.2. Unidad de investigación
2.1.3. Clase de investigación
2.1.4. Las fuentes de información

2.2. Recolección

2.2.1. Según la cobertura
2.2.2. Según la forma de observación

2.3. Organización de la información

2.3.1. Combinación o arreglo ordenado
2.3.2. Arreglo tallo y hojas
2.3.3. Tabulación de la información
2.3.4 Distribuciones de frecuencia

2.4. Presentación de la información

2.4.1. Componentes de una gráfica
2.4.2. Diagrama de frecuencias
2.4.3. Histograma de frecuencias
2.4.4. Polígono de frecuencias
2.4.5. Ojiva
2.4.6. Gráficos de línea
2.4.7. Diagramas de barra
2.4.8. Diagrama circular
2.4.9. Pictogramas
2.4.10. Mapas estadísticos o cartogramas

INTRODUCCIÓN A LA UNIDAD

La investigación estadística es necesaria para cualquier individuo en el mundo de hoy, cualquiera que sean sus actividades siempre hay aplicaciones estadísticas en ellas. Pero cualquier investigación estadística requiere seguir unos pasos y procedimientos establecidos para que esta tenga validez. En esta unidad se desarrollarán en forma introductoria y general algunos conceptos preliminares con el fin de utilizar un mismo lenguaje en cuanto se refiere a esta disciplina. De igual manera, se presentan los elementos iniciales básicos y necesarios para la comprensión y aplicación de la estadística en cualquier campo.

En el capítulo uno se ampliarán algunas definiciones de términos básicos de la estadística como población, muestra, variable, dato, etc., buscando que el estudiante los identifique en ejemplos sencillos de la vida diaria. En el capítulo dos se reconocerán cada uno de los pasos a seguir para una correcta y acertada investigación estadística como son la planeación, la recolección de la información, su organización y su presentación gráfica.

OBJETIVOS ESPECÍFICOS

- Conocer el significado de la palabra estadística.
- Diferenciar entre los conceptos de Estadística Descriptiva y Estadística Inferencial.
- Establecer los conceptos de población, muestra, variable, dato y parámetro.
- Identificar las etapas que sugiere una investigación estadística.
- Manejar los diferentes métodos de recolección de información para la investigación estadística.
- Advertir la importancia de las distribuciones de frecuencias para la descripción de datos.
- Aplicar los conceptos de frecuencia, marca de clase y distribución de frecuencias a un conjunto de datos estadísticos.
- Construir diferentes tipos de distribuciones de frecuencias para conjuntos de datos.
- Reconocer algunas características que debe tener una gráfica para que represente mejor una situación.
- Representar gráficamente distribuciones de frecuencias dadas o calculadas.

1. GENERALIDADES Y CONCEPTOS BÁSICOS

1.1. CONCEPTUALIZACIÓN DE TÉRMINOS ESTADÍSTICOS

1.1.1. ¿Qué es la Estadística?

Antes de dar a conocer los conceptos de los términos estadísticos que lleven a entablar el lenguaje común que se utilizará en adelante, es necesario saber qué es la Estadística y en qué consiste la Estadística Descriptiva.

Empíricamente se sabe que la Estadística tiene que ver con datos y la manera en que estos son agrupados. Esto se reconoce en muchos casos de la vida cotidiana que involucran información numérica y el contexto en que esta información es dada a conocer. Aunque también puede darse en muchos casos que, si bien están relacionados con la estadística, obedecen a otros fenómenos de disciplinas relacionadas con —pero que no conforman— la Estadística propiamente dicha.

La ***Estadística*** es un método científico de operar con un grupo de datos y de interpretarlos.

Si bien esta definición parece un poco ambigua, se verá más adelante el marco en que éste método se desarrolla y las “leyes” que lo rigen. Pero, por ahora, se deja abierta al cuestionamiento del estudiante la gama de posibilidades que abarca esta definición.

La Estadística, o el método de la estadística, se divide en dos ramas: la Estadística Descriptiva o deductiva y la Inferencia Estadística o estadística inductiva. Este curso se dedica a la Estadística Descriptiva, por lo que se hace necesario dar a conocer, en términos generales, en qué consiste la Inferencia Estadística.

La ***Inferencia Estadística*** comprende en un todo articulado el método y las técnicas necesarias para explicar el comportamiento de un grupo de datos en un nivel superior de lo que estos datos pueden dar a conocer por sí mismos. Es decir, se puede concluir sobre el grupo de datos sobrepasando los límites del conocimiento inicial que estos suministran, examinando solamente una parte de la población denominada muestra. Es por ello que a la Inferencia Estadística también se le conoce como Estadística Analítica.

Si esto es así, ¿qué le corresponde entonces a la ***Estadística Descriptiva***?

Esta tiene por fin elevar los aspectos característicos del grupo de datos pero sin intentar obtener más conocimiento del que pueda adquirirse por sí mismos. Es por ello que la Estadística Descriptiva es el punto de partida del análisis de un grupo de datos que involucran una cierta complejidad, o bien puede ser el todo de un análisis básico y limitado del grupo de datos.

1.1.2. Conceptos básicos

Población es el conjunto de medidas, individuos u objetos que comparten una característica en común. La población se basa en cuatro características: contenido, tipo de unidades y elementos, ubicación espacial y ubicación temporal. De la población es extraída la muestra.

Muestra es un conjunto de elementos extraídos de la población. Los resultados obtenidos en la muestra sirven para estimar los resultados que se obtendrían con el estudio completo de la población. Para que los resultados de la muestra puedan generalizarse a la población, es necesario que la muestra sea seleccionada adecuadamente, es decir, de modo que cualquiera de los elementos de la población tengan la misma posibilidad de ser seleccionados. A este tipo de muestra se le denomina ***muestra aleatoria***.

La ***unidad estadística*** es el elemento de la población que reporta la información y sobre el cuál se realiza un determinado análisis.

Los ***datos*** son todas aquellas características o valores susceptibles de ser observados, clasificados y contados. Estos pueden ser ***experimentales***, cuando se le aplica un tratamiento especial a las unidades muestreadas; ***de encuesta,*** cuando son tomadas sin ningún tratamiento; ***clasificados***, cuando están agrupados según una característica determinada; ***originales***, información que no ha recibido ningún tratamiento estadístico; ***primarios***, cuando son recogidos, anotados u observados por primera vez; o ***secundarios***, cuando son recopilados por otra persona o entidad diferente al investigador.

Variable es una característica susceptible de tener distintos valores en los elementos de un grupo o conjunto. Si la variable tiene la capacidad de tomar cualquier valor que exista entre dos magnitudes dadas, entonces esta variable será ***continua***. Si por el contrario, sólo puede tener un valor de entre cierta cantidad de valores dados, entonces será ***discreta***.

Parámetro son aquellos valores que caracterizan numéricamente a la población como tal. El parámetro poblacional de interés es único (media, varianza, etc.), pero una población puede tener muchas características —o parámetros— de interés. Por el contrario, un ***estadístico*** es una magnitud correspondiente a una

muestra aleatoria extraída de la población, por lo que cambiando la muestra cambiará entonces el estadístico (media muestral, varianza muestral, etc.). En pocas palabras se puede decir que parámetro es a población como estadístico es a muestra. Es común designar los parámetros con letras minúsculas del alfabeto griego y los estadísticos con letras de nuestro alfabeto. En la Unidad Didáctica Dos, se ampliará más estos dos conceptos.

EJEMPLO 1.1.

La Universidad Nacional Abierta y a Distancia UNAD desea establecer cuántos estudiantes hacen uso de la biblioteca en el CEAD de San Juan de Pasto. El coordinador zonal de biblioteca es designado para este trabajo y decide hacer la investigación el día 14 de mayo de 2005.

- En esta investigación se considera que el total de estudiantes del CEAD que hacen uso de la biblioteca es la ***población*** en estudio.
- Cada uno de los estudiantes seleccionados para la observación representa la ***unidad estadística*** de estudio
- El día 14 de mayo de 2005 indica la ***ubicación temporal***.
- El CEAD de San Juan de Pasto, identifica la ***ubicación espacial.***
- Como el coordinador zonal de biblioteca no puede revisar todo el día quienes acceden a la biblioteca, decide entonces establecer períodos de tiempo para realizar el conteo. En otras palabras, selecciona una ***muestra***.
- Identificada la población y la muestra, se ubica la ***unidad estadística***, en este caso el objeto de medición es cada uno de los estudiantes seleccionados de la muestra.
- Y la ***variable*** será el número de estudiantes seleccionados de la muestra, como se puede ver, una ***variable discreta***.
- Después de esto el coordinador selecciona los ***datos*** necesarios para el estudio, en este caso específico sólo requerirá del número de estudiantes que acceden a la biblioteca. Sin embargo, el coordinador zonal puede además, tomar otro tipo de datos como sexo, edad, razón por la cual visita la biblioteca, libros más consultados, etc.

EJERCICIOS CAPÍTULO 1.

1. Elabore un mapa conceptual en donde diferencie claramente los conceptos de Estadística Descriptiva e Inferencia Estadística.

2. Un equipo de fútbol profesional está compuesto de jugadores y cuerpo técnico.

- **a.** Si se desea conocer el promedio de edad de la selección Colombia para establecer una correlación entre edad y rendimiento físico, ¿tiene sentido registrar la edad del cuerpo técnico?
- **b.** Si sólo se está interesado en el grupo de jugadores, ¿qué datos pueden extraerse de ellos que tengan relevancia en el aspecto deportivo?
- **c.** Si se toma un jugador y se registra la velocidad con que recorre la cancha y la cantidad de goles anotados en un campeonato, ¿cuál de estas variables es continua y cuál es discreta?

3. De las siguientes variables, diga cuáles son continuas y cuáles discretas:

- **a.** Velocidad de un automóvil en kilómetros por hora.
- **b.** Valor total de acciones vendidas cada día en el mercado de valores.
- **c.** El volumen de gasolina que se pierde por evaporación durante el llenado de un tanque de combustible.
- **d.** El número de moléculas en una muestra de gas.
- **e.** La medida de la cantidad de lluvia caída en una localidad en un mes.
- **f.** Candidatos a la presidencia de la República.

4. En las siguientes situaciones, identifique: población, muestra, unidad estadística, dato y variable, y diga si esta última es discreta o continua.

- **a.** En la UNAD la matrícula en un año es de 10.458 estudiantes distribuidos en las cuatro facultades. Se desea conocer el número de estudiantes matriculados en la facultad de Ciencias Agrarias.
- **b.** Las temperaturas registradas en la ciudad de Pereira el 29 de junio de 2005 entre las 6 horas y las 18 horas.
- **c.** Se realiza un estudio a 250 hogares en la ciudad de Medellín para conocer si se hace uso adecuado del Manejo Integrado de Residuos Sólidos (MIRS).
- **d.** Las exportaciones mensuales de café colombiano durante el año 2004, en millones de dólares.

2. INVESTIGACIÓN ESTADÍSTICA

2.1. PLANEACIÓN

La planeación de una investigación estadística debe abarcar el conjunto de lineamientos, procedimientos y acciones que conlleven a la resolución satisfactoria para la cual se estableció la investigación. Es por ello que el plan de investigación debe fijar concretamente su objeto, el fin que persigue, la fuente o fuentes de información, los procedimientos a seguir y resolver los aspectos logísticos, físicos y humanos siguiendo un presupuesto de costos establecido.

La investigación estadística puede ser tan sencilla y poco compleja como la recopilación ordenada y coherente de datos que se encuentren en instituciones estatales o privadas que las suministren, o bien pueden ser tan elaboradas y complejas como lo son los censos poblacionales, los censos agrícolas o industriales que tengan importancia estratégica para una región, o inclusive para un país. Pero, sea como fuere, la investigación debe seguir una orientación en su planteamiento y resolución.

A continuación se presentan algunos aspectos básicos que se deben seguir para desarrollar un trabajo así.

2.1.1. Definición del objeto de investigación

Debe responder el qué, el cómo y establecer el momento correcto para hacerse, debe también restringir el espacio físico o geográfico donde se llevará a cabo. Es este punto el núcleo de la investigación, es por ello que no puede haber ambigüedad en sus planteamientos y alcances.

2.1.2. Unidad de investigación

Se trata del elemento de la población que origina la información. La unidad o elemento de investigación debe ser clara, adecuada, mesurable y comparable.

Debe determinarse la naturaleza cuantitativa o cualitativa de esta unidad, es decir, definir qué aspectos de la unidad de investigación son cuantitativos (registrados por medio de números) o cualitativos (recogidos mediante anotaciones literarias). También ha de considerarse la posibilidad o viabilidad de la investigación y si estos aspectos pueden ser conocidos con precisión. De igual manera, es necesario delimitar esta unidad en el tiempo y en el espacio, y a veces en el número.

2.1.3. Clase de investigación

En la planeación, debe también tenerse en cuenta el tipo de investigación que se va a realizar. Ésta puede ser ***descriptiva***, que consiste en obtener información con respecto a grupos; ***experimental*** o ***controlada***, provocada por el investigador en condiciones controladas, en la que se busca conocer por qué causa se produce un caso particular; o bien, ***explicada*** o ***analítica***, que permite establecer comparaciones y verificar hipótesis.

2.1.4. Las fuentes de información

Después de determinar el qué y el por qué de la investigación estadística, se debe preguntar el dónde conseguir la información requerida. Se trata entonces de definir las fuentes de información. Estas pueden ser directas o indirectas.

Una fuente de información estadística ***directa*** es aquella en donde el hecho se produce. Este tipo de fuentes son las mejores, pero no siempre son posibles. Cuando no sea posible, se emplea una fuente de información estadística ***indirecta***, aquella donde el hecho se refleja. En muchos casos este tipo de fuentes son complementarias de las primeras.

Cuando los datos son primarios, ellos pueden provenir de muchas fuentes como: ***hechos***, información cotidiana y fácil de identificar; ***opiniones***, referidos a lo que la gente piensa respecto a algo; ***motivos***, razones que explican por qué se actúa de una manera u otra. Cuando son secundarios ellos provienen de una ***fuente interna***, cuando los datos son recopilados por la misma entidad en los registros básicos de la misma organización, o bien pueden provenir de una ***fuente externa***, cuando los datos se recopilan por otra entidad diferente a la que hace la investigación.

2.2. RECOLECCIÓN

Después de planeada la investigación, comienza la recolección de los datos. Esta consiste en un conjunto de operaciones de toma de datos que puede ser por observación, por encuesta o tomada de publicaciones y/o fuentes confiables que han efectuado investigaciones estadísticas. Para esto se selecciona el método de recolección de la información acorde a las necesidades de la investigación, que se clasifican según su cobertura y según su forma de observación.

2.2.1. Según la cobertura

Se trata de decidir si se va a estudiar a la población en su totalidad o sólo una parte de ella. Si lo que se desea es atender a una cobertura total, es decir contar con todos los elementos de las fuentes de información, se usa el ***censo***. Si, en cambio, se hace una enumeración parcial de las fuentes de información, se usa el ***muestreo***.

Por su menor costo, mayor rapidez y menor número de personas que intervienen en la investigación, el muestreo es el método más utilizado. El muestreo puede ser de dos tipos: ***muestreo probabilístico*** o ***al azar***, cuando cada uno de los elementos tiene la misma probabilidad de ser escogido obteniendo así una ***muestra aleatoria***; y ***muestreo no probabilístico***, cuando el investigador selecciona los datos a su propio criterio, de manera caprichosa, por conveniencia o por cuotas, de manera que las muestras no son seleccionadas aleatoriamente y los resultados no ofrecen confiabilidad alguna.

2.2.2. Según la forma de observación

En este método se tiene en cuenta la forma de medición del dato. Si se hace de manera que la fuente de información se da cuenta de la medición que efectúa, se dice que se toman los datos por ***encuesta***. Éstas se pueden realizar por correo, entrega personal de cuestionario, entrevista, motivación, teléfono, etc.

El otro método de recolección de información es por ***observación***, en donde la medición se realiza sin que la fuente de información se dé cuenta del hecho. Este método se basa en el registro de los eventos que ocurren, por ejemplo cuando se examina el número de estudiantes que entran a la biblioteca con el fin de hacer una consulta referida a las Ciencias Sociales, simplemente se observa la acción del estudiante al entrar a la biblioteca: si hace o no la consulta que se investiga. Este método puede ser también indirecto cuando la recolección consiste en corroborar los datos que otros han observado.

2.3. ORGANIZACIÓN DE LA INFORMACIÓN

Luego de tomar la información necesaria en la investigación que se sigue, se obtiene una gran cantidad de datos que requieren ser interpretados y sobre los cuales se busca concluir algo específico. Para esto se debe depurar y clasificar la información aplicando técnicas adecuadas.

La organización y el resumen de la información son dos procesos distintos que se desarrollan por separado. La organización hace referencia al arreglo de los datos en un formato lógico para su interpretación. En cambio, el resumen implica

la condensación de varias mediciones en una forma compacta, ya sea gráfica o numéricamente. De ahí que se tome primero la forma de organizar la información tomada en una investigación estadística.

La información estadística puede organizarse de diversas maneras: ordenando el conjunto de datos como una combinación ordenada o en un arreglo denominado tallo y hojas, otro de los métodos usados es el uso de tablas y más específicamente la tabla de frecuencias. A continuación se hace un acercamiento a las diferentes formas de organizar los datos estadísticos tomados en el proceso de recolección de una investigación estadística.

2.3.1. Combinación o arreglo ordenado

El sólo hecho de tener ordenado un conjunto de datos en forma ascendente o descendente, permite un rápido análisis e interpretación de estos.

EJEMPLO 2.1.

Los siguientes datos representan la evaluación de los latidos cardíacos de un grupo de 30 personas después de cierta actividad física.

82	95	92	62	85	92
82	95	70	85	84	95
91	82	94	76	88	91
87	80	68	58	76	85
110	60	75	88	64	74

Es muy poca la información que arroja este conjunto de datos cuando se encuentran sin un tratamiento. A continuación estos datos son presentados como una ***combinación ordenada*** en forma ascendente (de menor a mayor):

58	70	80	85	88	94
60	74	82	85	91	95
62	75	82	85	91	95
64	76	82	87	92	95
68	76	84	88	92	110

A partir de esta lista ordenada se pueden concluir varias cosas:

- La más alta evaluación de latidos es 110
- La más baja evaluación de latidos es 58
- La mitad de la combinación se encuentra entre 82 y 85

- Hay una predominancia en los latidos con una evaluación entre 80 y 95
- Hay un “vacío” entre el valor 95 y el valor 110, es decir hay una cierta continuidad en los valores entre 58 y 95, pero 110 se encuentra más alejado del grupo de datos.
- Hay una evaluación atípica dentro del grupo de 30 personas, el que registra el valor 110. Es posible que esta persona tenga perturbaciones cardíacas. Sin embargo, es necesario ampliar la información antes de lanzar un juicio apresurado.

2.3.2. Arreglo de tallo y hojas

El arreglo de tallo y hojas es una técnica que resume de manera simultánea los datos en forma numérica y presenta una ilustración gráfica de la distribución.

Se trata de organizar los datos numéricos en dos columnas divididas por una línea vertical. La primera, denominada tallo, corresponderá a las decenas, centenas o unidades que representan el grupo de datos y en la segunda, llamada hojas, irán las correspondientes decenas, unidades o décimas. Para una mejor ilustración, en el siguiente ejemplo se continuará con los datos del ejemplo 2.1. para construir el correspondiente arreglo de tallo y hojas.

EJEMPLO 2.2.

Tomando la serie de datos del ejemplo 2.1., se puede observar que éstos tienen un rango desde los cincuentas hasta los ciento diez. Ellos se pueden presentar como un arreglo de tallo y hojas en una columna de números del 5 al 11 y trazando una línea vertical a su derecha. Estos números representarán el tallo. En la columna de las hojas, se enlistan las unidades (de manera ordenada) de cada uno de los datos registrados y correspondientes con su respectiva decena.

Figura 2.1.

Diagrama de tallo y hojas para los datos de pulsaciones del ejemplo 2.1.

Tallo	Hojas										
5	8										
6	0	2	4	8							
7	0	4	5	6	6						
8	0	2	2	2	4	5	5	5	7	8	8
9	1	1	2	2	4	5	5	5			
10											
11	0										

Observe que el diagrama de tallo y hojas al mismo tiempo que ordena los datos de forma ascendente, permite una visualización del comportamiento de estos. Con este se pueden confirmar muchas de las afirmaciones que se hacían en el ejemplo 2.1.

- La mayoría de los registros de latidos cardiacos del grupo de 30 personas se encuentra entre los ochentas.
- La forma general del conjunto de mediciones es asimétrico.
- Se ve más claramente el "vacío" que existe entre los valores 95 y 110, y se resalta cómo el valor de 110 se encuentra aislado del resto de conjunto de datos.

Si se quisiera clasificar más ampliamente los datos, se usa un ***diagrama de doble tallo***. Que consiste en dividir en dos cada posición del tallo, en grupos de cinco hojas. La primera posición del tallo dispone las hojas 0, 1, 2, 3, 4; y la segunda posición dispone las hojas 5, 6, 7, 8, 9.

Figura 2.2.
Diagrama de doble tallo y hojas para los datos de pulsaciones del ejemplo 2.1.

Tallo	Hojas					
5						
5	8					
6	0	2	4			
6	8					
7	0	4				
7	5	6	6			
8	0	2	2	2	4	
8	5	5	5	7	8	8
9	1	1	2	2		
9	4	5	5	5		
10						
10						
11	0					

Observe ahora que esta subdivisión más fina entrega más detalles del conjunto de datos. ¿Qué puede concluir usted?

2.3.3. Tabulación de la información

Una de las mejores técnicas usadas en la estadística es la elaboración de

tablas o cuadros. En ellos se plasman las series estadísticas, una sucesión de datos referentes a un fenómeno observado a través del tiempo y del espacio.

Una ***serie cronológica*** es aquella donde se analiza un fenómeno a través del tiempo en un espacio determinado. Por ejemplo, el número de egresados de la UNAD en el período 1994-2004 (ver tabla 2.1.)

Tabla 2.1.
Número de egresados de la UNAD en el período 1994-2004

Año	Número de egresados
1994	338
1995	424
1996	556
1997	971
1998	1358
1999	2119
2000	3328
2001	4357
2002	3400
2003	3697
2004	4774
Total	25322

Una ***serie espacial*** es aquella donde se estudia un fenómeno a través del espacio en un tiempo determinado. Por ejemplo, el total de estudiantes de la UNAD en la Zona Occidente en el primer semestre de 2005.

Tabla 2.2.
Clasificación de estudiantes por CEAD en la Zona Occidente
durante el primer semestre de 2005

CEAD	Número de estudiantes
Medellín	1507
Pereira	1850
La Dorada	350
Turbo	371
Total	4078

Una ***serie cualitativa*** es aquella donde se estudia un atributo o característica de la población investigada, independiente del tiempo y del espacio.

Por ejemplo, los empleados de una empresa clasificados por cargo.

Tabla 2.3.
Clasificación de empleados por cargo

Cargo	Cantidad
Administrador	1
Jefe de producción	1
Contador	1
Secretaria	2
Supervisor	5
Operario	45
Vigilante	3
Total	58

Una ***serie cuantitativa*** es aquella donde se expresa numéricamente un atributo o característica de la población en estudio, independiente del tiempo y del espacio. Por ejemplo, la estatura en centímetros de un grupo de estudiantes de quinto grado.

Tabla 2.4.
Clasificación de la estatura de los estudiantes de un grupo de quinto grado

Estatura (en centímetros)	**Número de estudiantes**
125 — 129	1
129 — 133	4
133 — 137	9
137 — 141	24
141 — 145	28
145 — 149	22
149 — 153	12
Total	100

2.3.4. Distribuciones de frecuencias

Una tabla de frecuencias es otro de los formatos que se usan para organizar y resumir los datos. Para comprender la técnica de la distribución de frecuencias y dominar sus aplicaciones, es necesario manejar algunos conceptos con suficiente claridad. Y para ello se parte del concepto básico en la distribución de frecuencias: el número de veces que un dato se repite de un conjunto de datos se le denomina ***frecuencia***.

Un conjunto de datos puede organizarse de diferentes maneras. Una de ellas es construir una ***distribución de frecuencias simple***, que indica las frecuencias con que aparecen los datos. Es este el tipo de distribución de frecuencias más utilizado en estadística, pues permite conocer el comportamiento de un conjunto determinado de datos y no se ocupa de detalles individuales que, en muchos casos, poco puede ayudar en la toma de decisiones.

EJEMPLO 2.3.

Continuando con la serie de datos del ejemplo 2.1., organice los datos en una distribución de frecuencias simple.

Tabla 2.5.
Distribución de frecuencias simple de latidos cardiacos de 30 personas

Velocidad de pulsaciones	Fre-cuencia	Velocidad de pulsaciones	Fre-cuencia	Velocidad de pulsaciones	Fre-cuencia
58	1	76	2	94	1
59	0	77	0	95	3
60	1	78	0	96	0
61	0	79	0	97	0
62	1	80	1	98	0
63	0	81	0	99	0
64	1	82	3	100	0
65	0	83	0	101	0
66	0	84	1	102	0
67	0	85	3	103	0
68	1	86	0	104	0
69	0	87	1	105	0
70	1	88	2	106	0
71	0	89	0	107	0
72	0	90	0	108	0
73	0	91	1	109	0
74	1	92	2	110	1
75	1	93	0		

Observe que esta manera de agrupar se vuelve engorrosa cuando se tienen muchísimos datos. Otra forma de organizar un conjunto de datos es construir una ***distribución de frecuencias agrupadas***, que indica las frecuencias con que aparecen los datos agrupados en lo que se denomina ***intervalos de clase***. Cada

intervalo de clase está limitado por dos valores, llamados ***límites de clase*** (límite inferior y límite superior). La diferencia entre estos límites en cada intervalo de clase se denomina ***ancho***, ***tamaño*** o ***amplitud*** del intervalo de clase.

Clase es, entonces, un grupo que presenta una característica común cuantificable del conjunto de datos. El valor correspondiente al punto medio de un intervalo de clase es la ***marca de clase*** y su valor es igual a la mitad de la suma de los límites de clase del intervalo de clase. Y se interpreta como el valor que corresponde asignar a cada uno de los elementos del intervalo de clase.

El ***rango*** o ***recorrido*** es la diferencia entre los valores extremos de todo el conjunto de datos; en él se encuentran distribuidos todos los datos.

En la construcción de la distribución de frecuencias se deben responder a estos interrogantes fundamentales: ¿Cuántos intervalos de clase crear?, ¿Cuál debe ser el tamaño de cada intervalo?, ¿Qué propiedades posee cada intervalo? Las siguientes pautas resuelven estas inquietudes, además en el ejemplo 2.4 estas serán despejadas.

- Hallar el rango (***R***) o recorrido del conjunto de datos.

- Seleccionar el número de intervalos de clase (***k***). Este número depende de la cantidad de datos disponibles. Una de las técnicas usadas es la ***Regla de Sturges*** (desarrollada por H. A. Sturges en 1926). Esta regla afirma que el número de intervalos de clase (***k***), viene dado por:

  $$k = 1 + 3.322 \log n$$

  donde ***n*** es el tamaño de la muestra. Si de este cálculo resulta un número decimal, éste de redondearse al entero superior.
  Esta fórmula ha sido usada para obtener los números de intervalos de clase que aparecen en la tabla 2.6. y que permite sugerir el número de intervalos de clase que debe usarse de acuerdo al tamaño de la muestra. De esta manera, el cálculo del número de intervalos de acuerdo al tamaño de la muestra, puede determinarse bien por la Regla de Sturges o bien por la tabla 2.6.

- Hallar el ancho o amplitud del intervalo de clase (***A***). Los intervalos de clase tienen por lo general el mismo ancho, de modo que al fijarse el número de clases se obtiene éste por una operación aritmética simple:

  $$A = \frac{R}{k}$$

donde $\boldsymbol{R}$ es el rango o recorrido y $\boldsymbol{k}$ es el número de clases. Si este cociente no es un entero, conviene redondear al entero superior. De manera que el rango es alterado y requiere, por tanto, efectuar un ajuste:

$$R^* = (A)(k)$$

- Con este nuevo rango, se tendrá entonces un exceso que deberá distribuirse entre el límite superior y el límite inferior. Este exceso es calculado restando el rango del nuevo rango.

$$Exceso = R - R^*$$

  Este valor debe distribuirse lo más equitativo posible, esto no quiere decir que sea repartido en partes iguales a los datos extremos, se trata de distribuir el exceso entre el límite inferior y el límite superior de modo que sea considerado la tendencia general de los datos.

- Formar los intervalos de clase. Se agrega $A-1$ al límite inferior de cada clase, iniciando por el límite inferior del rango.

- Fijar los ***límites reales*** de cada intervalo de clase. Dado que los intervalos de clase son mutuamente excluyentes, es decir, no permiten ambigüedad en los límites cuando estos se repiten como inferior de un intervalo y como superior en el siguiente intervalo, se determinan los límites reales de clase. Estos corresponden al punto medio entre el límite superior de una clase y el límite inferior de la clase siguiente.
  En muchos casos se permite que se repita el límite superior de una clase y el límite inferior de la clase siguiente, haciendo la salvedad de cuál clase será tomada por dicho límite. En general, es considerado el límite superior de la clase como valor de esta.

- Determinar la frecuencia de clase. Contando el número de observaciones que cae dentro de cada intervalo de clase.

- Construir la tabla de distribución de frecuencias agrupadas.

Tabla 2.6.
Número de intervalos de clases sugerido en función del tamaño de la muestra

Tamaño muestral	Número de intervalos de clase
Menos de 16	Datos insuficientes
16 – 31	5
32 – 63	6
64 – 127	7
128 – 255	8
256 – 511	9
512 – 1023	10
1024 – 2047	11
2048 – 4095	12
4096 - 8190	13

EJEMPLO 2.4.

Para los datos del ejemplo 2.1. elabore una tabla de distribución de frecuencias agrupada. Para esto, se seguirán los pasos propuestos:

- **Rango** = 110 – 58 = 52

- **Número de clases**. Aplicando la Regla de Sturges:

$$k = 1 + 3.322 \log 30 = 5.91 \approx 6$$

  Si se usa la tabla 2.6., esta indica que deben usarse 5 clases. Queda a criterio del investigador la decisión. En este caso se trabajará con el resultado que arroja la Regla de Sturges.

- **Amplitud de los intervalos de clase**.

$$A = \frac{52}{6} = 8.67 \approx 9$$

- Como se ha redondeado, debe hallarse el **nuevo rango**:

$$R^* = (9)(6) = 54$$

- Existe pues un **exceso** de 2, [54 – 52 = 2]. Este exceso debe distribuirse quitando 1 al límite inferior y agregando 1 al límite superior:

$$X_{mín} = 58 - 1 = 57$$
$$X_{máx} = 110 + 1 = 111$$

Si en el cálculo del exceso, este hubiera sido un número impar, la distribución entre los límites se calcularía considerando hacia dónde se agrupan más los datos. En este caso, los datos tienen una mayor tendencia hacia el límite inferior de modo que el exceso mayor se repartiría en él.

- **Intervalos de clase**. Se agrega $A - 1 = 9 - 1 = 8$ al límite inferior de cada clase, iniciando por el límite inferior del rango. Así:
  57 + 8 = 65
  66 + 8 = 74
  75 + 8 = 83
  84 + 8 = 92
  93 + 8 =101
  102+8 =110

- **Límites reales**. 56.5, 65.5, 74.5,…, 110.5. Que se obtiene de calcular la suma de cada límite y dividirlo entre dos. Así:

$$\frac{56+57}{2} = 56.5 \quad \frac{65+66}{2} = 65.5 \quad \frac{74+75}{2} = 74.5 \quad ...$$

- Frecuencias de clase en cada intervalo.

Tabla 2.7.
Distribución de frecuencias agrupadas de la velocidad de pulsaciones

Intervalos de clase (Velocidad de pulsaciones)	**Frecuencia** (Número de personas)
56.5 – 65.5	4
65.5 – 74.5	3
74.5 – 83.5	7
83.5 – 92.5	11
92.5 – 101.5	4
101.5 – 110.5	1
Total	**30**

Al obtener la tabla de distribución agrupada como en el ejemplo 2.4., son muchos los interrogantes que continúan sin responderse como: ¿Qué porcentaje del grupo de personas evaluadas presentan pulsaciones entre 74.5 y 83.5? La

tabla 2.7. indica que son 7 personas pero ¿Qué porcentaje es ese? Y, más aún ¿Qué porcentaje de la muestra presentan, por ejemplo, pulsaciones menores de 92.5?

Cuando se habla de la frecuencia de una clase, se refiere a la ***frecuencia absoluta***, pero si ésta se da en términos del total de frecuencias se tiene entonces la ***frecuencia relativa***. Esta se obtiene en porcentaje al dividir la frecuencia de clase entre el número total de frecuencias (o tamaño de la muestra).

$$f_r = \frac{f}{n} \times 100$$

donde f_r es la frecuencia relativa de clase, f es la frecuencia absoluta de clase y n es el tamaño de la muestra. En la tabla 2.8. de distribución de frecuencias agrupadas de los datos del ejemplo 2.1., se calculan las correspondientes frecuencias relativas de cada intervalo de clase.

Tabla 2.8.
Distribución de frecuencias absolutas, relativas y acumuladas ascendentes de la velocidad de pulsaciones

Intervalos de clase (Velocidad de pulsaciones)	**Frecuencia** (Número de personas)	**Frecuencia relativa** (%)	**Frecuencia absoluta acumulada** *Ascendente*	**Frecuencia relativa acumulada** *Ascendente*
56.5 – 65.5	4	13.3%	4	13.3%
65.5 – 74.5	3	10%	7	23.3%
74.5 – 83.5	7	23.4%	14	46.7%
83.5 – 92.5	11	36.7%	25	83.4%
92.5 – 101.5	4	13.3%	29	96.7%
101.5 – 110.5	1	3.3%	30	100%
Total	**30**	**100%**		

La ***distribución de frecuencias acumuladas*** se construye con el cálculo de la ***frecuencia absoluta acumulada*** y la ***frecuencia relativa acumulada***. La primera es la acumulación sucesiva en forma descendente o ascendente de las frecuencias absolutas. Si la frecuencia absoluta acumulada es ascendente, la primera frecuencia absoluta corresponderá a la primera frecuencia absoluta acumulada. La segunda acumulada se obtiene sumando las dos primeras absolutas, y así sucesivamente. La última frecuencia absoluta acumulada corresponderá al número total de frecuencias.

De la misma manera, la frecuencia relativa acumulada es una acumulación sucesiva en forma ascendente o descendente de frecuencias relativas. Si es ascendente, la última frecuencia relativa acumulada tendrá un valor del 100%. En la tabla 2.8. se expresan estos tipos de frecuencia tomando los datos del ejemplo 2.1.

Esta tabla arroja información tan completa que permite concluir afirmaciones tales como:

- El 36.7% de las personas evaluadas registran pulsaciones entre el 83.5 y 92.5 y sólo el 3.3% registran valores altos, entre 101.5 y 110.5.
- De las 30 personas, 25 de ellas no superan registros de 92.5 pulsaciones; esto corresponde al 83.4% del total.

Construya la distribución de frecuencias absoluta descendente y relativa acumuladas descendente con los datos de la velocidad de pulsaciones. ¿Qué porcentaje de la muestra registra valores superiores a 92.5? ¿A cuántas personas corresponde? ¿Qué porcentaje registra valores de más de 75?

EJERCICIOS TEMA 2.3.

1. Los siguientes datos representan las calificaciones en una prueba de coordinación física aplicada a un grupo de 20 personas después de haber ingerido una cantidad de alcohol equivalente a 0.1% de su peso. Organice los datos como una combinación ordenada.

69	84	52	93	61	74	79	65	88	63
57	64	67	72	74	55	82	61	68	77

2. Elabore una lista de los valores de datos que aparecen en el diagrama de tallo y hoja siguiente.

Tallo	**Hojas**					
4	0	2	3			
5	1	1	8	9		
6	2	3	3	7	7	9
7	0					

3. En un estudio sobre el crecimiento de los varones se obtuvieron estas observaciones sobre el perímetro craneal en centímetros de un niño al nacer. Elabore un diagrama de tallo y hojas y haga un breve comentario de los resultados que este arroja.

33.1	34.6	34.2	35.1	34.2	35.6
34.5	35.8	34.5	34.7	34.3	35.2
33.7	36.0	34.2	33.6	34.6	34.3
33.4	34.9	33.8	34.7	35.2	34.6
33.7	34.8	33.9	34.2	35.1	34.2
36.5	34.1	34.0	36.1	35.3	34.3

4. Los siguientes datos muestran el número de huevos en cada uno de los nidos de 30 tortugas sobre la playa de Florida. Existen dos tipos definidos de tortugas en el área. ¿Un arreglo de tallo y hojas revela la existencia de dos poblaciones? ¿Lo hará uno de doble tallo?

206	167	175	204	123	138
197	187	193	124	137	141
142	192	197	109	126	127
181	171	163	146	124	184
101	201	133	141	152	132

5. Las siguientes son el número de llamadas semanal que recibe un call center.

1959	4534	7020	6725	6964	7428
2802	2462	4000	3378	7343	4189
2412	7624	1548	4801	737	5321
6837	8639	7417	6082	10241	962
5099	6627	4484	5633	4148	6588
6472	8327	8225	6142	12130	9166
5749	1801	4632	9359	8973	849
3894	5847	4327			

a. Organice los datos como una combinación ordenada.
b. Determine el dato mayor y el menor
c. Determine el rango
d. ¿Cuántas clases se necesitan para agrupar estos datos?
e. ¿Cuál es la amplitud mínima necesaria por clase para cubrir el intervalo, si se emplean el número de clases hallado en el numeral d?
f. Verifique si es necesario hallar un nuevo rango y hacer el ajuste de exceso.
g. Determine los intervalos de clase para este conjunto de datos
h. Halle los límites reales de dichos intervalos.
i. Construya la tabla de frecuencias absoluta, relativa y acumulada ascendente y descendente.

6. Los siguientes datos corresponden al total de ventas semanales (en cientos de dólares) de una tienda de accesorios para dama. Construya una tabla completa de distribución de frecuencias agrupadas. ¿Qué concluye?

192.3	192.1	98.7	99.1	99.6
102.3	191.5	93.1	102.8	96.4
102.1	97.8	97.6	95.4	94.2
90.5	103.4	92.9	102.5	97.3
99.8	96.3	113.2	98.5	114.1

7. Tome los datos del ejercicio 1 sobre las calificaciones en una prueba de coordinación física aplicada a un grupo de 20 personas después de haber ingerido una cantidad de alcohol equivalente a 0.1% de su peso y construya una tabla completa de distribución de frecuencias agrupadas.

8. Los siguientes son los números de venados observados en 72 sectores de tierra en un conteo de vida silvestre. Complete la siguiente tabla de distribución de frecuencias.

18	8	9	22	12	16	20	33	15	21	18	13
13	19	0	2	14	17	11	18	16	13	12	6
8	12	13	21	8	11	19	1	14	4	19	16
2	16	11	18	10	28	15	24	8	20	6	7
21	0	16	12	20	17	13	20	10	16	5	10
15	10	16	14	29	17	4	18	21	10	16	9

Intervalo de clase	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada ascendente	Frecuencia relativa acumulada ascendente
0 - 4					
5 – 9					
10 – 14					
15 – 19					
20 – 24					
25 – 29					
30 - 34					

9. La siguiente es la distribución de los pesos de 125 muestras de minerales recolectadas en una investigación de campo.

Peso en gramos	Número de especimenes
0.0 – 19.9	16
19.9 – 39.9	38
39.9 – 59.9	35
59.9 – 79.9	20
79.9 – 99.9	11
99.9 – 119.9	4
119.9 – 139.9	1
TOTAL	**125**

Si es posible, encuentre cuántas de las muestras pesan:

- **a.** Como máximo 59.9 gramos.
- **b.** Más de 59.9 gramos.
- **c.** Más de 80.0 gramos.
- **d.** 80.0 gramos o menos.
- **e.** Exactamente 70.0 gramos.
- **f.** Cualquier valor de 60.0 a 100 gramos.
- **g.** ¿Qué porcentaje pesa menos de 79.9 gramos?
- **h.** ¿Qué porcentaje pesa más 19.9 gramos?
- **i.** ¿Qué porcentaje pesa exactamente 39.9?

10. Tome los datos del ejercicio 3 sobre el perímetro craneal en centímetros de un niño al nacer y construya una tabla completa de distribución de frecuencias agrupadas. Emita conclusiones de peso a partir de lo obtenido.

11. La siguiente tabla corresponde a la estatura (en centímetros) de los estudiantes de un grupo de quinto grado. Complete la tabla de frecuencias agrupadas y a partir de este, emita conclusiones.

Estatura (en centímetros)	**Número de estudiantes**	**Frecuencia relativa**	**Marca de clase**	**Frecuencia absoluta acumulada ascendente**
125 — 129	1			
129 — 133	4			
133 — 137	9			
137 — 141	24			
141 — 145	28			
145 — 149	22			
149 — 153	12			
Total	100			

2.4. PRESENTACIÓN DE LA INFORMACIÓN

Anteriormente se mencionó que la organización y el resumen de la información son dos procesos distintos que se ejecutan en forma independiente. Ya se ha desarrollado todo cuanto tiene que ver con la organización de la información, se verá ahora lo que implica el resumen o presentación de la información. Se trata pues de conocer algunas técnicas de construcción de gráficas, que es la mejor manera para resumir una investigación estadística.

A continuación, se tratarán las partes más fundamentales de una gráfica y los aspectos a tener en cuenta para su construcción, luego se presentarán los distintos tipos de gráficas usadas más comúnmente en estadística entre las cuales se encuentran el histograma, el polígono de frecuencias, la ojiva, los gráficos de puntos, lineales, de barras y circulares y los pictogramas.

2.4.1. Componentes de una gráfica

Cuando se diseña una gráfica, sea esta cual fuere, deben tenerse en cuenta ciertos aspectos con el fin de mejorar su apariencia y mostrar con claridad lo que se quiera que ella refleje.

Una gráfica siempre debe poseer un ***título*** que indique la descripción del contenido de ella. En muchas ocasiones, es importante indicar la ***escala*** con la que se trabaja. Es decir, identificar los ejes coordenados (X y Y) e indicar sus magnitudes correspondientes. La escala se aplica para saber la dimensión del fenómeno graficado. Otro aspecto importante a tener en cuenta es la ***fuente*** de información, que indique de dónde han sido tomados los datos incluyendo el tipo de publicación, el año del registro y otros indicadores que resulten importantes para la investigación.

La forma y el tipo de la gráfica que se seleccione depende en gran parte del investigador o de quien la elabora, sin embargo debe tenerse en cuenta para quién va dirigida ésta, el lugar de exposición y otros factores de logística que intervienen en la decisión del mejor diseño. Existen ciertos principios generales que se deben tener en cuenta en el logro de una buena gráfica:

- Si en la investigación se tienen varias gráficas, estas deben estar enumeradas en forma consecutiva.
- Toda gráfica debe tener un título que aclare su contenido.
- En los diagramas, las líneas de la ordenada y la abscisa que llevan escala, deben ser más gruesas que las demás.
- La mejor gráfica es la más sencilla. Evite saturar la gráfica de datos o textos innecesarios. Haga uso de sólo lo estrictamente necesario.

- La gráfica no sustituye el cuadro o la tabla, debe ser el complemento.
- Toda gráfica debe ir acompañada de convenciones para identificar las características que se grafican.
- La lectura de la escala del eje horizontal se hace de izquierda a derecha y la del eje vertical se hace de abajo hacia arriba.
- La representación del hecho debe variar sólo en una dimensión.
- En toda gráfica se debe explicar la fuente de donde fueron obtenidos los datos, aclarar las escalas, leyendas, notas, llamadas y convenciones que ayuden a identificar e interpretar las características presentadas.
- Las gráficas nunca preceden al texto.

2.4.2. Diagrama de frecuencias

Los ***diagramas de frecuencia*** se representan por medio de líneas verticales, cuya altura está dada por los valores de las frecuencias, ya sean absolutas o relativas. Si la representación se refiere a las frecuencias acumuladas (absolutas o relativas), esta se hará por medio de líneas horizontales, ubicando en el eje vertical los valores de la frecuencia acumulada. Este último diagrama, denominado ***diagrama de frecuencias acumuladas***, genera una serie de líneas horizontales que dan la sensación de los peldaños de una escalera.

EJEMPLO 2.5.

Tabla 2.9.
Distribución de frecuencias simple
de visita al odontólogo de niños entre los 6 y 12 años

Edad del niño (Años)	**Frecuencia absoluta** (Número de visitas)	**Frecuencia absoluta acumulada**
2	3	3
3	1	4
4	2	6
5	3	9
6	7	16
7	9	25
8	4	29
9	0	29
10	1	30
11	0	30
12	0	30
Total	**30**	

Esta tabla de frecuencias indica las veces que un grupo de 30 niños de 6 a 12 años de edad visitó en los últimos 6 meses al odontólogo. Construya un diagrama de frecuencias absolutas y un diagrama de frecuencias absolutas acumuladas.

Figura 2.3.
Diagrama de frecuencias absolutas
de visita al odontólogo de niños entre los 6 y 12 años

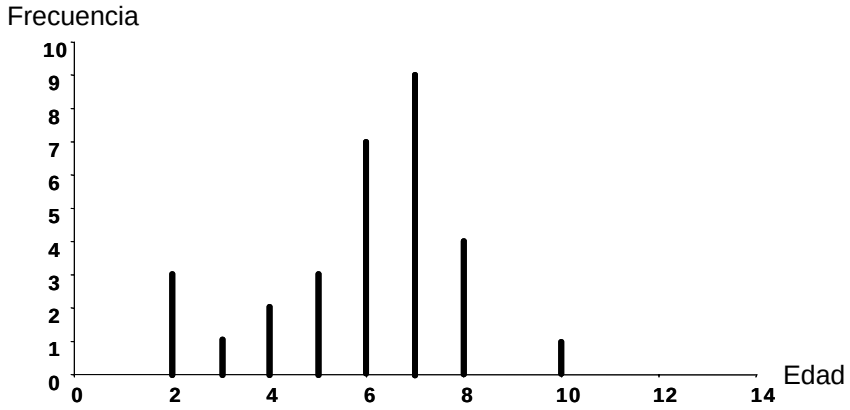
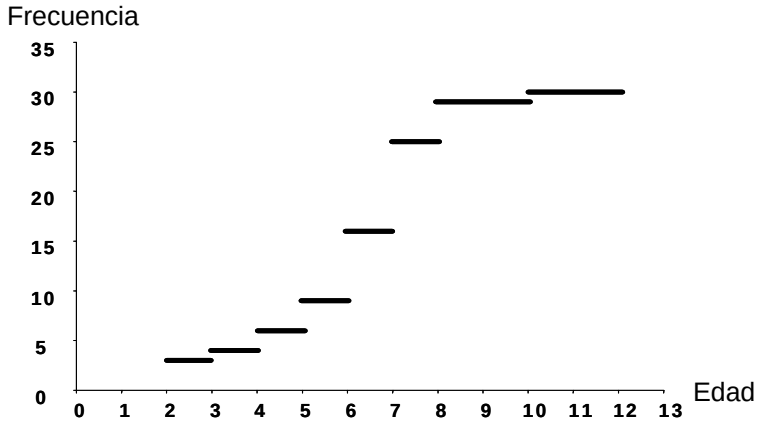


Figura 2.4.
Diagrama de frecuencias absolutas acumuladas
de visita al odontólogo de niños entre los 6 y 12 años



En las figuras 2.3. y 2.4. se reflejan los diagramas de frecuencia absoluta y frecuencia absoluta acumulada, respectivamente.
Obsérvese que a partir de la figura 2.3. rápidamente se puede concluir que los niños de 7 años de edad son los que más han asistido al odontólogo en los últimos seis meses de la muestra tomada.
De igual manera se percibe un agrupamiento a la izquierda de los datos, es decir no es simétrica la gráfica. Este tipo de gráficos suelen llamarse ***asimétricos***

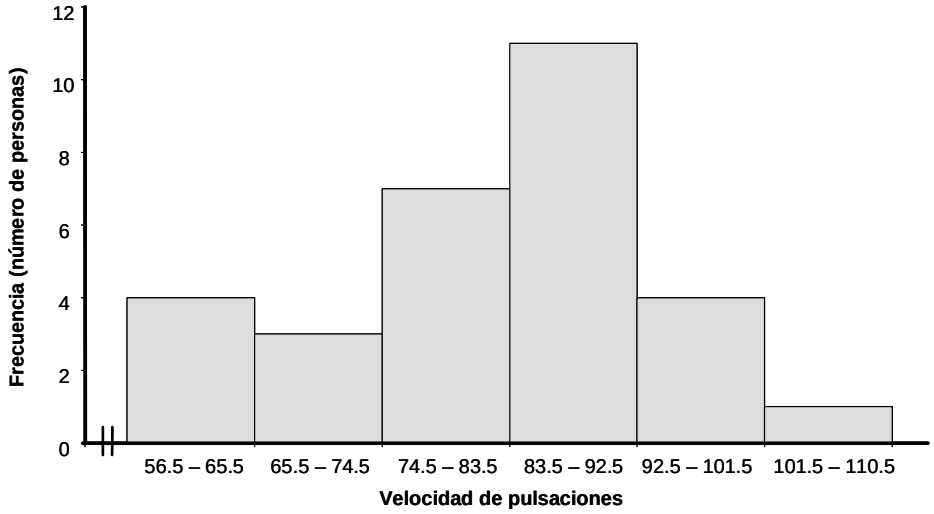
sesgados a la izquierda.
En la figura 2.4. las dos últimas líneas horizontales son de mayor tamaño que las demás, esto se debe a que no hay registro de niños que asisten al odontólogo con edades de 9, 11 y 12 años.
Se puede ver también que estas dos últimas líneas están menos separadas que las otras, pues el "salto" se debe a que existe un niño de la muestra de 10 años que sí ha asistido al odontólogo. En cambio, existe un gran "salto" a los 7 años, ¿sabe usted por qué?

2.4.3. Histograma de frecuencias

En el caso de las distribuciones de frecuencia agrupada, la forma de representación gráfica más común, se conoce con el nombre de ***histograma de frecuencias***. Estos se construyen representando los intervalos de clase en la escala horizontal y las frecuencias de clase (absolutas o relativas) en la escala vertical y trazando rectángulos cuyas bases equivalen a la amplitud de los intervalos de clase y sus alturas corresponden a las frecuencias de cada clase.

En la figura siguiente se registra el diagrama de frecuencias absolutas del grupo de datos del ejemplo 2.1. Nótese el origen o punto de partida de la variable es cero y luego aparece un ***corte*** o ***puente***, de manera que permite acortar la distancia entre el origen y el primer valor de la variable. Esta convención también puede usarse en el eje vertical u ordenada.

Figura 2.5.
Histograma de frecuencias absolutas de la velocidad de pulsaciones

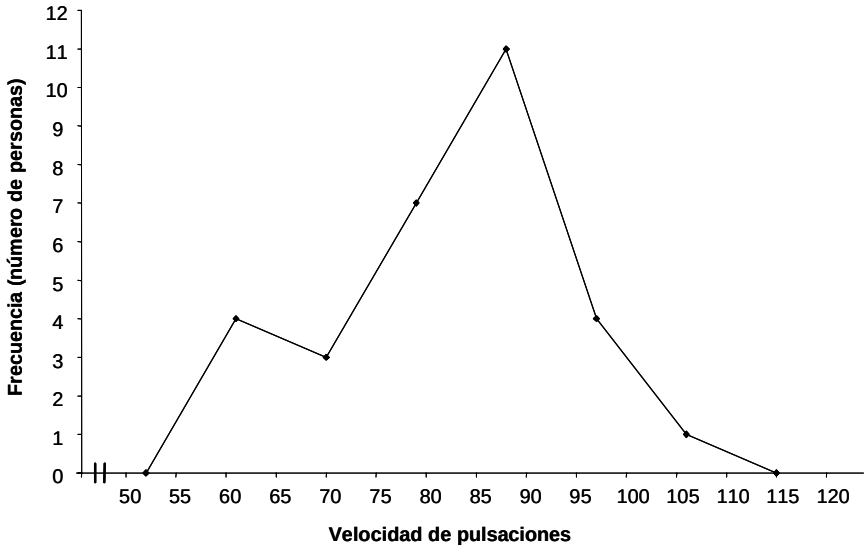


2.4.4. Polígono de frecuencias

Describe también la información de la distribución de frecuencias absolutas o relativas. Pero se grafican las marcas de clase de cada intervalo, generando una secuencia de puntos que se unen en segmentos de recta para formar un polígono, de ahí el nombre.

El polígono puede dibujarse sobre el histograma de frecuencias o de manera independiente. En el primer caso, se unen los centros de las bases superiores de los rectángulos; en el segundo caso, se unen los puntos de intersección de la abscisa, que corresponde a la marca de clase, con la ordenada correspondiente a la frecuencia relativa o absoluta. La figura 2.6. representa el polígono de frecuencias de los datos graficados en el histograma de la figura 2.5.

Figura 2.6.
Polígono de frecuencias absolutas
de la velocidad de pulsaciones



2.4.5. Ojiva

Contrario al polígono de frecuencias, la ojiva es una ***curva suavizada***[2]. Las curvas en estadística tienen diversas formas: estas se clasifican de acuerdo a la forma en ***simétricas*** y ***asimétricas*** siendo estas últimas sesgadas a la derecha o a la izquierda; y, según los máximos o picos que presenten, en ***unimodales***,

[2] Algunos autores consideran que la ojiva no es una curva suavizada, que está compuesta de segmentos rectilíneos. No se trata aquí de crear una discusión sobre ello pero queda al lector la decisión si elabora la ojiva como curva suavizada o como la unión de segmentos de líneas. En este módulo se trabajará como curva suavizada.

bimodales o ***multimodales.***

La ojiva es el gráfico de una distribución de frecuencias acumuladas (relativas o absolutas) y puede ser descendente o ascendente. Ella permite presentar en un mismo gráfico, diferentes curvas lo que no permite el histograma de frecuencias. En el eje horizontal se ubican el límite superior de cada intervalo de clase y en el vertical, las respectivas frecuencias acumuladas, ya sean relativas o absolutas. Luego se unen estos puntos en una curva suavizada, partiendo desde el límite inferior del primer intervalo. Observe las siguientes figuras, que representan la ojiva ascendente y descendente de los datos tomados de velocidad de pulsaciones de una muestra de 30 personas (ejemplo 2.1.)

Figura 2.7.

Ojiva ascendente de la velocidad de pulsaciones

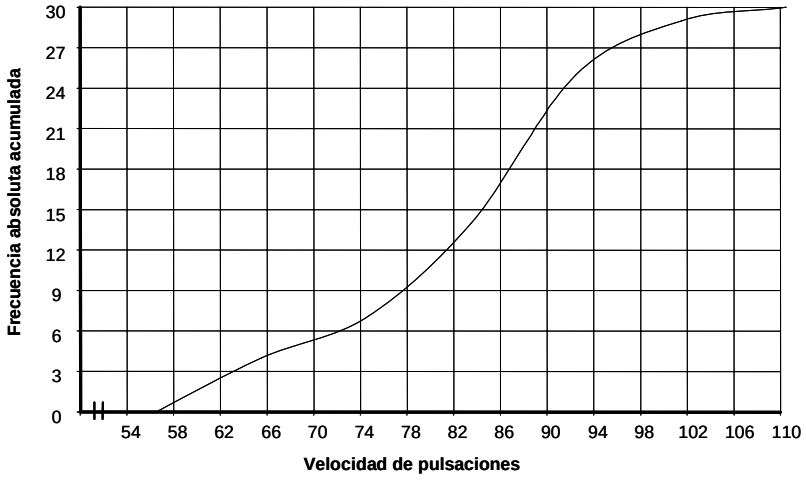
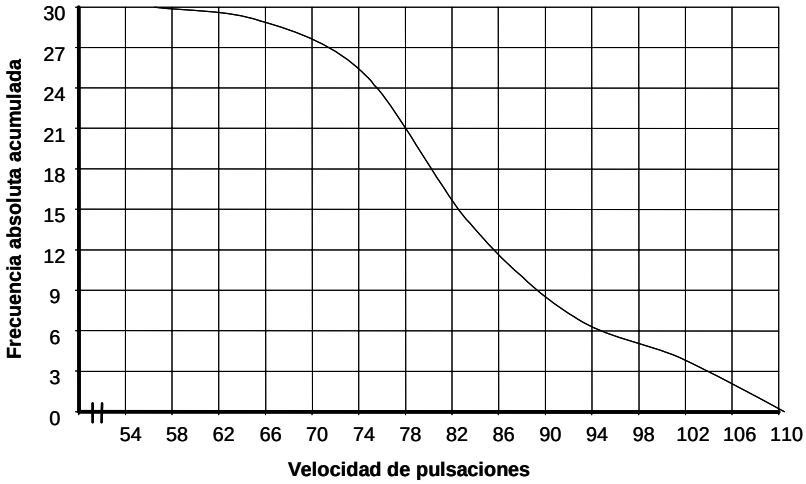


Figura 2.8.

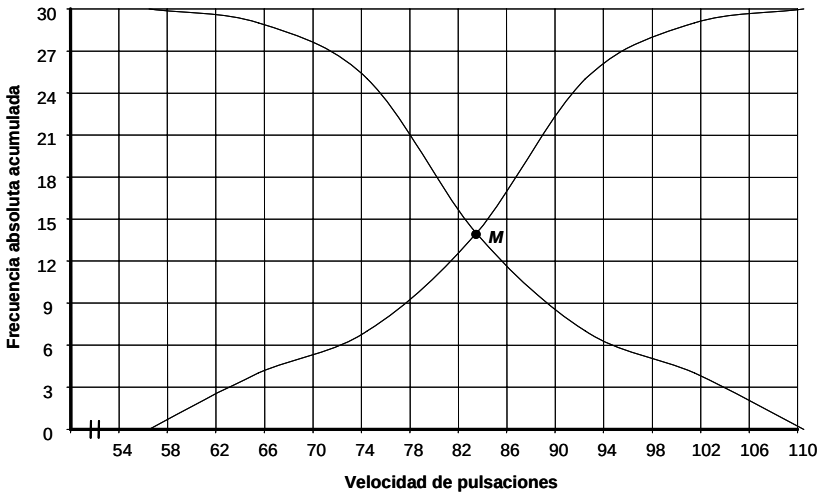
Ojiva descendente de la velocidad de pulsaciones



Si ambas ojivas se dibujan en un mismo gráfico, se obtiene la figura 2.9. Obsérvese que ellas se cortan en un punto ***M***, este punto se denomina ***mediana***, concepto que se discutirá en la siguiente unidad didáctica y que representa el valor del término de la mitad de la distribución.

Figura 2.9.

Ojiva ascendente y descendente de la velocidad de pulsaciones



2.4.6 Gráficos de línea

Está compuesta de segmentos de líneas que unen los pares ordenados a representar. Sirven para describir los cambios o fluctuaciones que sufre un fenómeno, generalmente durante un tiempo. Pueden ser ***simples***, cuando se dibuja una sola serie de datos o ***compuestos***, cuando se comparan dos o más series de datos, generalmente a través del tiempo (series cronológicas).

EJEMPLO 2.6.

Tabla 2.10.

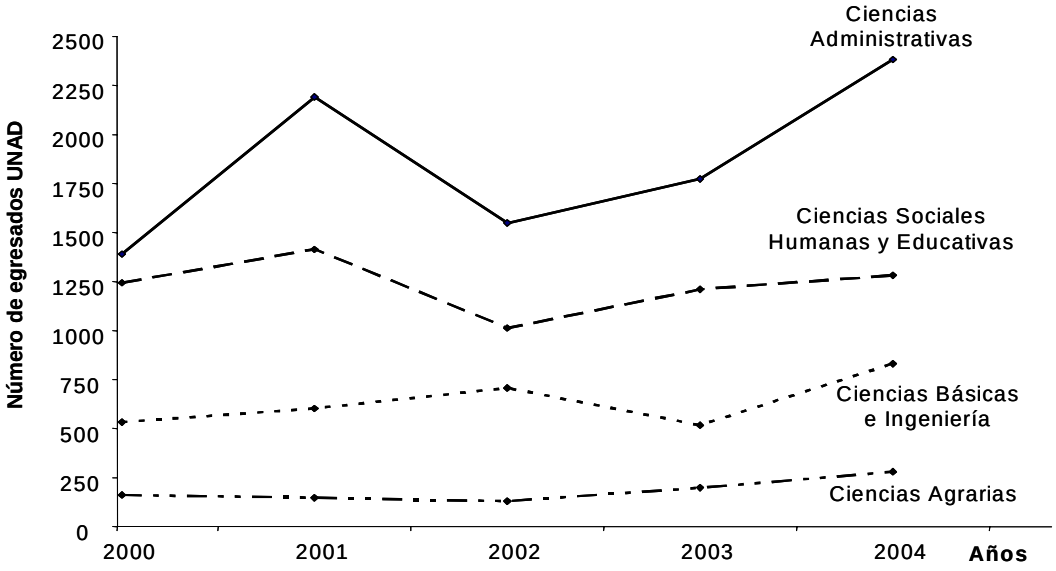
Egresados de la UNAD en el período 2000-2004

FACULTAD	2000	2001	2002	2003	2004
Ciencias Administrativas	1391	2192	1549	1773	2383
Ciencias Básicas e Ingeniería	533	603	708	517	830
Ciencias Agrarias	161	147	130	197	280
C. Soc. Humanas y Educ.	1243	1415	1013	1210	1281

La tabla 2.10. indica el número de egresados de la UNAD en el período 2000-2004, discriminados por facultad.

En el siguiente gráfico de puntos, se ve claramente el comportamiento y fluctuación en el tiempo de cada facultad respecto a sus egresados.

Figura 2.10.
Diagrama de líneas
Egresados de la UNAD en el período 2000-2004



De allí se puede ver cómo en 2004 hubo un aumento considerado en todas las facultades, de igual forma en 2002 disminuyó estrepitosamente el número de egresados en las facultades de Ciencias Administrativas y Ciencias Sociales Humanas y Educativas, mientras que en Ciencias Básicas e Ingeniería se daba un ascenso.
También se puede leer de este tipo de gráficos que, independiente de las fluctuaciones en el tiempo, la Facultad de Ciencias Administrativas es la que reporta mayor número de egresados anuales, seguida de Ciencias Sociales Humanas y Educativas, Ciencias Básicas e Ingeniería y por último Ciencias Agrarias.

2.4.7. Diagramas de barras

Es una de las gráficas más usadas para representar tanto características

cuantitativas como cualitativas. Es muy semejante al histograma de frecuencias, pero el diagrama de barras no requiere que la información esté agrupada en tablas de frecuencias.

Las barras son rectángulos con alturas proporcionales a las frecuencias o magnitudes correspondientes, pueden construirse en forma vertical u horizontal, sin embargo son más comunes las verticales; en este tipo de gráficos se ubica la variable o atributo en el eje horizontal y la altura está dada por los valores o cantidades que toma dicha variable.

El diagrama de barras se puede trabajar para describir una sola característica de la variable, ***diagrama de barras simple***, o bien describir y comparar dos o más características de ella de forma ***segmentada*** o ***agrupada***. Para diferenciar una característica de otra en la misma barra se recurre a diferenciarlas usando colores, sombrándolas o rellenándolas con tramas.

EJEMPLO 2.7.

La siguiente información corresponde a las ventas por departamento, al contado y a crédito, de un almacén de cadena en la ciudad de Bucaramanga en el mes de marzo de 2005. Los valores representan las ventas en millones de pesos.

Tabla 2.11.
Ventas por departamento al contado y a crédito en marzo de 2005

Departamento	Contado	Crédito	Total
Alimentos	200	120	320
Ropa	180	110	290
Calzado	150	90	240
Electrodomésticos	300	210	510

Los siguientes diagramas de barras verticales describen las ventas por departamento del almacén. Obsérvese que tanto la figura 2.11. y 2.12., aunque sean visualmente diferentes, ofrecen los mismos resultados. Inténtelo haciendo los diagramas de forma horizontal, ¿es clara la información? ¿Cuál tipo de diagrama de barras elegiría usted para una investigación? ¿Por qué?

Obsérvese además, en la figura 2.11., que también se puede graficar una barra más, la correspondiente al total de ventas, la cual permitiría una comparación eficiente de las ventas del almacén. ¿Cómo sería esta gráfica?

Figura 2.11.
Diagrama de barras agrupadas de las ventas
por departamento al contado y a crédito en marzo de 2005

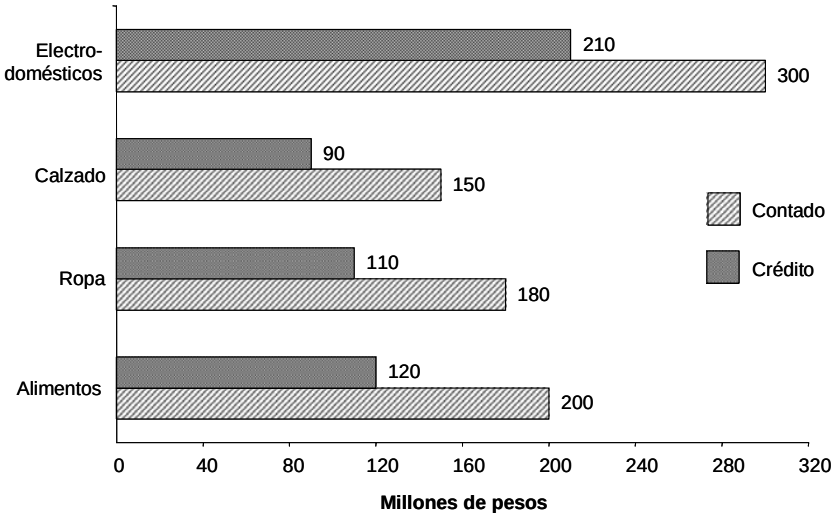
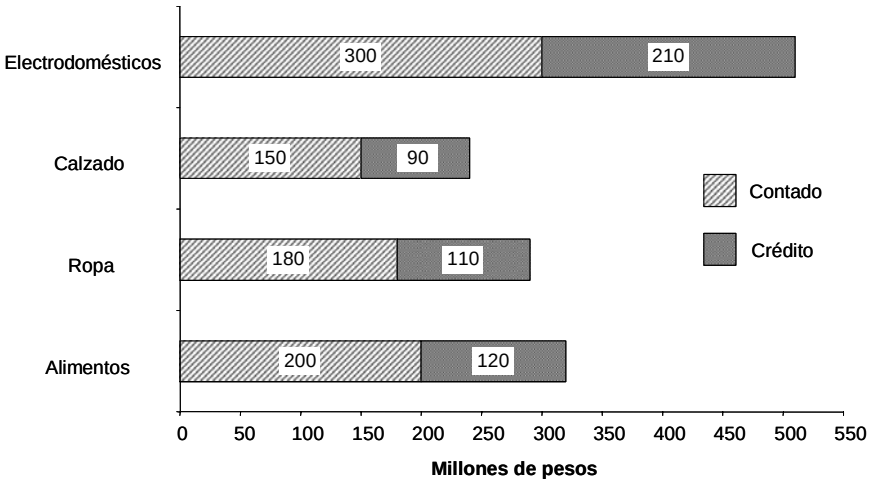


Figura 2.12.
Diagrama de barras segmentadas de las ventas
por departamento al contado y a crédito en marzo de 2005



Construya una tabla de frecuencias relativas para los datos de la tabla 2.11. y con ella elabore por lo menos dos diagramas de barra diferentes en los que muestre el porcentaje de ventas de contado y a crédito alcanzadas durante ese mes en el almacén de cadena para cada uno de los departamentos evaluados. Elabore una pequeña síntesis de los resultados que arrojan las gráficas que ha construido.

2.4.8. Diagrama circular

Es otro tipo de gráfico que permite observar los componentes de un total, como sectores de un círculo. Se utiliza para representaciones gráficas de distribuciones porcentuales. Es una forma efectiva de representar distribuciones de frecuencias en las que la característica es cualitativa.

Los ángulos de los sectores son proporcionales a los componentes del total. Se construye subdividiendo los 360º de un círculo, proporcionalmente al número o al porcentaje de cada una de las clases en que se ha dividido la observación. Una mayor apreciación se logra coloreando distintivamente los sectores o dándole una trama a cada sector.

EJEMPLO 2.8.

En una entrevista masiva de una multinacional, asistieron 1250 personas con expectativas de emplearse. De ellas el 50% eran casados, 25% solteros, 15% separados y 10% en unión libre. Si se quisiera mostrar en un diagrama circular estas proporciones, se debe tener en cuenta que los 360º del círculo equivalen al 100%, debe pues plantearse una regla de tres simple:

Porcentaje		**Grados**
100%		360º
50%	⟶	X

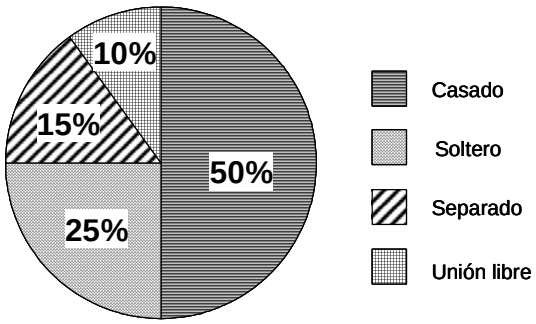
Donde:

$$X = \frac{50 \times 360}{100} = 180^{o}$$

De la misma manera, el 25% equivale a 90º en el círculo, 15% a 54º y 10% a 36º. Compruébelo. Así pues, se grafica el diagrama circular:

Figura 2.13.

Diagrama circular para el estado civil de 1250 aspirantes a empleo



Este tipo de gráficos es inconveniente cuando se tienen varias partes y cada una representa una pequeña proporción o cuando son muchas las partes que se van a representar. Si se le quiere emplear en secuencias cronológicas, se dibujan círculos de igual radio, tantos como años, meses o días se quieran representar en la secuencia, mostrando en cada uno la correspondiente distribución porcentual.

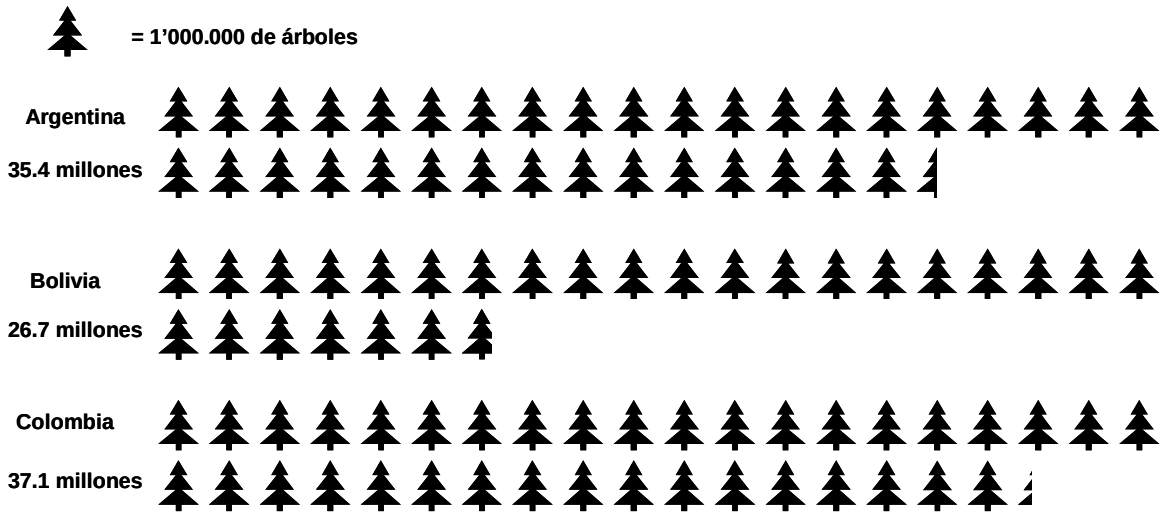
2.4.9. Pictogramas

Es una forma de representar los datos por medio de símbolos o dibujos donde cada uno representa la misma información con un valor fijo. Los pictogramas son usados comúnmente en el diseño publicitario, ya que se consideran más expresivos. Así es como se encuentran pictogramas señalando la población de un país, donde una figura humana representaría un millón de personas, por ejemplo.

En la siguiente figura se indica por medio de un pictograma los millones de árboles talados en Argentina, Bolivia y Colombia. Obsérvese que un árbol representará un millón de árboles talados anuales. Si la cantidad no es exacta, se presenta una fracción de la figura.

Figura 2.14.

Pictograma para el número de árboles talados en Argentina, Bolivia y Colombia



2.4.10. Mapas estadísticos o cartogramas

Este tipo de gráficos muestra la información cuantitativa o cualitativa sobre

bases geográficas dentro de las cuales se ubican símbolos o figuras como puntos, barras, círculos, colores, etc. Es muy común en la prensa o boletines de información, cuando se indica por ejemplo, el informe del estado del tiempo o, en un mapa de Colombia, se indican con figuras humanas las zonas en conflicto o en disputa con los diversos grupos armados del país.

EJERCICIOS TEMA 2.4.

1. Construya un diagrama de frecuencias absolutas y de frecuencias absolutas acumuladas con los datos reportados en la tabla 2.1., sobre el número de egresados de la UNAD en el período 1994-2004.

2. Con la tabla de frecuencia construida en el punto 5 de los ejercicios del tema 2.3., sobre el número de llamadas semanal que se recibe en un call center, construya un histograma de frecuencias absolutas, un polígono de frecuencias absolutas y las respectivas ojivas ascendente y descendente.

3. Un profesor decide registrar el mes de nacimiento de cada uno de los 40 estudiantes del tercer grado. Construya una tabla de frecuencias relativa y un diagrama de barras para los datos recolectados.
Junio, julio, noviembre, abril, enero, febrero, septiembre, julio, agosto, septiembre, diciembre, julio, junio, noviembre, mayo, abril, febrero, agosto, junio, mayo, octubre, agosto, noviembre, enero, junio, abril, septiembre, diciembre, agosto, junio, julio, marzo, diciembre, marzo, junio, noviembre, septiembre, junio, marzo, noviembre.

4. Construya un diagrama de barras para la tabla 2.2. en donde se clasifica el número de estudiantes por CEAD en la Seccional Occidente durante el primer semestre de 2005. Elabore también un pictograma.

5. La siguiente tabla indica las superficies de los distintos continentes del mundo en kilómetros cuadrados (km^2). Represente estos datos en un diagrama circular.

Continente	Área en km^2
Asia	44`391.200
África	30`244.000
Norteamérica	24`247.000
Suramérica	17`821.000
Antártica	13`338.500
Europa	10`354.600
Oceanía	8`547.000

6. Elabore por lo menos dos gráficas adecuadas para presentar la siguiente información: Durante 5 meses un escritor escribió una novela de 198 páginas de la siguiente manera: en el primer mes, 10.5% del total; en el segundo mes, 12.3% del total; en el tercer mes; 20.8%, en el cuarto mes, 17.4% del total y en el último mes, el 39% restante.

7. Dibuje en un gráfico de línea las ventas de un almacén en el primer semestre del año para sus tres sucursales en el país. El reporte contable fue:

Cartagena: $3’452.000 en enero; $2’125.600 en febrero; $2`058.400 en marzo; $3`032.300 en abril; $4`875.600 en mayo; $5`468.700 en junio.

Medellín: $2’301.500 en enero; $2’100.600 en febrero; $1`998.400 en marzo; $2`932.700 en abril; $3`985.100 en mayo; $4`500.700 en junio.

Bogotá: $4’750.500 en enero; $3’400.100 en febrero; $2`985.600 en marzo; $3`002.700 en abril; $4`923.100 en mayo; $6`130.700 en junio.

Haga un pequeño reporte escrito de las fluctuaciones de venta en las tres sucursales al administrador del almacén.

El administrador del almacén le solicita conocer las ventas totales mes a mes y le pide que entregue un informe escrito y gráfico de los resultados. ¿Qué tipo de grafico usaría? Elabórelo y escriba un pequeño reporte.

8. Con la tabla de frecuencia construida en el punto 10 de los ejercicios del tema 2.3., sobre el perímetro craneal en centímetros de un niño al nacer, construya un histograma de frecuencias relativas, un polígono de frecuencias relativas y las respectivas ojivas ascendente y descendente.

INFORMACIÓN DE RETORNO DE LA UNIDAD

EJERCICIOS CAPÍTULO 1.

2.

a. No
b. Edad, peso, velocidad,… Todo cuanto pueda influir en su resistencia física.
c. Velocidad: variable continua. Goles: variable discreta.

3.

a. Continua
b. Continua
c. Continua
d. Discreta
e. Continua
f. Discreta

4.

Situa-ción	Población	Muestra	Unidad estadística	Datos	Variable	Tipo de variable
a	Estudiantes matriculados en la UNAD en ese año.	Los estudiantes matriculados en la Facultad de Ciencias Agrarias.	Cada uno de los estudiantes matriculados en ese año.	10458 estudiantes matriculados en la UNAD.	Número de estudiantes matriculados en la Facultad de Ciencias Agrarias.	Discreta
b	Las temperaturas registradas el 29 de junio de 2005 en Pereira.	Las temperaturas registradas el 29 de junio de 2005 en Pereira entre las 6 y las 18 horas.	Cada una de las temperaturas registradas en ese tiempo y espacio.	Temperaturas registradas, bien sea en ºC ó ºK.	Temperaturas registradas el 29 de junio de 2005 en Pereira entre las 6 y las 18 horas.	Continua
c	Hogares en la ciudad de Medellín.	250 hogares de Medellín.	Cada uno de los hogares.	X hogares de los 250 hogares encuestados.	Hogares que hacen uso adecuado de MIRS.	Discreta
d	Las exportaciones de café colombiano.	Las exportaciones mensuales de café colombiano en el 2004.	Millones de dólares por exportación mensual.	X millones de dólares mensuales.	Millones de dólares en exportaciones mensuales de café colombiano.	Continua

EJERCICIOS TEMA 2.3.

1.

52	57	61	64	67	69	74	77	82	88
55	61	63	65	68	72	74	79	84	93

2.

40	43	51	59	63	67	69
42	51	58	62	63	37	70

3.

Tallo	Hojas																		
33	.1	.4	.6	.7	.7	.8	.9												
34	.0	.1	.2	.2	.2	.2	.2	.3	.3	.3	.5	.5	.6	.6	.6	.7	.7	.8	.9
35	.1	.1	.2	.2	.3	.6	.8												
36	.0	.1	.5																

4.

Tallo	Hojas				
10	1	9			
11					
12	3	4	4	6	7
13	2	3	7	8	
14	1	1	2	6	
15	2				
16	3	7			
17	1	5			
18	1	4	7		
19	2	3	7	7	
20	1	4	6		

5. a.

737	2802	4632	6082	7020	8973
849	3378	4484	6142	7343	9166
962	3894	4801	6472	7417	9359
1548	4000	5099	6588	7428	10241
1801	4148	5321	6627	7624	12130
1959	4189	5633	6725	8225	
2412	4327	5749	6837	8327	
2462	4534	5847	6964	8639	

b. Dato mayor: 12130 Dato menor: 737

c. Rango=12130 – 737=11393

d. $k = 1 + 3.32 \log 45 = 6.49 \approx 7$

e. $A = \frac{11393}{7} = 1627{,}57 \approx 1628$

f. $R^* = 7 \times 1628 = 11396$

Exceso: 11396 – 11393=3

Este exceso se distribuye entre el límite inferior y el límite superior. Se distribuyen 2 en un extremo y 1 en el otro. La decisión la toma el investigador.

Puede ser: restar 2 al límite inferior y adicionar 1 al límite superior.

$737 - 2 = 735$ $\quad$ $12130 + 1 = 12131$

g. $A - 1 = 1628 - 1 = 1627$

735 + 1627 = 2362	(735, 2362)
2363 + 1627 = 3990	(2363, 3990)
3991 + 1627 = 5618	(3991, 5618)
5619 + 1627 = 7246	(5619, 7246)
7247 + 1627 = 8874	(7247, 8874)
8875 + 1627 = 10502	(8875, 10502)
10503 + 1627 = 12130	(10503, 12130)

h. Límites reales:

(734.5, 2362.5)
(2362.5, 3990.5)
(3990.5, 5618.5)
(5618.5, 7246.5)
(7246.5, 8874.5)
(8874.5, 10502.5)
(10502.5, 12130.5)

i.

Intervalos de clase	**Frec.**	**Frec. relativa** (%)	**Frec. abs. acumulada** *Ascendente*	**Frec. abs. acumulada** *Descendente*	**Frec. relat. acumulada** *Ascendente*	**Frec. relat. acumulada** *Descendente*
734.5 – 2362.5	6	13.3	6	45	13.3	100
2362.5 – 3990.5	5	11.1	11	39	24.4	86.7
3990.5 – 5618.5	10	22.2	21	34	46.6	75.6
5618.5 – 7246.5	12	26.7	33	24	73.3	53.4
7246.5 – 8874.5	7	15.6	40	12	88.9	26.7
8874.5 – 10502.5	4	8.9	44	5	97.8	11.1
10502.5 – 12130.5	1	2.2	45	1	100	2.2
Total	**45**	**100%**				

$R = 192{,}5 - 90{,}5 = 101$

6. $k = 1 + 3.322 \log 25 = 5{,}64 \approx 6$

$A = \frac{101}{6} = 16{,}83 \approx 17$

Intervalos de clase	Frec.	Frec. relativa (%)	Frec. abs. acumulada *Ascendente*	Frec. abs. acumulada *Descendente*	Frec. relat. acumulada *Ascendente*	Frec. relat. acumulada *Descendente*
90 – 107	20	80	20	25	80	100
107 – 124	2	8	22	5	88	20
124 – 141	0	0	22	3	88	12
141 – 158	0	0	22	3	88	12
158 – 175	0	0	22	3	88	12
175 - 192	3	12	25	3	100	12
Total	**25**	**100%**				

7. $R = 93 - 52 = 41$ $\quad k = 1 + 3.322 \log 20 = 5{,}32 \approx 5$

$A = \frac{41}{5} = 8{,}2 \approx 9$ $\quad R^* = 9 \times 5 = 45$

$Exceso = 45 - 41 = 4$

Intervalos de clase	Frec.	Frec. relativa (%)	Frec. abs. acumulada *Ascendente*	Frec. abs. acumulada *Descendente*	Frec. relat. acumulada *Ascendente*	Frec. relat. acumulada *Descendente*
49.5 – 58.5	3	15	3	20	15	100
58.5 – 67.5	6	30	9	17	45	85
67.5 – 76.5	5	23	14	11	70	55
76.5 – 85.5	4	20	18	6	90	30
85.5 – 94.5	2	10	20	2	100	10
Total	**20**	**100%**				

8.

Intervalo de clase	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada ascendente	Frecuencia relativa acumulada ascendente
0 - 4	2	7	9,72	7	9,72
5 – 9	7	10	13,89	17	23,61
10 – 14	12	20	27,78	37	51,39
15 – 19	17	22	30,56	59	81,94
20 – 24	22	10	13,89	69	95,83
25 – 29	27	2	2,78	71	98,61
30 - 34	32	1	1,39	72	100,00
Total		**72**	**100%**		

9.

a. 89

b. 36

c. No es posible

d. No es posible
e. No es posible
f. No es posible
g. 87.2%
h. 87.2%
i. No es posible

10. $R = 36{,}5 - 33{,}1 = 3{,}4$ $\qquad k = 1 + 3.322 \log 36 = 6{,}17 \approx 6$

$A = \dfrac{3.4}{6} = 0{,}56 \approx 0{,}6$ $\qquad R^* = 6 \times 0{,}6 = 3{,}6$

$Exceso = 3{,}6 - 3{,}4 = 0{,}2$

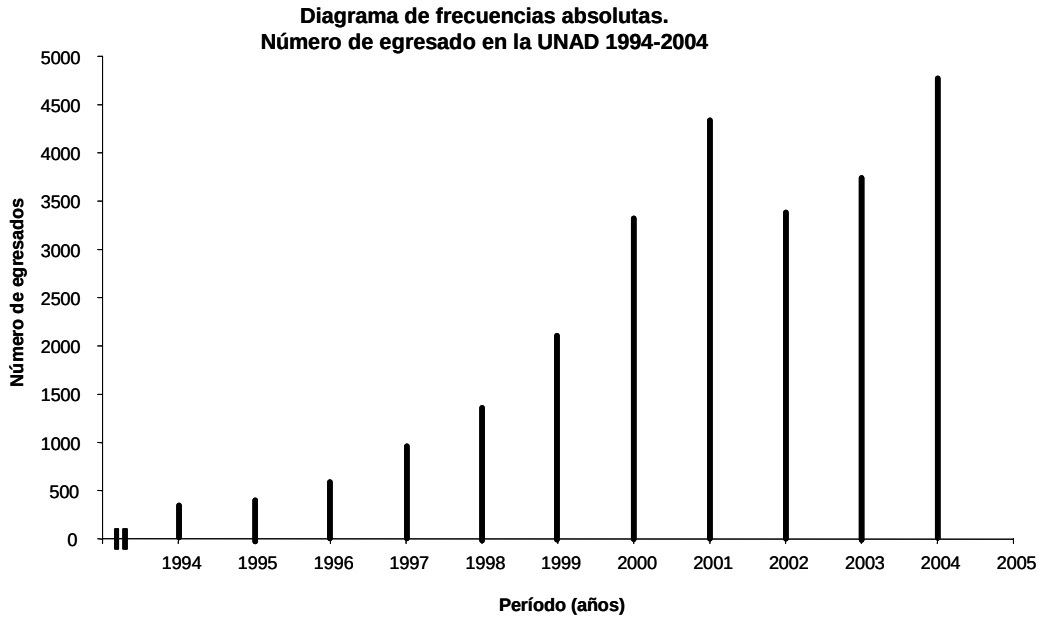
Intervalos de clase	**Frec.**	**Frec. relativa** (%)	**Frec. abs. acumulada** *Ascendente*	**Frec. abs. acumulada** *Descendente*	**Frec. relat. acumulada** *Ascendente*	**Frec. relat. acumulada** *Descendente*
32.95 – 33.55	2	5.6	2	36	5.6	100
33.55 – 34.15	7	19.4	9	34	25	94.4
34.15 – 34.75	15	41.7	24	27	66.7	75
34.75 – 35.35	7	19.4	31	12	86.1	33.3
35.35 – 35.95	2	5.6	33	5	91.7	13.9
35.95 – 36.55	3	8.3	36	3	100	8.3
Total	**36**	**100%**				

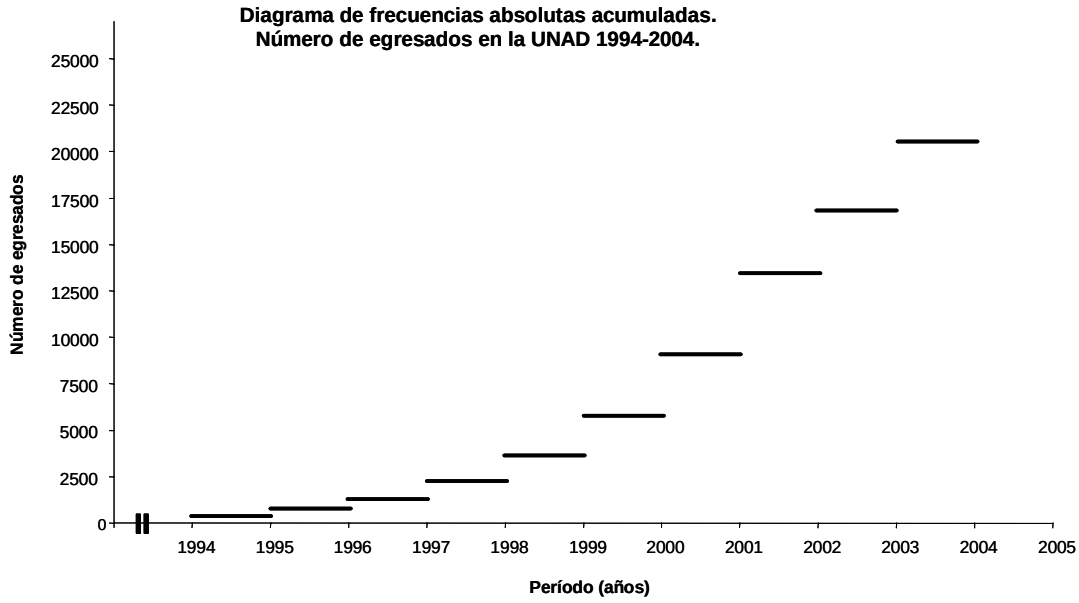
11.

Estatura (en centímetros)	**Número de estudiantes**	**Frecuencia relativa**	**Marca de clase**	**Frec. abs. acumulada ascendente**
125 — 129	1	1%	127	1
129 — 133	4	4%	131	5
133 — 137	9	9%	135	14
137 — 141	24	24%	139	38
141 — 145	28	28%	143	66
145 — 149	22	22%	147	88
149 — 153	12	12%	151	100
Total	100			

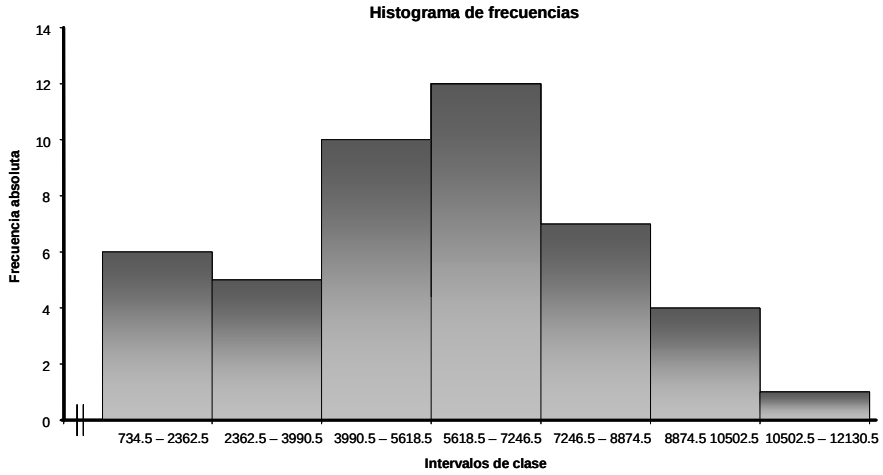
EJERCICIOS TEMA 2.4.

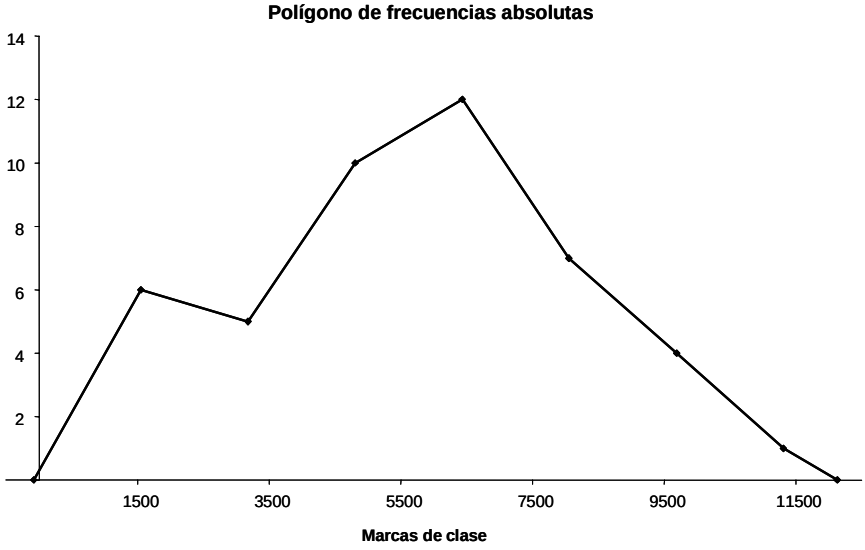
1.

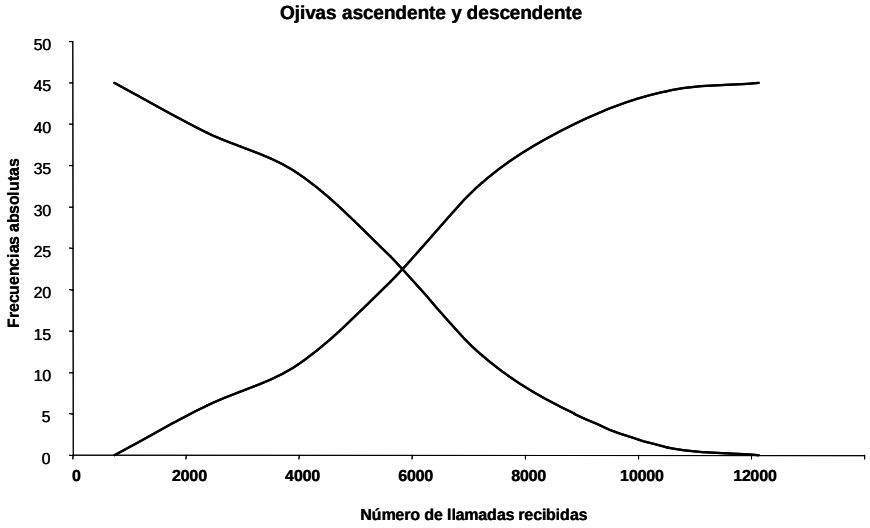




2.

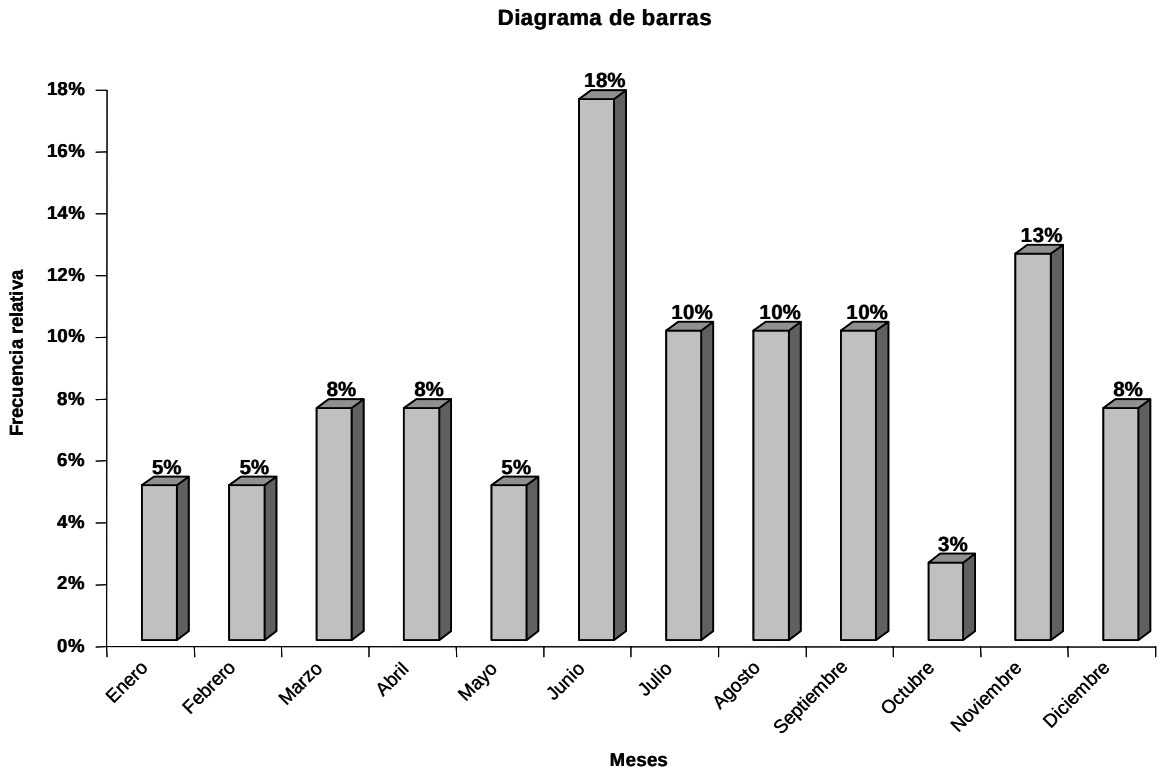

Histograma de frecuencias
Frecuencia absoluta
0
2
4
6
8
10
12
14
734.5 – 2362.5
2362.5 – 3990.5
3990.5 – 5618.5
5618.5 – 7246.5
7246.5 – 8874.5
8874.5 10502.5
10502.5 – 12130.5
Intervalos de clase


Polígono de frecuencias absolutas
2
4
6
8
10
12
14
1500
3500
5500
7500
9500
11500
Marcas de clase


Ojivas ascendente y descendente
Frecuencias absolutas
0
5
10
15
20
25
30
35
40
45
50
0
2000
4000
6000
8000
10000
12000
Número de llamadas recibidas

3.

Mes	Frecuencia absoluta	Frecuencia relativa
Enero	2	5%
Febrero	2	5%
Marzo	3	7,5%
Abril	3	7,5%
Mayo	2	5%
Junio	7	17,5%
Julio	4	10%
Agosto	4	10%
Septiembre	4	10%
Octubre	1	2,5%
Noviembre	5	12,5%
Diciembre	3	7,5%
Total	**40**	**100%**



4.

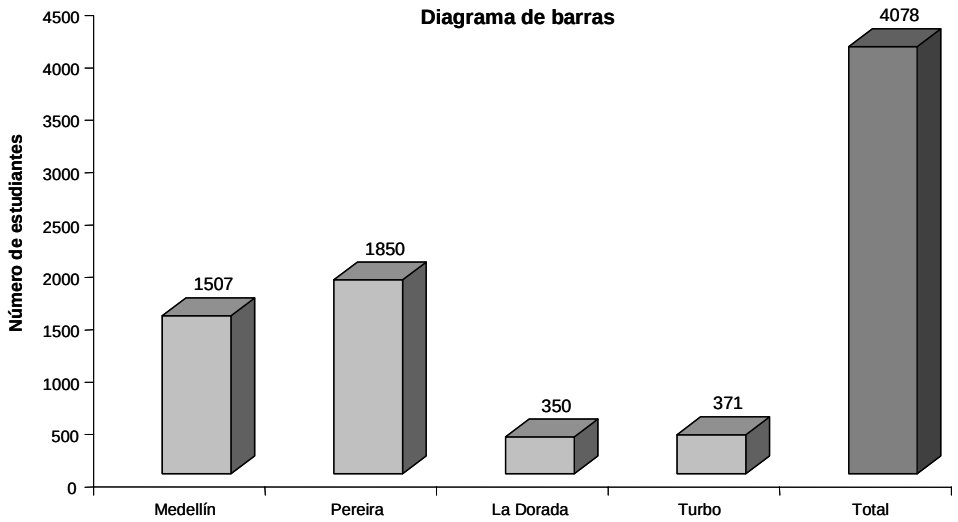


Diagrama de barras
Número de estudiantes
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1507
1850
350
371
4078
Medellín
Pereira
La Dorada
Turbo
Total

5.

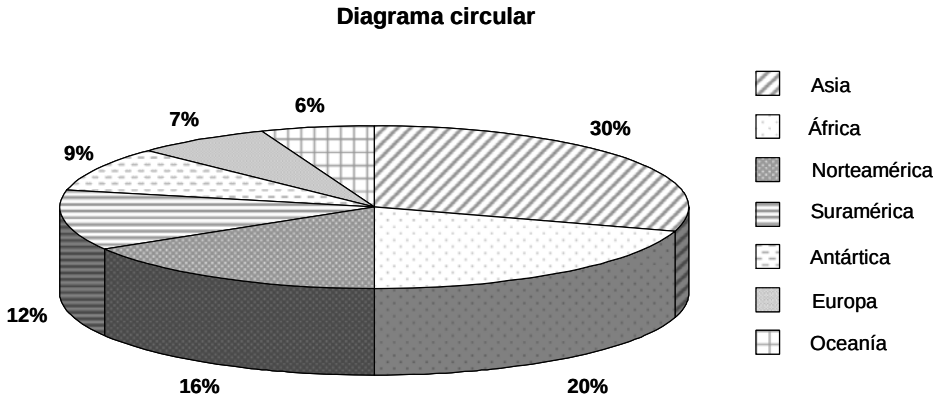


Diagrama circular
7%
6%
30%
9%
12%
16%
20%
Asia
África
Norteamérica
Suramérica
Antártica
Europa
Oceanía

6.

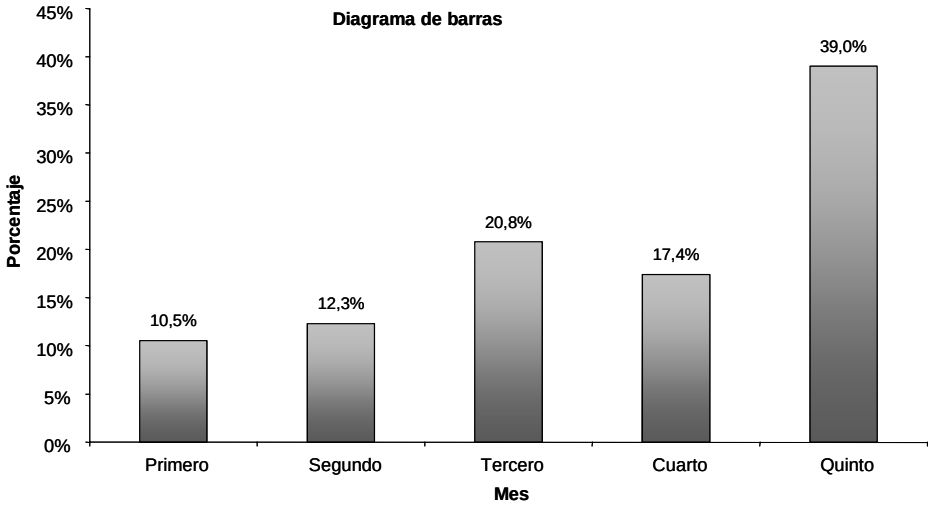


Diagrama de barras
Porcentaje
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
10,5%
12,3%
20,8%
17,4%
39,0%
Primero
Segundo
Tercero
Cuarto
Quinto
Mes

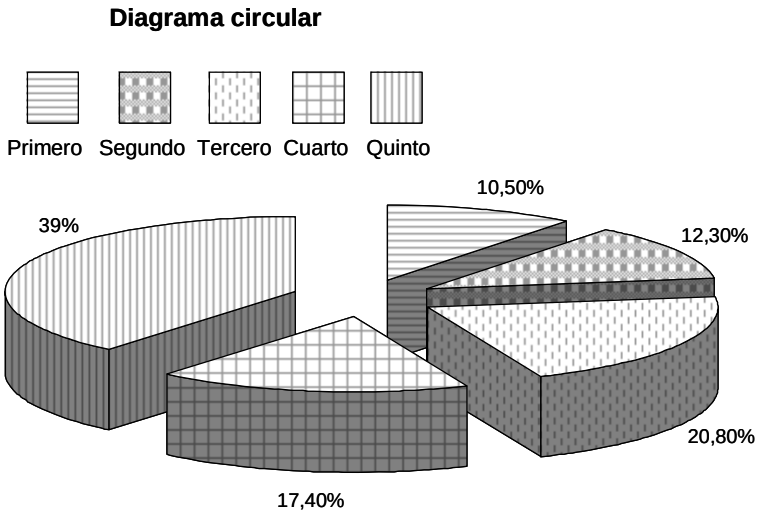


Diagrama circular
Primero
Segundo
Tercero
Cuarto
Quinto
39%
10,50%
12,30%
20,80%
17,40%

7.

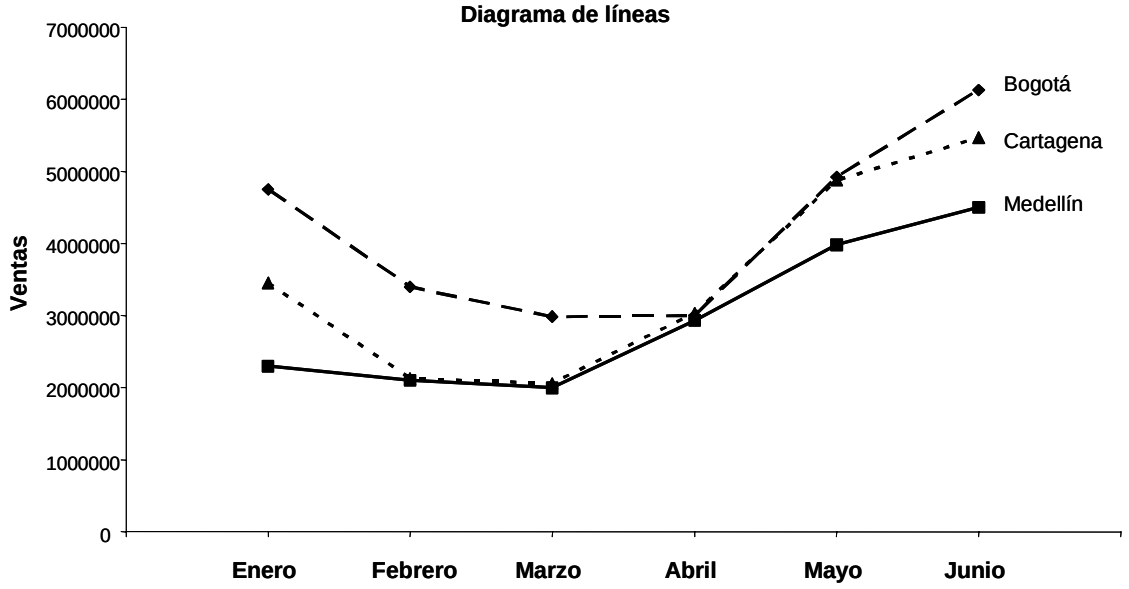
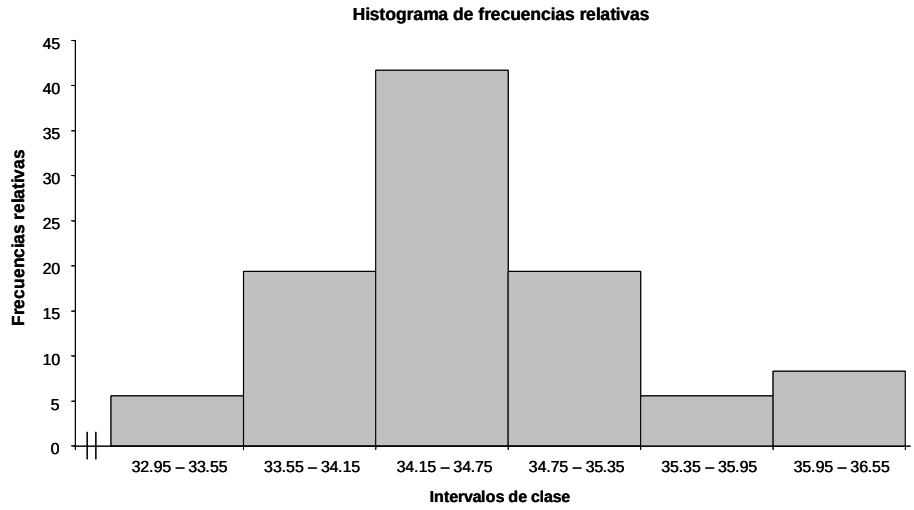
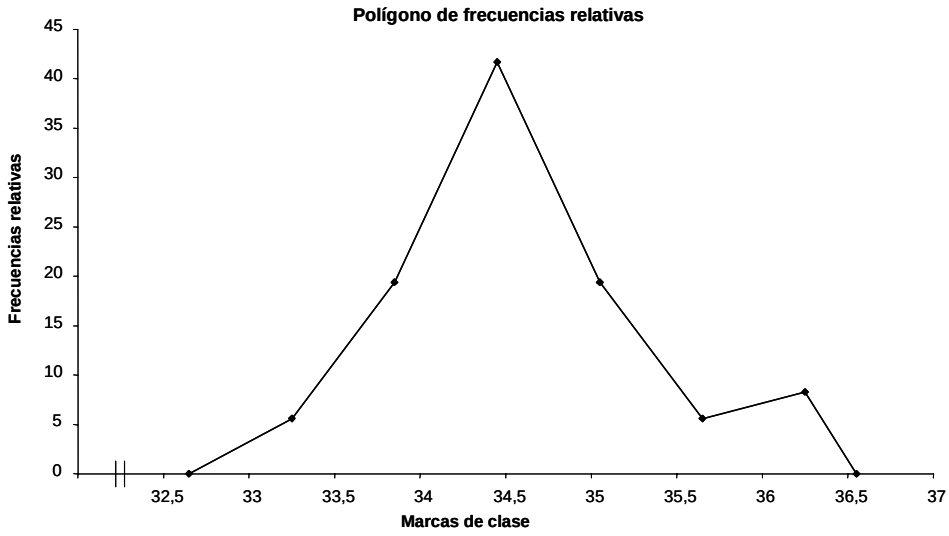
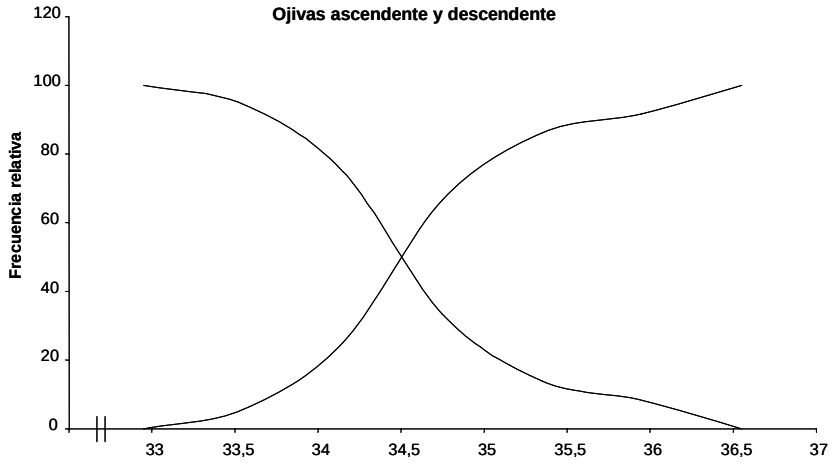


Diagrama de líneas
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Ventas
Enero
Febrero
Marzo
Abril
Mayo
Junio
Bogotá
Cartagena
Medellín

8.


Histograma de frecuencias relativas
Frecuencias relativas
45
40
35
30
25
20
15
10
5
0
32.95 – 33.55
33.55 – 34.15
34.15 – 34.75
34.75 – 35.35
35.35 – 35.95
35.95 – 36.55
Intervalos de clase


Polígono de frecuencias relativas
Frecuencias relativas
45
40
35
30
25
20
15
10
5
0
32,5
33
33,5
34
34,5
35
35,5
36
36,5
37
Marcas de clase


Ojivas ascendente y descendente
Frecuencia relativa
120
100
80
60
40
20
0
33
33,5
34
34,5
35
35,5
36
36,5
37

BIBLIOGRAFÍA DE LA UNIDAD

BEJARANO BARRERA, Hernán (1995). *Estadística Descriptiva.* Santa fe de Bogotá: UNISUR.

CHRISTENSEN, Howard B. (1999). *Estadística Paso a Paso.* México: Editorial Trillas.

MARTÍNEZ BENCARDINO, Ciro (2004). *Estadística Básica Aplicada.* Santa fe de Bogotá: ECOE Ediciones.

MARTÍNEZ BENCARDINO, Ciro (2003). *Estadística y Muestreo.* Santa fe de Bogotá: ECOE Ediciones.

MILTON, J. Susan (1999). *Estadística para biología y ciencias de la salud.* Madrid: McGraw Hill — Interamericana.

PORTUS GOVINDEN, Lincoyán (2001). *Introducción a la Estadística.* Segunda edición. Santa fe de Bogotá. McGraw Hill.

PORTILLA CHIMAL, Enrique (1980). *Estadística, Primer Curso.* Bogotá: Nueva Editorial Interamericana.

SPIEGEL, Murria R. (1991). *Estadística. Serie de compendios Schaum*. México: McGraw Hill.

SMITH, A. Stanley. (1992). *Curso de Estadística Elemental para las ciencias aplicadas.* Primera edición. Santa fe de Bogotá. Editorial Addison – Wesley Iberoamericana.

http://www.educarchile.cl/eduteca/estadistica/

http://www.aulafacil.com/CursoEstadistica/CursoEstadistica.htm

http://www.uaq.mx/matematicas/estadisticas/xu3.html

http://www.ing.unp.edu.ar/estadisitio/estaddes.htm

http://www.elosiodelosantos.com/descriptiva.html

http://thales.cica.es/rd/Recursos/rd98/Matematicas/01/matematicas-01.html

http://148.216.10.83/estadistica/descriptiva.htm

http://www.eneayudas.cl/estentrada.htm

http://www.universidadabierta.edu.mx/SerEst/MAP/METODOS%20CUANTITATIVOS/Pye/tema_11.htm

http://html.rincondelvago.com/estadistica_15.html

http://www.hrc.es/bioest/M_docente.html#tema2

http://personal5.iddeo.es/ztt/Tem/T11_Estadistica_Introduccion.htm

Unidad Didáctica Dos

MEDIDAS ESTADÍSTICAS

Unidad Didáctica Dos
MEDIDAS ESTADÍSTICAS

1. Medidas Estadísticas Univariantes

1.1. Medidas de tendencia central

1.1.1. Media aritmética
1.1.2. Mediana
1.1.3. Moda
1.1.4. Otras medidas de tendencia central

1.2. Medidas de dispersión

1.2.1. Rango o recorrido
1.2.2. Varianza
1.2.3. Desviación típica o estándar
1.2.4. Coeficiente de variación
1.2.5. Desviación media
1.2.6. Puntaje típico o estandarizado

1.3. Medidas de asimetría y apuntamiento

1.3.1. Asimetría
1.3.2. Apuntamiento o curtosis

2. Medidas Estadísticas Bivariantes

2.1. Regresión y correlación

2.1.1. Diagrama de dispersión
2.1.2. Regresión lineal simple
2.1.3. Correlación
2.1.4. Regresión múltiple

2.2. Números índice

2.2.1. Construcción de números índice
2.2.2. Tipos de números índice
2.2.3. Índices simples
2.2.4. Índices compuestos
2.2.5. Usos de los números índices

INTRODUCCIÓN A LA UNIDAD

La Unidad Didáctica 1 se dedicó a explicar los métodos que deben aplicarse en una investigación estadística tales como la planeación, recolección, organización y presentación de ella. Esta unidad tiene como propósito indicar otros métodos para medir e interpretar el comportamiento de un conjunto de datos dados.

Se ha visto que tanto las tablas como las muy diversas formas de graficar la información describen fenómenos de una población o muestra, pero no siempre lo hacen en forma satisfactoria; es allí donde se hace visible la importancia de las medidas estadísticas bien sean univariantes, en donde interviene una variable, o bivariantes cuando lo hacen dos.

Esta Unidad Didáctica se ha dividido en dos grandes capítulos: Medidas Estadísticas Univariantes y Medidas Estadísticas Bivariantes, obedeciendo al número de variables que intervienen en estos cálculos aritméticos. En el primer capítulo, se considerarán cuatro clases de medidas: de posición o de tendencia central, de dispersión o variabilidad, de asimetría o de deformación y de apuntamiento o curtosis.

En el segundo capítulo, se estudiará el comportamiento de dos variables, a fin de determinar si existe alguna relación entre sí y de cuantificar dicho grado de relación. Se desarrollarán aquí los conceptos de regresión y correlación de dos variables y el concepto y usos de los números índices.

Pero antes de iniciar con estos nuevos conceptos, se hace indispensable recordar algunas nociones aritméticas y algebraicas básicas en estadística, es por esto que se recomienda al lector iniciar el capítulo repasando la sumatoria como propiedad aritmética fundamental para entender las medidas estadísticas de una población o muestra. Todo cuanto tiene que ver con sumatoria y productoria puede ser repasado y consultado en el anexo A, que se encuentra al final del texto.

OBJETIVOS ESPECÍFICOS

- Ejecutar las operaciones indicadas por la notación sumatoria y productoria.
- Desarrollar destrezas para calcular algunas medidas de tendencia central.
- Interpretar las medidas de tendencia central y comprender sus aplicaciones.
- Comparar las medidas de tendencia central y seleccionar la más útil según las circunstancias.
- Desarrollar destrezas para calcular algunas medidas de dispersión.
- Comparar las medidas de dispersión y seleccionar la más útil para una determinada aplicación.
- Reconocer que las medidas de dispersión complementan la descripción que proporcionan las medidas de tendencia central.
- Interpretar y utilizar las medidas de dispersión.
- Identificar los tipos de asimetría y apuntamiento en una distribución de datos.
- Identificar hechos que admitan intuitivamente un comportamiento lineal simple.
- Interpretar y manejar los conceptos de regresión y correlación.
- Dibujar y aplicar gráficos de dispersión.
- Calcular el coeficiente de correlación entre dos variables.
- Calcular la ecuación de regresión para dos variables.
- Identificar e interpretar correctamente números índices.
- Desarrollar destrezas necesarias para elaborar y aplicar números índices en circunstancias específicas.

1. MEDIDAS ESTADÍSTICAS UNIVARIANTES

1.1. MEDIDAS DE TENDENCIA CENTRAL

Al ver la forma de representar los conjuntos de datos en histogramas y polígonos de frecuencia se puso de relieve un comportamiento peculiar de estos, y es el de mostrar una tendencia a agruparse alrededor de los datos más frecuentes, haciendo de esta forma que estas representaciones adquieran una forma de campana. Esta tendencia al agrupamiento de los datos hacia la parte central de los gráficos que los representan da lugar a lo que se conoce como medidas de tendencia central, correspondientes a la media, mediana y moda

1.1.1. Media aritmética

Es la medida más conocida y la más fácil de calcular. Se define como la suma de los valores de una cantidad dada de números dividido entre la cantidad de números.

$$\overline{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

donde:

n = cantidad de elementos

X_i = valor de cada elemento

$\overline{x}$ = media aritmética, o simplemente media

EJEMPLO 1.1.

El precio de la bolsa de un litro de leche en diferentes supermercados fue: $1.300, $1.350, $1.250, $1.400 y $1.325. El valor promedio o media aritmética es entonces:

$$\overline{x} = \frac{1.300 + 1.350 + 1.250 + 1.400 + 1.325}{5} = \$1.325$$

La media aritmética tiene la propiedad de asignar a cada elemento de la suma el mismo valor, o sea el valor promedio.

Si se conoce el valor de la media y el número n de elementos u observaciones, se puede conocer el valor de la suma total multiplicando la media

por el número de elementos. Esto es:

$$\sum_{i=1}^{n} X_i = n \cdot \bar{x}$$

EJEMPLO 1.2.

Las ventas de un almacén durante el primer semestre del año fueron $3'422.000; hallar el total de ventas de este período de tiempo.

Venta total primer semestre = 6 x (3'422.000) = $20'532.000

También puede suceder que los elementos que se analizan se encuentren agrupados, en este caso para encontrar el valor de la media aritmética se debe realizar la ponderación de estos elementos agrupados, es decir, encontrar el peso que le corresponde a cada valor. Esto da lugar a la ***media aritmética ponderada***.

EJEMPLO 1.3.

Un agricultor vende la cosecha de papas de la siguiente forma: 30 sacos a $256.000, 18 sacos a $264.000 y 25 sacos a $261.500. ¿Cuál es el precio promedio del saco de papa vendida por el agricultor?

Precio promedio saco de papa = $\frac{30(256.000)+18(264.000)+25(261.500)}{30+18+25}$ =$259.856

La media ponderada se halla al realizar el cociente entre la suma de los productos de los valores por sus respectivos pesos y la suma de los pesos. El caso general se expresa así:

$$\bar{x} = \frac{m_1 X_1 . + m_2 X_2 . + \ldots + m_n X_n .}{m_1 + m_2 + \ldots + m_n} = \frac{\sum_{i=1}^{n} m_i X_i}{\sum_{i=1}^{n} m_i}$$

Siendo X_1 X_2,… X_n, las cantidades ponderadas y m_1, m_2,,…, m_n los pesos o ponderaciones.

Un caso similar al anterior consiste en la ***media de una distribución de frecuencias agrupadas***, donde los pesos o ponderaciones corresponderían a las frecuencias de los valores de las marcas de clase, recordando que la marca de clase es el valor promedio de un intervalo de clase. Esta similitud entre la media de una distribución de frecuencias agrupadas y la media aritmética ponderada se muestra en el siguiente ejemplo.

EJEMPLO 1.4.

Dada la siguiente distribución de frecuencias agrupadas, calcule su correspondiente media aritmética:

Tabla 1.1.
Distribución de frecuencias agrupadas

Intervalo	Marca de clase X	Frecuencia f	$f.X$
16-20	18	4	72
21-25	23	6	138
26-30	28	7	196
31-35	33	5	165
36-40	38	3	114
Total		**25**	**685**

$$\bar{x} = \frac{\sum f \cdot X.}{\sum f} = \frac{685}{25} = 27.4$$

De lo anterior puede verse que:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{\sum f \cdot X}{n}$$

Dada la importancia que tiene el cálculo de la media aritmética y su frecuente uso, se hace necesario considerar algunas de sus propiedades:

- *La suma de las desviaciones respecto a la media aritmética es igual a cero.*

Una desviación es la diferencia que se presenta entre los valores que toma la variable y un valor constate, en este caso es la media aritmética. Esta propiedad, al igual que las demás, es válida para datos agrupados o no agrupados. Y en términos aritméticos ella plantea:

$$\sum (X - \bar{x}) = 0$$

Tenga en cuenta que cuando los datos están agrupados en una tabla de frecuencias, las desviaciones con respecto a la media deben ponderarse. Si la distribución es simétrica no hay necesidad de ponderar.

- *La suma de los cuadrados de las desviaciones respecto a la media es siempre menor que la suma de los cuadrados de las desviaciones con respecto a cualquier otro valor.*

  Esto quiere decir que sólo la media aritmética hace mínima la suma de los cuadrados de las desviaciones en torno a ella. Esta importante propiedad se retomará más adelante cuando se estudie regresión lineal y el método de los mínimos cuadrados para ajuste de curvas.

En síntesis, la media o promedio aritmético es la medida de tendencia central más comúnmente usada, además de ser la única de las medidas de tendencia central que permite un tratamiento algebraico. Sin embargo no siempre es recomendable usarla como un promedio, ya que es muy sensible a los valores extremos del conjunto de datos. Por otra parte, la media es ligeramente más difícil de calcular a mano que las otras medidas que se verán en seguida, puesto que requiere sumar todo el conjunto de datos, que bien podrían ser bastantes, y dividir entre el número de elementos del conjunto.

1.1.2. Mediana

Se define como el valor que divide una distribución de datos ordenados en dos mitades, es decir, se encuentra en el centro de la distribución.

La mediana se simboliza como ***Me***. Es menos usada que la media aritmética. Para su cálculo es necesario que los datos estén ordenados. Cuando la cantidad de datos es impar, fácilmente se identifica la mediana; pero cuando el número de datos es par, la mediana se calcula hallando el valor medio entre los dos valores centrales y no coincidirá con ninguno de los valores del conjunto de datos.

EJEMPLO 1.5.

a. Dados los valores: 19, 15, 23, 28, 14, 26, 18, 20, 30, determinar su media. Lo primero que debe hacerse es ordenar los datos:

14 15 18 19 **20** 23 26 28 30

Como el número de datos es 9, el valor del medio de estos datos es la mediana, puesto que deja cuatro valores por debajo y cuatro valores por encima. Este valor es **20**.

b. Hallar la media del siguiente conjunto de datos ordenados:

14 15 18 19 **20** **23** 26 28 30 32

Observe que son 10 datos, un número par de datos. En este caso se toman los dos valores del medio y se promedian:

$$Me = \frac{20 + 23}{2} = 21.5$$

Cuando los datos se encuentran agrupados, se calcula el valor de $\frac{n}{2}$ y con él se busca, en las frecuencias acumuladas, el intervalo de clase en donde este se encuentra o se aproxime mejor. Esta clase recibe el nombre de ***clase de la mediana***. Identificada la clase de la mediana, se considera que los valores en esa clase se distribuyen uniformemente de modo que se pueda calcular la mediana por el método de la interpolación lineal. En el siguiente ejemplo se describe paso a paso el cálculo de esta medida de tendencia central.

EJEMPLO 1.6.

Tomando la tabla 1.1 de distribución de frecuencias agrupadas del ejemplo 1.4. de esta unidad didáctica, calcular la mediana del conjunto de datos.

Primero se identifica la clase de la mediana (la clase que contiene a la mediana).

$$\frac{n}{2} = \frac{25}{2} = 12.5$$

La clase de la mediana es (26-30), pues el número de frecuencias acumuladas es el valor más cercano a 12.5.

Tabla 1.2.
Distribución de frecuencias agrupadas

Intervalo	Frecuencia f	Frecuencia acumulada
16-20	4	4
21-25	6	**10**
26-30	7	17
31-35	5	22
36-40	3	25
Total	**25**	

Clase de la mediana →

Hay 10 observaciones por debajo del límite inferior de la clase de la mediana.

$$12.5 - 10 = 2.5$$

El valor de 2.5 se interpola en el ancho o amplitud de la clase de la mediana que es 4.

Frecuencia absoluta		Ancho de clase
7	→	4
2.5	→	**X**

$$X = \frac{2.5 \times 4}{7} = 1.4$$

Así pues, la mediana se encontrará 1.4 unidades más del límite inferior de la clase de la mediana:

$$Me = 26 + 1.4 = 27.4$$

En muchas referencias bibliográficas se expone una ecuación para el cálculo de la mediana cuando los datos se encuentran agrupados. Ella se deriva del análisis hecho en el ejemplo anterior y se describe de la siguiente manera:

$$Me = \frac{\frac{n}{2} - F_{k-1}}{f_k} \times A_k + L_k$$

Donde:
n es el tamaño de la muestra o la suma de todas las frecuencias.
F_{k-1} es la frecuencia absoluta acumulada de la clase anterior de la clase de la mediana.
f_k es la frecuencia absoluta de la clase de la mediana.
A_k es la amplitud de la clase de la mediana.
L_k es el límite real inferior de la clase de la mediana.

EJEMPLO 1.7.

Determine la mediana de la distribución de frecuencias agrupadas del ejemplo 1.6., haciendo uso de la ecuación para su cálculo.

Primero, se identifica cada valor:
n = 25
F_{k-1} = 10
f_k = 7
A_k = 4
L_k = 26

$$Me = \frac{\frac{n}{2} - F_{k-1}}{f_k} \times A_k + L_k \qquad \Rightarrow \qquad Me = \frac{\frac{25}{2} - 10}{7} \times 4 + 26 = 1.4 + 26 = 27.4$$

Otra manera para hallar la mediana de un conjunto de datos agrupados es el método gráfico. Ya se vio algo cuando se estudiaba la ojiva: al graficar en un mismo eje coordenado la ojiva ascendente y descendente, el punto donde estas dos curvas se encuentren corresponde a la mediana de los datos agrupados, leyendo el valor en el eje horizontal.

Si se trabaja en cambio con la ojiva porcentual, es decir con la distribución de frecuencias relativas, la mediana será el valor de la abscisa cuya ordenada es el 50%.

Se concluye entonces que la mediana no está afectada por los valores

extremos del conjunto de datos, sean estos grandes o pequeños. No influencian en lo absoluto como sí lo hacen en el cálculo de la media. Cuando la distribución de los datos es muy simétrica, no hay casi diferencia entre la media y la mediana. El cálculo de la mediana es simple, pero siempre requiere que los datos se encuentren ordenados, condición que no requiere el cálculo de la media. Finalmente, se podría decir que la mediana no es una medida muy confiable para describir el conjunto de datos, pues en su cálculo sólo intervienen los valores más centrales sin tener en cuenta los demás y su comportamiento general.

1.1.3. Moda

Se trata del valor más frecuente en un conjunto de datos. Se considera como el valor más representativo o típico de una serie de valores. Es simbolizada como ***Mo***. Si dos valores tienen la misma frecuencia se dice que el conjunto es ***bimodal***. Cuando más de dos valores ocurren con la misma frecuencia y ésta es la más alta, todos los valores son modas, por lo que el conjunto de datos recibe el nombre de ***multimodal***.

Cuando los datos se encuentran agrupados la moda es la marca de clase del intervalo de clase que contiene la mayor frecuencia.

La moda también puede determinarse gráficamente, usando un histograma de frecuencias o un polígono de frecuencias. La barra más alta o el pico más alto corresponde al valor que más se repite. Generalmente las curvas de frecuencia presentan un solo pico, pero a veces se encuentran series con dos o más picos, es decir puntos que corresponden a una mayor densidad de frecuencias. Esto sucede cuando se trabaja con grupos de datos heterogéneos.

EJEMPLO 1.8.

Las siguientes tablas de frecuencias indican el número de personas de acuerdo a su edad que asistieron al estreno de una película.

En la tabla 1.3., donde los datos están sin agrupar, la moda es **22**, valor correspondiente a la mayor frecuencia que es 5.

En la tabla 1.4., los datos se encuentran agrupados, la moda se encuentra en el intervalo de clase 19.5 – 22.5 y corresponde a la marca de clase que es **21**.

Obsérvese que aunque sean el mismo conjunto de datos, la moda varía dependiendo de su tratamiento, es decir, de cómo estos se agrupan. En este caso, debe considerarse el valor obtenido con la tabla de frecuencias de los datos sin

agrupar.

Tabla 1.3.
Distribución de frecuencias
de la asistencia a cine

X	f	X	f
14	1	23	4
15	0	24	3
16	1	25	2
17	2	26	4
18	3	27	3
19	4	28	2
20	4	29	0
21	4	30	0
22	5	31	1
Total			**43**

Tabla 1.4.
Distribución de frecuencias agrupadas
de la asistencia a cine

Intervalos de clase	Marca de clase	Frec.
13.5 – 16.5	15	2
16.5 – 19.5	18	9
19.5 – 22.5	**21**	13
22.5 – 25.5	24	9
25.5 – 28.5	27	9
28.5 – 31.5	30	1
Total		**43**

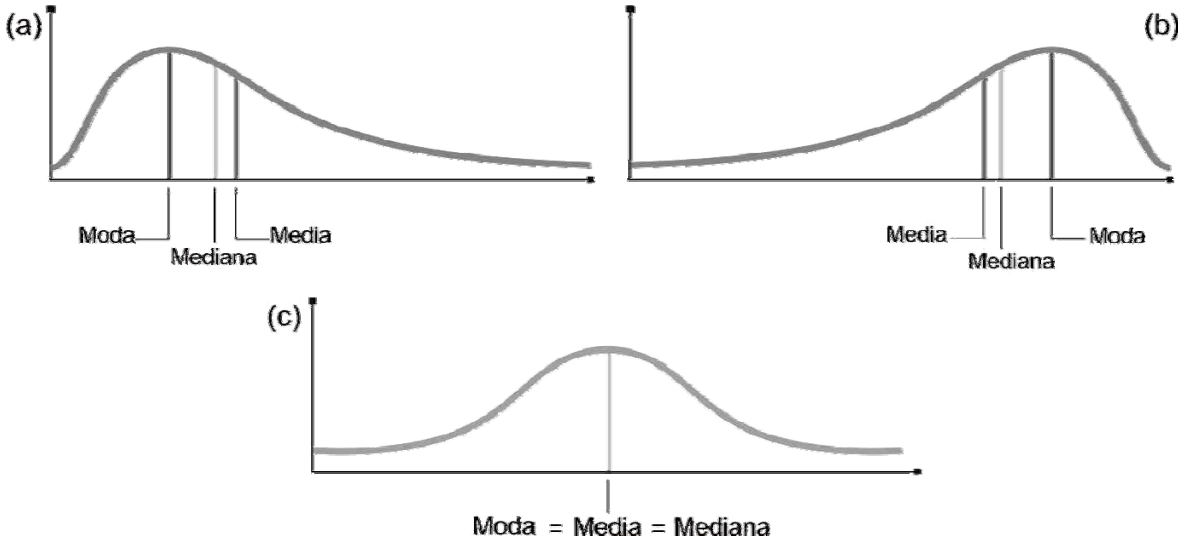
La moda no es tan usada como la media o la mediana. Para encontrarla se requiere que los datos estén ordenados. Su cálculo es poco preciso debido a que no se puede expresar en términos algebraicos.

Se han visto hasta ahora tres medidas de tendencia central: media, mediana y moda. Determinar cuál de ellas usar en un tratamiento estadístico depende mucho de la información que se tenga y del objetivo que se persigue. La media, a diferencia de la mediana y la moda, presenta una ligera estabilidad en el muestreo, es por eso que su uso es más frecuente. Si la distribución es casi simétrica, cualquiera de ellas puede usarse y resultarán aproximadamente iguales. Cuando los datos no están ordenados, puede resultar más fácil calcular la media aritmética que la mediana. Cuando los datos no están agrupados, el cálculo de la moda se hace más preciso. Si la distribución no es simétrica, es más recomendable emplear la mediana o la moda como medidas de posición.

En cualquier distribución el valor de la mediana se localiza entre la media y la moda. Cuando la distribución es asimétrica a la derecha se cumple que $Mo < Me < \bar{x}$; si en cambio es asimétrica a la izquierda $\bar{x} < Me < Mo$. Se dice entonces, que una distribución está ***sesgada*** si no es simétrica y si se extiende más hacia un lado que hacia el otro. Y será ***simétrica*** cuando la mitad de su histograma es aproximadamente igual a su otra mitad. Los datos sesgados a la izquierda (***sesgo negativo***) presentan una cola izquierda más larga y su media y mediana se encuentran a la izquierda de la moda. Mientras que los datos

sesgados a la derecha (***sesgo positivo***) poseen una cola derecha más larga y su mediana y media están a la derecha de la moda (ver figura 1.1.)

Figura 1.1.
Distribuciones sesgadas
(a) Sesgada a la derecha; (b) Sesgada a la izquierda; (c) Simétrica



La ***relación de Pearson*** afirma que la distancia entre la media y la moda es tres veces la distancia entre la media y la mediana. Esta relación es utilizada para calcular cualquiera de ellas, conociendo las otras dos medidas.

$$\bar{x} - Mo = 3(\bar{x} - Me) \quad \Rightarrow \quad Mo = 3Me - 2\bar{x}$$

En resumen, se puede entender la *media aritmética* como el punto de equilibrio del conjunto de datos (como el centro de gravedad de un cuerpo); la *mediana* como la medida que permite dividir el área bajo la curva de distribución en dos parte iguales y la *moda* como el pico más alto de la curva de distribución.

El cuadro siguiente[3] resume y compara de una manera didáctica y práctica la media, mediana y moda en términos de ventajas y desventajas para su cálculo y uso en la investigación estadística. Ellas tres son las medidas de tendencia central más comúnmente usadas, en el tema siguiente se estudiarán otras medidas no menos importantes pero si menos usadas en el tratamiento estadístico.

[3] Modificado de *Probabilidad y estadística*, Mario F. Triola. Novena edición. Pearson & Addison Wesley. México. 2004.

Tabla 1.5.
Comparación de la media, mediana y moda

Medida de tendencia central	¿Qué tan común es?	¿Existe siempre?	¿Toma en cuenta cada valor?	¿Se ve afectada por los valores extremos?	¿Requiere que los datos estén ordenados?	Ventajas y desventajas
Media	Es la más común	Si	Si	Si	No	Presenta una ligera estabilidad frente al muestreo.
Mediana	De uso común	Si	No	No	Si	No es muy confiable para describir el conjunto de datos, pues en su cálculo sólo intervienen los datos más centrales.
Moda	Usada en ocasiones	Podría no existir o haber más de una	No	No	Si	Es más precisa cuando los datos no están agrupados.

1.1.4. Otras medidas de tendencia central

La ***media geométrica*** se utiliza para promediar crecimientos geométricos de la variable, o cuando se quiere dar importancia a valores pequeños, o cuando se quiere determinar el valor medio para un conjunto de porcentajes. Suele utilizarse en negocios y economía para calcular las tasas de cambio promedio, las tasas de crecimiento promedio o tasas promedio. Se simboliza ***Mg*** y se define como la raíz n-ésima de la productoria de los n valores de la variable.

Cuando los datos no son agrupados, la media geométrica se calcula hallando el producto de todos los elementos y extrayendo la raíz del orden del número de observaciones.

$$Mg = \sqrt[n]{\prod_{i=1}^{n} X_i} = \sqrt[n]{X_1 \cdot X_2 \cdot ... \cdot X_n}$$

Cuando los datos están agrupados, la media geométrica se define como la raíz n-ésima de la productoria de los valores de la variable (marca de clase) elevadas cada una de ellas a su correspondiente frecuencia absoluta.

$$Mg = \sqrt[n]{\prod_{i=1}^{n} X_i^{n_i}} = \sqrt[n]{X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_n^{n_i}}$$

EJEMPLO 1.9.

a. Hallar la media geométrica de 2, 4, 6, 9, 12, 15

$$Mg = \sqrt[6]{\prod_{i=1}^{6} X_i} = \sqrt[6]{2 \cdot 4 \cdot 6 \cdot 9 \cdot 12 \cdot 15} = \sqrt[6]{77.760} = 6.53$$

b. Hallar la media geométrica de la siguiente distribución de frecuencias agrupadas.

Tabla 1.6.
Distribución de frecuencias agrupadas

Intervalos de clase	Marcas de clase	Frecuencias
0.5 – 1.5	1	2
1.5 – 2.5	2	5
2.5 – 3.5	3	8
3.5 – 4.5	4	5
Total		**20**

$$Mg = \sqrt[4]{\prod_{i=1}^{4} X_i^{n_i}} = \sqrt[4]{1^2 \cdot 2^5 \cdot 3^8 \cdot 4^5} = 121.1$$

La ***media armónica*** de un conjunto de datos es el recíproco de la media aritmética de los recíprocos de los números de la serie de datos. Se simboliza ***Mh*** y se define como:

$$\frac{1}{Mh} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{Xn}}{n} = \frac{\sum \frac{1}{x}}{n} \qquad \Rightarrow \qquad Mh = \frac{n}{\sum \frac{1}{x}}$$

La media armónica es muy influenciable por los valores extremos de la

serie, especialmente los más pequeños. Se utiliza preferiblemente para conjuntos de datos que consisten en tasas de cambios, como la velocidad.

EJEMPLO 1.10.

Un obrero se gasta 50 minutos en terminar un producto y otro lo hace en 40 minutos. ¿Cuál es el tiempo medio requerido para terminar dicho producto?

$$\frac{1}{Mh} = \frac{\frac{1}{50} + \frac{1}{40}}{2} = \frac{0.045}{2} = 0.0225 \Rightarrow Mh = \frac{1}{0.0225} = 44.44$$

44.44 minutos es el tiempo medio requerido.

Los ***cuartiles, deciles*** y ***percentiles*** son medidas que se utilizan para determinar los intervalos dentro de los cuales quedan proporcionalmente repartidos los términos de la distribución.

Para calcular los ***cuartiles*** se divide la distribución en cuatro partes iguales, de manera que cada una tendrá el 25% de las observaciones. Los tres puntos de separación de los valores son los cuartiles. El cuartil inferior (Q_1) es aquel valor de la variable que representa el 25% de las observaciones y a la vez, es superado por el 75% restante. El segundo cuartil (Q_2) corresponderá a la mediana de la distribución. El tercer cuartil (Q_3) es aquel valor que representa el 75% y es superado por el 25% restante de las observaciones.

Para calcular estos tres promedios se procede de manera semejante al cálculo de la media aritmética.

EJEMPLO 1.11.

Hallar los cuartiles de la distribución de frecuencias de la tabla 1.2., del ejemplo 1.6.

Primero se identifica la clase en donde se encuentra el primer cuartil.

$$\frac{n}{4} = \frac{25}{4} = 6.25$$

El intervalo de clase donde se encuentra el primer cuartil es (21-25), pues el número de frecuencias acumuladas es el valor más cercano a 6.25.

	Intervalo	Frecuencia f	Frecuencia acumulada
	16-20	4	4
Clase del Q_1	21-25	6	10
	26-30	7	17
Clase del Q_3	31-35	5	22
	36-40	3	25
	Total	**25**	

Hay 4 observaciones por debajo del límite inferior de la clase del primer cuartil.

$$6.25 - 4 = 2.25$$

El valor de 2.25 se interpola en la amplitud de la clase del primer cuartil que es 4.

Frecuencia absoluta		Ancho de clase
6	⟶	4
2.25	⟶	**X**

$$X = \frac{2.25 \times 4}{6} = 1.5$$

Así pues, el primer cuartil se encontrará 1.5 unidades más del límite inferior de la clase correspondiente:

$$Q_1 = 21 + 1.5 = 22.5$$

El segundo cuartil corresponde al punto medio de la distribución, esto es la mediana del grupo de datos.

$$Q_2 = Me = 27.4$$

Para el tercer cuartil se procede de la misma manera.

$$\frac{3n}{4} = \frac{3 \times 25}{4} = 18.75$$

El intervalo de clase donde se encuentra el tercer cuartil es (31-35) y hay 17 observaciones por debajo del límite inferior de la clase de este cuartil.

$$18.75 - 17 = 1.75$$

Frecuencia absoluta		Ancho de clase
5	⟶	4
1.75	⟶	**X**

$$X = \frac{1.75 \times 4}{5} = 1.4$$

El tercer cuartil se encontrará 1.4 unidades más del límite inferior de su clase:

$$Q_3 = 31 + 1.4 = 32.4$$

Lo que quiere decir que el 25% de los valores está por debajo de 22.5; el 50% está por debajo de 27.4 y el 75% está por debajo de 32.4.

Para calcular los ***deciles*** se divide el conjunto de datos en 10 partes iguales, de manera que se obtienen nueve valores que dividen la frecuencia total en diez partes iguales. El primer decil (D_1) es igual al valor que supera al 10% de las observaciones y es superado por el 90% restante y así para cada uno de los deciles. Su cálculo es muy semejante al de los cuartiles.

De igual manera se puede calcular el ***centil*** o ***percentil*** al dividir en cien partes iguales la distribución. El primer percentil (P_1) es igual al valor que supera al 1% de las observaciones y es superado por el 99% restante y así sucesivamente. Obsérvese que $D_1=P_{10}$; $D_2=P_{20;}$...

El método más sencillo para identificar tanto cuartiles, deciles y percentiles es el gráfico, haciendo uso de la ojiva porcentual ascendente. Sólo requiere buscar en el eje vertical el porcentaje que se busca y leer en el eje horizontal su correspondiente valor.

EJEMPLO 1.12.

A partir de la ojiva porcentual de la distribución de frecuencias agrupadas de la tabla 1.2., determine el valor de: $Q_1, Q_2, Q_3, D_1, D_5, D_9, P_5, P_{95}$.

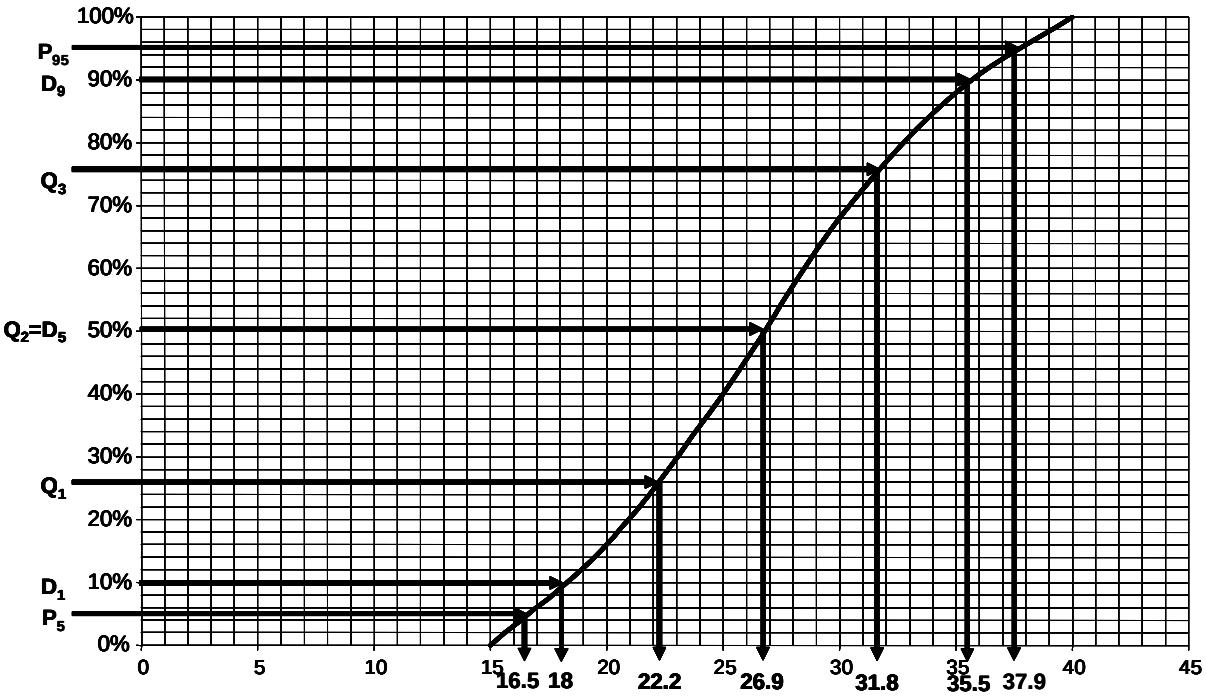
Para construir la ojiva, se debe completar la tabla de distribución de frecuencias

agrupadas.

Tabla 1.7.
Distribución de frecuencias agrupadas

Intervalo	Frecuencia absoluta	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
16-20	4	4	16%	16%
21-25	6	10	24%	40%
26-30	7	17	28%	68%
31-35	5	22	20%	88%
36-40	3	25	12%	100%
Total	**25**		**100%**	

Figura 1.2.
Ojiva porcentual ascendente



Con las frecuencias relativas acumuladas se construye la ojiva porcentual ascendente. Una vez construida, se inicia el proceso de identificar cada valor pedido, teniendo en cuenta qué porcentaje representa. Es decir, el primer cuartil representa el 25%, el segundo 50%, el tercero 75%, el primer decil representa el 10%, del quinto es el 50% y el noveno corresponderá al 90%, mientras que el

percentil 5 representa al 5% y el 95 al 95%.

Observe en la figura 1.2. que los valores teóricos (calculados en ejemplos anteriores) no son completamente coincidentes. Esto demuestra que el método gráfico no es el más apropiado para su determinación, sin embargo es muy útil y sus valores se aproximan al teórico entre mejor esté graficada la ojiva.

Tabla 1.8.
Resumen de cálculos, ejemplo 1.12.

Medida	Porcentaje que representa	Valor teórico calculado	Valor gráfico obtenido
Q_1	25%	22.5	22.2
Q_2	50%	27.4	26.9
Q_3	75%	32.4	31.8
D_1	10%		18
D_5	50%	27.4	26.9
D_9	90%		35.5
P_5	5%		16.5
P_{95}	95%		37.9

Ahora intente lo siguiente: determine los valores teóricos de las medidas que aún no ha calculado y compárelas con las obtenidas por el método gráfico. ¿Son muy diferentes?

El cálculo de percentiles para datos no agrupados se hace más sencillo. Para ello se requiere que los datos se encuentren ordenados de manera ascendente. Luego se determina el valor de la expresión:

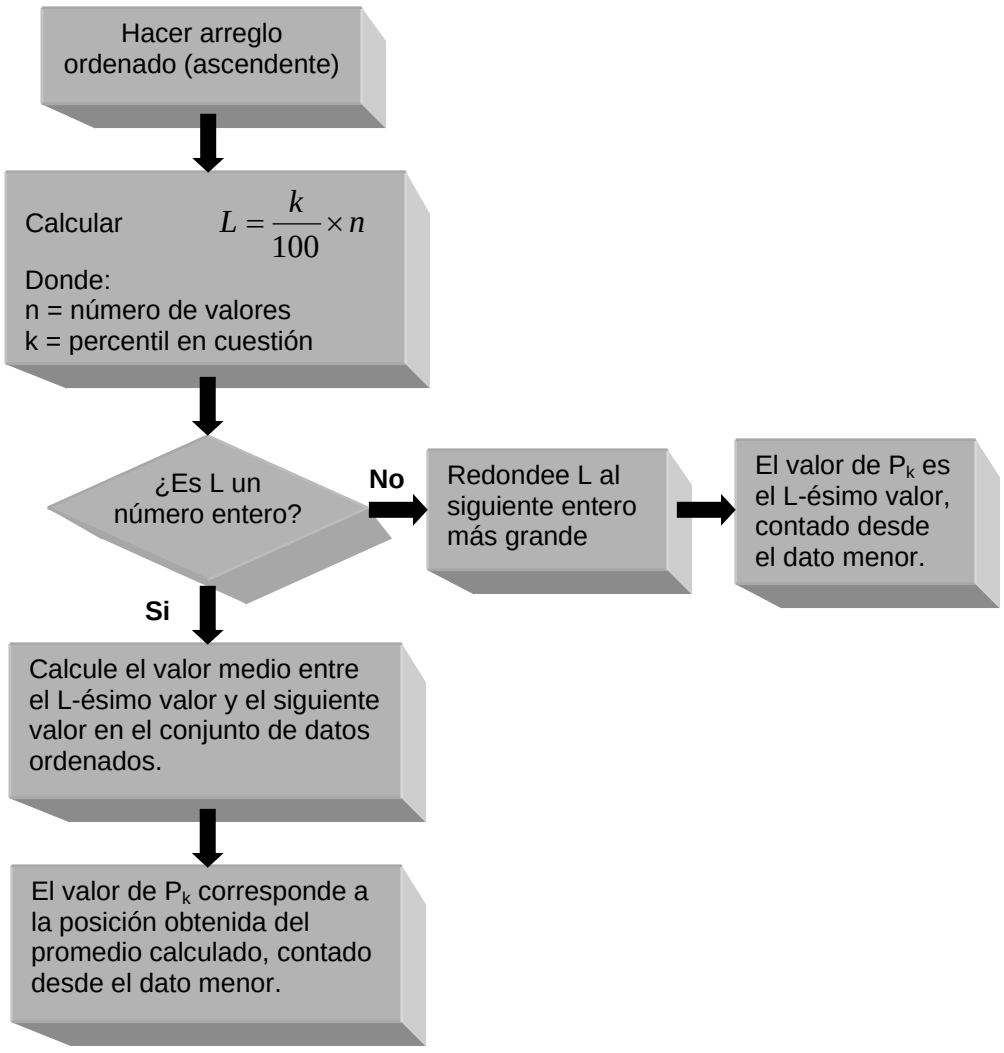
$$L = \frac{k}{100} \times n$$

Donde:
n es el número de valores del grupo de datos
k es el percentil en cuestión

Si el valor de L es un número entero, el valor del k-ésimo percentil estará por el valor medio entre el L-ésimo valor y el siguiente valor. Si en cambio, el valor de L no es un número entero, este valor debe ser redondeado al siguiente entero más grande y el valor de Pk corresponderá a la posición L-ésima. El siguiente diagrama

de flujo[4] clarifica el procedimiento para el cálculo del k-ésimo percentil.

Figura 1.3.
Diagrama de flujo para el cálculo del k-ésimo percentil



EJEMPLO 1.13.

Tome el arreglo ordenado del ejemplo 2.1., de la Unidad Didáctica Uno sobre la evaluación de los latidos cardíacos de un grupo de 30 personas después de cierta actividad física y calcule los siguientes percentiles.

[4] Modificado de *Probabilidad y estadística*, Mario F. Triola. Novena edición. Pearson & Addison Wesley. México. 2004.

58	70	80	85	88	94
60	74	82	85	91	95
62	**75**	**82**	85	91	95
64	76	82	87	92	95
68	76	84	88	**92**	110

a. El valor del percentil 10, $\boldsymbol{P_{10}}$

Para esto, se sigue el procedimiento planteado en el diagrama de flujo de la figura 1.3. Los datos se encuentran ordenados de forma ascendente; se procede entonces a calcular L, es decir el localizador que da la posición del valor 10.

$$L = \frac{10}{100} \times 30 = 3$$

Después, se verifica si el valor de L es un entero o no. En este caso, L es entero así que se sigue hacia abajo en el diagrama de flujo. De manera que el décimo percentil está a la mitad entre el valor L-ésimo (tercero) y el siguiente valor (cuarto). Es decir, el valor del percentil 10 se ubica entre 62 y 64, que corresponden al tercer y cuarto valor del grupo de datos ordenados, respectivamente. Se tiene entonces:

$$P_{10} = \frac{62 + 64}{2} = 63$$

b. El valor del percentil 43, $\boldsymbol{P_{43}}$

Se calcula el valor de L:

$$L = \frac{43}{100} \times 30 = 12.9 \approx 13$$

Como el valor de L no es entero, se redondea al siguiente entero más grande. El valor del percentil 43 es el valor 13º del grupo de datos ordenados contado desde el dato menor. Así:

$$P_{43} = 82$$

c. El valor del percentil 81, $\boldsymbol{P_{81}}$

Se calcula el valor de L:

$$L = \frac{81}{100} \times 30 = 24.3 \approx 25$$

Como el valor de L no es entero, se redondea al siguiente entero más grande. Observe que no se redondea al entero más cercano sino al entero mayor. El valor del percentil 81 es el valor 25º del grupo de datos ordenados contado desde el dato menor. Así:

$$P_{81} = 92$$

d. El valor del cuartil 1, $\boldsymbol{Q_1}$

Recuerde que $\boldsymbol{Q_1}$ es igual que $\boldsymbol{P_{25}}$, por lo que se procede a calcular el valor del percentil 25. Se halla el valor de L:

$$L = \frac{25}{100} \times 30 = 7.5 \approx 8$$

El valor del percentil 25 es el valor 8º del grupo de datos ordenados contado desde el dato menor. Así:

$$Q_1 = P_{25} = 75$$

EJERCICIOS TEMA 1.1.

1. En la siguiente serie de números indicar:

$4.000	$4.500	$5.000	$5.000	$8.250
$9.300	$9.700	$12.000	$12.500	$35.000

a. La media
b. La mediana
c. La moda
d. ¿Cuál de las medidas es más representativa? ¿Por qué?
e. ¿Qué valor de esta serie afecta a la media aritmética?

2. Calcule la media aritmética, mediana y moda de los siguientes conjuntos de datos:

a. 6, 5, 7, 6, 5, 4, 7, 4, 6, 8, 7, 6

b.

X	***f***
4	8
5	12
6	11
7	20
8	14
9	10
10	7
n	**82**

c.

Intervalos de clase	Frecuencia
39 – 49	5
49 – 59	8
59 – 69	10
69 – 79	9
79 – 89	8
89 – 99	6
99 - 109	4
Total	**50**

3. De un grupo de 100 obreros en una fábrica, 40 trabajan en el día y 60 en la noche. Se sabe que el salario promedio de los 100 obreros es $407.200 y que los del turno del día reciben en promedio $28.000 menos que los trabajadores nocturnos. ¿Cuál es el salario promedio en cada grupo?

4. Carlos obtiene calificaciones parciales de 65, 83, 80, y 90. En el examen final recibe una calificación de 92. Calcule la media ponderada, si cada uno de los exámenes parciales cuenta el 15% y el examen final cuenta 40% de la calificación total.

5. Antes del examen final de Estadística, un estudiante obtiene calificaciones de 3.5 en el 20%, 2.0 en el 30%, 4.2 en el 10%. Si la evaluación final equivale al 40% restante, ¿que calificación necesita para obtener un promedio final de 3.5?

6. En una industria se ha controlado el tiempo que tardan tres obreros en ensamblar un motor. Uno demora 6 horas, otro 8 horas y un tercero demora 5 horas. Halle el rendimiento de un obrero tipo, que sirva de base para análisis financieros.

7. Un hombre viaja desde Bogotá hasta Acacías a una velocidad de 60 km/h. Para evitar la noche en carretera, este decide acelerar a 80 km/h para llegar de nuevo a Bogotá. ¿Cuál es la velocidad promedio del viaje completo?

8. El factor de crecimiento promedio de dinero compuesto con tasa de interés anual del 10%, el 8%, el 9%, el 12% y el 7% se obtiene determinando la media geométrica de 1.10, 1.08, 1.09, 1.12 y 1.07. Calcule el factor de crecimiento promedio.

9. Para la siguiente tabla de distribución de frecuencias agrupadas, determine los tres cuartiles tanto teórica como gráficamente.

Intervalos de clase	Frecuencia
39 – 49	5
49 – 59	8
59 – 69	10
69 – 79	9
79 – 89	8
89 – 99	6
99 - 109	4
Total	**50**

10. Tome la combinación ordenada de los datos que corresponden al perímetro craneal de un niño al nacer, del numeral 3 de los ejercicios del tema 2.3., y calcule los siguientes percentiles:

a. 5 **b.** 15 **c.** 95 **d.** 25
e. 50 **f.** 10 **g.** 75 **g.** 30

1.2. MEDIDAS DE DISPERSIÓN

Se veía en el tema anterior la tendencia que tiene un conjunto de datos dado a agruparse hacia el centro, pero también se descubrió que los datos extremos podían estar bastante alejados de esa tendencia central. Medir esa variación respecto a los promedios es un cálculo importante en el tratamiento estadístico de datos, medidas a las que se les denomina de dispersión o de variación.

La información que arrojan las medidas de tendencia central no siempre proporcionan conclusiones contundentes frente al conjunto de datos. Por ejemplo, a un profesor de Estadística poco le dice la media aritmética al afirmar que el promedio de los estudiantes tiene el curso en 3.0 ya que no le termina de aclarar si el grupo completo está muy cerca de esa nota, sea por encima o por debajo de ella, o si al contrario existe tanta variabilidad en las notas de los estudiantes que puede ir desde 1.0 hasta 5.0. Se estudiará a continuación cómo resolver este tipo de problemas y qué medidas de dispersión usar.

1.2.1. Rango o recorrido

Sobre esta medida ya se había trabajado en la construcción de las tablas de frecuencia agrupada. Se trata de la diferencia entre el límite superior y el límite inferior de un conjunto de datos. Es la medida de dispersión más fácil de calcular, sólo requiere que los datos estén ordenados. Pero es poco usada como medida de dispersión porque se deja afectar fácilmente de los valores extremos de poca frecuencia.

EJEMPLO 1.14.

Un profesor de Estadística tiene a su cargo dos grupos de 40 estudiantes cada uno. La siguiente tabla de frecuencias reporta las calificaciones del grupo A y grupo B de estudiantes, después de la primera evaluación. ¿Hay diferencia alguna entre estos dos grupos?

Lo primero que se hace para verificar diferencias entre ambos grupos es calcular su media aritmética.

$$\bar{x}_A = \frac{\sum fX}{n} = \frac{174.4}{40} = 4.36 \qquad \bar{x}_B = \frac{\sum fX}{n} = \frac{174.4}{40} = 4.36$$

$$RangoA = 5.0 - 4.0 = 1.0 \qquad RangoB = 5.0 - 4.0 = 1.0$$

Tabla 1.9.
Distribución de frecuencias
de las calificaciones de estudiantes de Estadística

~~Calificación~~	Frecuencia	
	A	**B**
4.0	1	2
4.1	2	9
4.2	3	7
4.3	16	4
4.4	10	5
4.5	4	4
4.6	3	3
4.7	0	2
4.8	0	1
4.9	0	1
5.0	1	2
Total	**40**	**40**

Tanto la media como el rango de ambos conjuntos de datos son iguales. Sin embargo, ellos se distribuyen de forma muy diferente. Observe que el grupo A es más compacto hacia las notas entre 4.5 y 4.0. La nota de 5.0 de un solo estudiante interfiere muchísimo en el análisis verdadero del comportamiento académico de los estudiantes del grupo A.

Analice qué tanto cambian los valores de la media y el rango del grupo A de estudiantes si se elimina la nota de 5.0, observe que un dato extremo hace variar completamente el conjunto de datos y demuestra que, comparado con otro, el cálculo de la media y el rango son insuficientes para arrojar análisis certero de comparación.

$$\bar{x}_A = \frac{\sum fX}{n} = \frac{169.4}{39} = 4.34 \qquad RangoA = 4.6 - 4.0 = 0.6$$

En cambio, las calificaciones del grupo B se distribuyen mejor alrededor de todo el rango de datos.

Para eliminar la influencia de los extremos en el cálculo del rango, es común hacer uso del ***rango intercuartílico*** que consiste en determinar la diferencia entre el tercer cuartil y el primero.

$$Q_D = Q_3 - Q_1$$

El ***rango semiintercuartílico*** o ***desviación cuartil*** se obtiene calculando el rango intercuartílico y dividiendo este entre dos.

$$Q_{D2} = \frac{Q_3 - Q_1}{2}$$

Ambas medidas son más confiables como variabilidad comparadas con el rango, sin embargo presentan inconvenientes para su uso puesto que no consideran todos los valores de la distribución y puede ocurrir que los valores inferiores a Q_1 o superiores a Q_3 estén o muy compactos o muy dispersos sin que esto afecte a Q_D y no sea reflejado en su resultado.

De la misma manera, el ***rango interdecil*** corresponde a la diferencia entre el noveno y el primer decil:

$$D_R = D_9 - D_1$$

1.2.2. Varianza

Es una de las medidas más usadas en estadística, ella a su vez da origen a otra mucho más significativa: la desviación típica o estándar. Se define como la media aritmética de los cuadrados de las desviaciones respecto a la media aritmética. Se simboliza s^2 para la varianza muestral y σ^2 para la varianza poblacional.

Para datos no agrupados:

$$s^2 = \frac{\sum (X - \bar{x})^2}{n} \qquad \Rightarrow \qquad s^2 = \frac{\sum X^2}{n} - \bar{x}^2$$

Para datos agrupados:

$$s^2 = \frac{\sum f (X - \bar{x})^2}{n} \qquad \Rightarrow \qquad s^2 = \frac{\sum f \cdot X^2}{n} - \bar{x}^2$$

La varianza indica la desviación de los datos respecto a la media. Para comparar dos distribuciones, en cuanto a su variabilidad absoluta, se pueden utilizar sus varianzas de manera que el resultado indique cuál de ellas es más homogénea o cuál es más heterogénea.

EJEMPLO 1.15.

Se quiere conocer la verdadera calidad de producción en dos empresas fabricantes de tornillos para fuselaje. La siguiente tabla indica las longitudes de una muestra de tres tornillos tomados al azar. Haga un análisis de variabilidad de ambas empresas.

Empresa A	1,95 pulg.	2,03 pulg.	2,02 pulg.
Empresa B	1,70 pulg.	1,80 pulg.	2,50 pulg.

Es fácil calcular que ambas empresas tienen una media de $\bar{x} = 2{,}0$ pulgadas. Pero las muestras difieren mucho en sus tamaños, para visualizar mejor esto se analizan sus respectivas varianzas. Tenga en cuenta que los datos no están agrupados, por lo que se hace uso de la primera ecuación:

$$s^2{}_A = \frac{\sum X^2}{n} - \bar{x}^2 = \frac{1{,}95^2 + 2{,}03^2 + 2{,}02^2}{3} - 2{,}0^2 = 0{,}001$$

$$s^2{}_B = \frac{\sum X^2}{n} - \bar{x}^2 = \frac{1{,}70^2 + 1{,}80^2 + 2{,}50^2}{3} - 2{,}0^2 = 0{,}127$$

Observe que la empresa A tiene una variación mayor respecto a la empresa B en cuanto a la calidad en la fabricación de tornillos. Esto quiere decir que la empresa B varía mucho, en su producción, el tamaño de sus tornillos mientras que la empresa A mantiene un rango constante en el tamaño de los tornillos que produce.

Las unidades de la varianza son los cuadrados de las unidades de los datos: pesos cuadrados, alumnos cuadrados, etc., medidas difíciles de interpretar. De allí que la varianza de origen a la desviación típica o estándar.

1.2.3. Desviación típica o estándar

Esta medida se obtiene extrayendo la raíz cuadrada de la varianza, tomando siempre el valor positivo. Se simboliza por s en la muestra y σ en la población. Esta es la medida de dispersión más conocida y más utilizada en el análisis de datos estadísticos.

Para datos no agrupados:

$$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n}} \qquad \Rightarrow \qquad s = \sqrt{\frac{\sum X^2}{n} - \bar{x}^2}$$

Para datos agrupados:

$$s = \sqrt{\frac{\sum f (X - \bar{x})^2}{n}} \qquad \Rightarrow \qquad s = \sqrt{\frac{\sum f \cdot X^2}{n} - \bar{x}^2}$$

EJEMPLO 1.16.

Después de estudiar los conceptos de varianza y desviación estándar, se está en capacidad de hacer un análisis mucho más riguroso de la variabilidad de las calificaciones de los estudiantes de Estadística del ejemplo 1.14.

Tabla 1.10.
Distribución de frecuencias
de las calificaciones de estudiantes de Estadística

Calificación	Frecuencia		X^2	$f \cdot X^2$	
	A	B		A	B
4.0	1	2	16	16	32
4.1	2	9	16,81	33,62	151,29
4.2	3	7	17,64	52,92	123,48
4.3	16	4	18,49	295,84	73,96
4.4	10	5	19,36	193,6	96,8
4.5	4	4	20,25	81	81
4.6	3	3	21,16	63,48	63,48
4.7	0	2	22,09	0	44,18
4.8	0	1	23,04	0	23,04
4.9	0	1	24,01	0	24,01
5.0	1	2	25	25	50
Total	**40**	**40**	**223,85**	**761,46**	**763,24**

Para el grupo A se tiene:

$$s_A = \sqrt{\frac{\sum f \cdot X^2}{n} - \bar{x}^2} = \sqrt{\frac{761.46}{40} - 4.36^2} = \sqrt{0.0269} = 0.164$$

Y para el grupo B de estudiantes, se tiene:

$$s_B = \sqrt{\frac{\sum f \cdot X^2}{n} - \bar{x}^2} = \sqrt{\frac{763.24}{40} - 4.36^2} = \sqrt{0.0714} = 0.267$$

La varianza del grupo B es mayor que la del grupo A, se dice entonces que los datos del grupo B tiene mayor variabilidad que los del grupo A; en otras palabras, en el grupo B hubo mayor estabilidad en las notas alrededor de su media: 4.36.

Es importante tener en cuenta las siguientes propiedades de la desviación estándar:

- La desviación estándar es una medida de variación de todos los valores con respecto a la media.
- El valor de la desviación estándar siempre es positivo y sólo es igual a cero cuando los valores de los datos son iguales.
- Si el valor de la desviación estándar es muy grande, este indica mayor variación en el grupo de datos.
- El valor de la desviación estándar puede incrementarse drásticamente cuando se incluye uno o más datos distantes.
- Las unidades de la desviación estándar son las mismas de los datos originales (pulgadas, centímetros, etc.)

1.2.4. Coeficiente de variación

Las medidas de dispersión que se han estudiado son medidas absolutas y se expresan en las mismas unidades con las que se mide la variable. Cuando se comparan dos o más conjuntos de datos con unidades de medida de observación diferentes, no es posible compararlas con estas medidas absolutas. Si las unidades de observación de los conjuntos de datos son iguales, estos pueden compararse usando cualquiera de estos estadísticos (como en el ejemplo anterior) pero siempre y cuando la media aritmética sea la misma, de lo contrario estas apreciaciones no aportarán una buena conclusión sobre las series que se comparan.

Para efectuar comparaciones entre series de observaciones distintas, en estadística se usa el ***coeficiente de variación*** y así se puede determinar cuál serie tiene mayor o menor variabilidad relativa.

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Cuando el coeficiente de variación es muy alto se dice que la media

aritmética no es lo suficientemente representativa en la distribución.

1.2.5. Desviación media

Se define como la media aritmética de las desviaciones respecto a la media, tomadas en valor absoluto[5]. Es una de las medidas más fáciles de calcular y por ello, muy usada. Ella toma todos los valores de la variable y es menos afectada que la desviación estándar por los valores extremos. Su valor siempre será menor que la desviación estándar.

Para datos no agrupados:

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

Para datos agrupados:

$$DM = \frac{\sum f \cdot |X - \bar{x}|}{n}$$

Cuanto mayor sea el valor de la desviación media, mayor será la dispersión de los datos; sin embargo este valor no proporciona una relación matemática precisa con la posición de un dato dentro de la distribución y, puesto que se toman los valores absolutos, mide la desviación de una observación sin determinar si está por encima o por debajo de la media aritmética.

De la misma manera que la desviación estándar, a la desviación media puede calculársele el ***coeficiente de desviación media***:

$$CVM = \frac{DM}{\bar{x}} \times 100\%$$

EJEMPLO 1.17.

Los siguientes datos corresponden a los salarios de 10 empleados (en miles de pesos) de dos empresas de alimentos. Calcular los coeficientes de variación y de

[5] Recuerde que el valor absoluto de un número indica siempre su valor positivo. Por ejemplo: $|-2| = 2$; $|2| = 2$. Si requiere repasar este tema, se recomienda trabajar en los módulos de Matemáticas Básicas o Álgebra, Trigonometría y Geometría Analítica de la UNAD o cualquier otro texto de matemáticas básicas.

desviación media.

Empresa A:	$420	$680	$690	$720	$720	$720	$730	$740	$740	$760
Empresa B:	$415	$480	$510	$650	$700	$700	$730	$735	$750	$760

Empresa A:
Media aritmética: $\bar{x} = 692$
Varianza: $s^2 = 8716$
Desviación estándar: $s = 93{,}36$
Desviación media: $DM = 57{,}2$
Coeficiente de variación: $CV = \frac{93{,}36}{692} \times 100\% = 13{,}49\%$
Coeficiente de desviación media: $CVM = \frac{57{,}2}{692} \times 100\% = 8{,}27\%$

Empresa B:
Media aritmética: $\bar{x} = 643$
Varianza: $s^2 = 14396$
Desviación estándar: $s = 119{,}98$
Desviación media: $DM = 104{,}86$
Coeficiente de variación: $CV = \frac{119{,}98}{643} \times 100\% = 18{,}66\%$
Coeficiente de desviación media: $CVM = \frac{104{,}86}{643} \times 100\% = 16{,}31\%$

El CVM es menor que el CV debido a que la desviación media es menor que la desviación estándar.
Estos resultados llevan a las siguientes conclusiones:

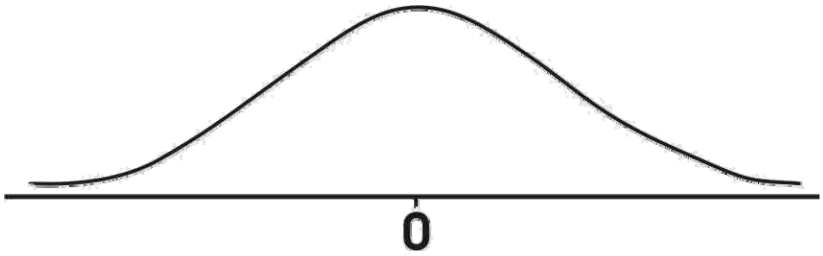
- El salario promedio de los 10 empleados de la empresa A es de $692.000, mientras que en la empresa B el salario promedio es de sólo $643.000.
- En la empresa B los salarios varían grandemente respecto al media: en 14396 miles de pesos cuadrados, que en términos de la desviación estándar esto es $119.980. En cambio, en la empresa A la variación es de $93.360.
- El coeficiente de variación y el coeficiente de variación media de la empresa B son menores a los coeficientes calculados para la empresa A, esto indica la variación relativa de los salarios en ambas empresas.

1.2.6. Puntaje típico o estandarizado

Cuando se tiene una distribución simétrica, su polígono de frecuencias

revelará una forma de campana muy común en estadística. Esta curva es llamada ***curva normal***, ***de error***, ***de probabilidad*** o ***campana de Gauss***. En ella la media aritmética se localiza en la mitad de la distribución. En el eje horizontal se ubican los valores que toma la variable y en el vertical la frecuencia absoluta o relativa. El área bajo la curva tendrá un valor del 100%

Figura 1.4.
Curva normal o campana de Gauss



El ***puntaje típico*** o ***estandarizado*** o ***variable normalizada***, es una medida de dispersión muy utilizada como variable estadística en este tipo de distribución, denominada ***distribución normal***. El puntaje estandarizado mide la desviación de una observación con respecto a la media aritmética, en unidades de desviación estándar, determinándose así la posición relativa de una observación dentro del conjunto de datos. Por lo general se simboliza por ***Z***, pero cuando el tamaño de la muestra es menor de 30, se simboliza por ***t***.

$$Z = \frac{X - \bar{x}}{s}$$

Por ser adimensional, el puntaje ***Z*** es útil para comparar datos individuales de distribuciones que tienen distintas unidades de medida, así como diferentes medias y desviaciones estándar. Dentro de sus propiedades, las más importantes son que su media es cero y su desviación estándar y varianza es uno.

EJEMPLO 1.18.

Al terminar el semestre, un grupo de 150 estudiantes de primer semestre de Regencia de Farmacia del CEAD de Medellín obtuvieron los siguientes resultados en el puntaje final de los cursos Lógica Matemática y Estadística Descriptiva:

- Lógica Matemática: puntuación media de 3.9 y varianza 3.2.
- Estadística Descriptiva: puntuación media de 3.7 y desviación estándar 1.7.

a. ¿En cuál curso hubo mayor ***dispersión absoluta***? ¿En cuál hubo mayor ***dispersión relativa***?

b. Si un estudiante obtuvo como nota final en Lógica Matemática 3.8 y en

Estadística Descriptiva 3.5. ¿En cuál curso fue su puntuación relativa superior?

a. Para determinar la dispersión absoluta, basta con hacer una comparación entre sus desviaciones estándar. Observe que en los datos suministrados, ya se tiene el valor de la desviación estándar de las calificaciones de Estadística Descriptiva en cambio, se tiene la varianza de las calificaciones de Lógica Matemática. Recuerde que la desviación estándar es la raíz cuadrada de la varianza.

Para Lógica Matemática: $s^2 = 3.2 \quad \rightarrow \quad s = \sqrt{3.2} = 1.79$

Se tiene entonces que en Lógica Matemática hubo una mayor dispersión absoluta, pues 1.79>1.7, aunque no es mucha la diferencia.

Para saber en cuál hubo mayor dispersión relativa, se recurre al coeficiente de variación:

Para Lógica Matemática: $CV = \frac{1.79}{3.9} \times 100 = 45.9\%$

Para Estadística Descriptiva: $CV = \frac{1.7}{3.7} \times 100 = 46\%$

En Estadística Descriptiva hubo una mayor dispersión relativa 46%>45.9%

b. Para el cálculo de la puntuación relativa, se hace uso del puntaje estandarizado. Es decir, se requiere estandarizar las calificaciones convirtiéndolas en puntuaciones ***Z***.

Lógica Matemática: $Z = \frac{x - \bar{x}}{s} = \frac{3.8 - 3.9}{1.79} = -0.06$

Estadística Descriptiva: $Z = \frac{x - \bar{x}}{s} = \frac{3.5 - 3.7}{1.7} = -0.12$

Estos valores de puntuación ***Z*** negativos indican que ambas calificaciones se encuentran por debajo de la media. Este es un principio del puntaje estandarizado: *Siempre que un valor sea menor que la media, su puntuación **Z** correspondiente será negativa.*

Estos resultados afirman entonces que el estudiante con calificaciones de 3.8 en Lógica Matemática y 3.5 en Estadística Descriptiva, está por debajo del promedio del grupo en ambos cursos.

Dado que -0.06 se encuentra más cera a 0 (la media de la variable estandarizada), se dice que la puntuación relativa del estudiante fue superior en Lógica Matemática.

EJERCICIOS TEMA 1.2.

1. En un café Internet, el rango de tiempo de uso en un mes es de 27 minutos, si el mayor tiempo de consulta en ese mes duró 1 hora y 12 minutos, halle el menor tiempo de consulta en ese mes.

2. Calcule el rango intercuartílico y semiintercuartílico de los datos agrupados en la tabla del numeral 9 de los ejercicios del tema 1.1. de esta Unidad Didáctica.

3. Halle el rango, la varianza, la desviación estándar, desviación media y el coeficiente de variación de las siguientes series:

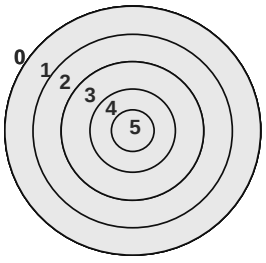
a.	5	6	3	8	0	1			
b.	2.35	3.16	1.20	2.10	5.32	4.8			
c.	3	1	0	2	1	0	2	0	3
d.	5.35	6.16	4.20	5.10	8.32	7.8			

4. Tome los datos de la tabla de distribución de frecuencias agrupadas del numeral 2c de los ejercicios del tema 1.1. de esta Unidad Didáctica y determine varianza y desviación estándar.

5. Tome los datos del ejemplo 2.1., de la Unidad Didáctica 1 que representan la evaluación de los latidos cardíacos de un grupo de 30 personas después de cierta actividad física. Continúe con esos datos para terminar el análisis completo de ese fenómeno y ahora calcule varianza, desviación estándar, desviación media.

6. En una prueba de tiro al blanco de cinco anillos, dos competidores Johan y Samantha obtuvieron los resultados que se indican a continuación. Determine, usando medidas estadísticas, quién es el mejor.

Johan	Samantha
1 Tiro de 5 Puntos	4 Tiros de 5 Puntos
8 Tiros de 4 Puntos	9 Tiros de 4 Puntos
14 Tiros de 3 Puntos	7 Tiros de 3 Puntos
5 Tiros de 2 Puntos	5 Tiros de 2 Puntos
1 Tiro de 1 Punto	3 Tiros de 1 Punto
1 Tiro de 0 Puntos	2 Tiros de 0 Puntos



7. Un fabricante de bombillas de neón tiene dos tipos de tubos, A y B. Los tubos tienen unas duraciones medias respectivas de 1.495 horas y 1.875 horas, y desviaciones estándar de 280 horas y 310 horas respectivamente.

a. ¿Qué tubo tiene la mayor dispersión absoluta?
b. ¿Qué tubo tiene la mayor dispersión relativa?
c. Si se extrajo un tubo de cada tipo y su duración fue de 1.350 horas y 1.750 horas respectivamente, ¿cuál tipo de tubo tiene menor posición relativa?

8. Dada la serie de puntuaciones 9, 5, 6, 11, 1, 2, 10, 4, hallar el puntaje estandarizado de cada puntuación

9. Las estaturas de los hombres adultos tienen una media de 1,75 metros y una desviación estándar de 7 centímetros. Calcule las puntuaciones Z que corresponden a las siguientes personas:
a. Carlos Alberto que mide 156 centímetros.
b. Juan José que mide 1,81 metros.
c. Francisco que mide 1,68 metros.

10. En un grupo de estudiantes la estatura promedio es 163,1 cm., con una desviación estándar de 9,38 cm. y su peso promedio es de 61,3 kg con desviación estándar 11,7 kg. Mauricio mide 1,70 metros y pesa 63 kg, calcule:
a. La puntuación estandarizada de cada medida.
b. ¿En cuál de las dos medidas hay mayor dispersión absoluta?
c. ¿En cuál de las dos medidas hay menor dispersión relativa?

1.3. MEDIDAS DE ASIMETRÍA Y APUNTAMIENTO

Después de conocer cómo varía un grupo de datos respecto a su media e identificar otras medidas de variación, se trabajará a continuación unas nociones básicas sobre curvas asimétricas. En cursos más avanzados, como Probabilidad, este tema se profundiza más, pero para los objetivos que se trazan en este curso basta con las nociones que se desarrollan a continuación.

1.3.1. Asimetría

Ya se ha mencionado algo sobre los efectos de la asimetría respecto a la media, mediana y moda (ver sección 1.1.3. de esta Unidad Didáctica). En una distribución ***simétrica*** se tiene que:

$$\bar{x} = Me = Mo$$

En las distribuciones ***asimétricas*** la media se corre en el sentido del alargamiento o ***sesgo*** por efecto de las frecuencias y de los valores extremos de la variable; la mediana también se corre pero menos que la media ya que en ella sólo influyen las frecuencias; en tanto que la moda no es influenciada ni por las frecuencias ni por los valores extremos (ver figura 1.1. de la presente Unidad Didáctica). La distribución es ***asimétrica positiva*** cuando presenta un alargamiento o sesgo a la derecha y:

$$Mo < Me < \bar{x}$$

Será ***asimétrica negativa*** cuando presenta un alargamiento o sesgo a la izquierda y:

$$\bar{x} < Me < Mo$$

Las asimetrías positivas son las más frecuentes que las sesgadas hacia la izquierda, porque con frecuencia es más fácil obtener valores excepcionalmente grandes que valores excepcionalmente pequeños. Ejemplo de ello es la distribución de valores en los consumos de servicios públicos, las calificaciones en pruebas, los sueldos, etc.

Se reconocen, entre otras, las siguientes medidas para calcular el grado de la asimetría:

- ***Coeficiente de Pearson***. Asimetría en función de la media y la moda. Varía entre ±3 y es 0 en la distribución normal.

$$As = \frac{\bar{x} - Mo}{s} \quad \Leftrightarrow \quad As = \frac{3 \cdot (\bar{x} - Me)}{s}$$

- ***Media cuartil de asimetría*** o ***media de Bowley***. Varía entre ±1 y es 0 en la distribución normal.

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Si $As = 0$ la distribución es *simétrica.*
Si $As > 0$ la distribución es *asimétrica positiva.*
Si $As < 0$ la distribución es *asimétrica negativa.*

1.3.2 Apuntamiento o curtosis

Las curvas de distribución, comparadas con la curva de distribución normal, pueden presentar diferentes grados de ***apuntamiento*** o altura de la cima de la curva. Esta agudeza en la cima se observa en la moda.

Si la curva es más plana que la normal se dice que la curva es ***platicúrtica***; si es más aguda que la normal, recibe el nombre de apuntada o ***leptocúrtica.*** Si la distribución es normal, la curva se conoce también como ***mesocúrtica***.

La ***curtosis*** es la medida de la altura de la curva y esta dada por:

$$Ap = \frac{\sum Z_i^{\,4} \cdot f_i}{n \cdot s^4}$$

Si $Ap = 3$ la distribución es *normal* o *mesocúrtica.*
Si $Ap > 3$ la distribución es *apuntada* o *leptocúrtica.*
Si $Ap < 3$ la distribución es *achatada* o *platicúrtica.*

Otra medida de curtosis que se emplea está basada en el rango semiintercuartílico y los percentiles 10 y 9:

$$Ap = \frac{Q_{D2}}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

En el siguiente ejemplo se explicarán mejor las medidas de asimetría y apuntamiento.

EJEMPLO 1.19.

El coordinador académico del CEAD de Valledupar desea conocer el rendimiento académico de los estudiantes de primer semestre en el 2005, en los cursos de Lógica Matemática, Competencias Comunicativas, Cultura Política, Estadística Descriptiva y Herramientas Informáticas. Para esto selecciona una muestra de 55 estudiantes de los distintos programas que se ofrecen en el CEAD. La siguiente tabla, arroja los resultados de la investigación realizada por el funcionario.

Tabla 1.11.
Distribución de frecuencias
de las calificaciones de primer semestre en Valledupar

Calificación	Lógica Matemática	Competencias Comunicativas	Cultura Política	Estadística Descriptiva	Herramientas Informáticas
0,0	1	3	2	1	1
0,5	4	3	2	1	2
1,0	7	5	3	2	3
1,5	9	6	4	4	7
2,0	9	7	6	11	9
2,5	8	7	8	14	11
3,0	6	7	9	12	9
3,5	4	6	9	6	7
4,0	3	5	7	3	3
4,5	2	3	4	1	2
5,0	2	3	1	0	1
Total	**55**	**55**	**55**	**55**	**55**

En la tabla siguiente se reporta un resumen de las medidas estadísticas por cada uno de los cursos (¡compruébelo!):

Medida	Lógica Matemática	Competencias Comunicativas	Cultura Política	Estadística Descriptiva	Herramientas Informáticas
$\overline{x}$	2.25	2.5	2.75	2.53	2.5
Me	2.0	2.5	3.0	2.5	2.5
Mo	1.5 y 2.0	2.0, 2.5 y 3.0	3.0 y 3.5	2.5	2.5
s^2	1.45	1.84	1.45	0.76	1.12
s	1.20	1.36	1.20	0.87	1.06
Q_1	1.5	1.5	2.0	2.0	2.0
Q_2	2.0	2.5	3.0	2.5	2.5
Q_3	3.0	3.5	3.5	3.0	3.4

Lógica Matemática

Se observa que $Mo < Me < \bar{x}$, lo que indica que la distribución es asimétrica positiva. Para confirmarlo se hace uso del coeficiente de Pearson y la media de Bowley: En este caso se trabajará con la media de Bowley, pues la distribución tiene dos modas y no permite un resultado seguro con el coeficiente de Pearson.

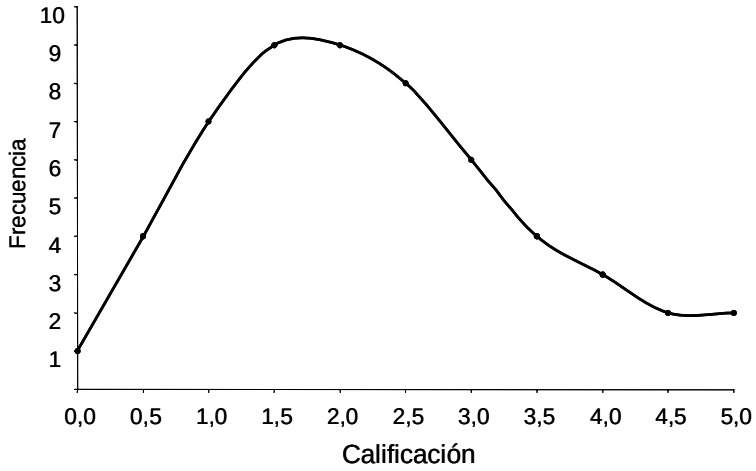
$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{1.5 + 3 - 2(2)}{3 - 1.5} = 0.33 > 0$$

El polígono de frecuencias de las calificaciones de Lógica Matemática confirma los resultados.

Figura 1.5.

Curva asimétrica positiva

Polígono de frecuencias de calificaciones de Lógica Matemática



La curva lleva a concluir que la mayoría de los estudiantes están por debajo de la media en el curso de Lógica Matemática y son pocos los estudiantes que la superan.

Competencias Comunicativas

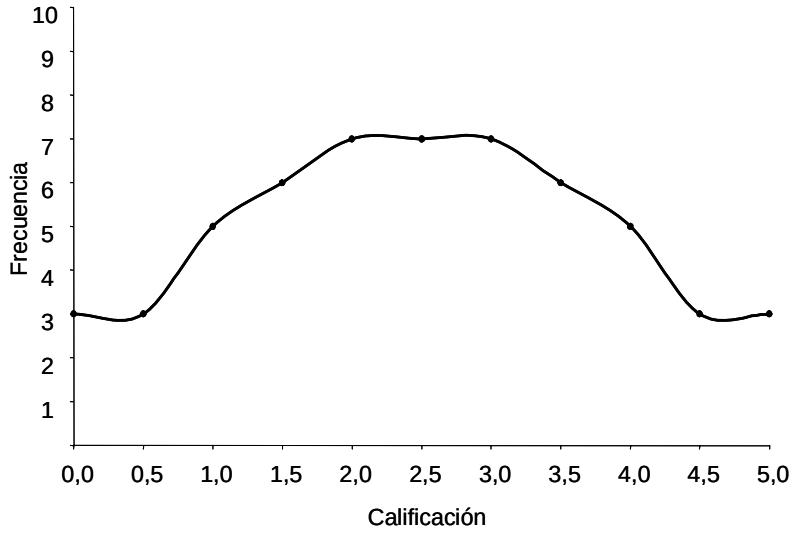
Se observa que $Mo = Me = \bar{x}$, lo que indica que la distribución es simétrica. Para confirmarlo se hace uso del coeficiente de Bowley, pues la distribución tiene tres modas y no permite un resultado seguro con el coeficiente de Pearson.

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{1.5 + 3.5 - 2(2.5)}{3.5 - 1.5} = 0$$

El polígono de frecuencias de las calificaciones de Competencias Comunicativas

confirma los resultados.

Figura 1.6.
Curva simétrica platicúrtica
Polígono de frecuencias de calificaciones de Competencias Comunicativas



Para determinar el grado de apuntamiento o curtosis, se debe determinar el puntaje típico o estandarizado de cada clase y luego aplicar la fórmula que lo calcula. En la siguiente tabla se indican estos valores.

Tabla 1.12.
Cálculo de ***Z*** para la distribución de frecuencias
de las calificaciones de Competencias Comunicativas

Calificación	f	Z	$Z_i^4 f_i$
0,0	3	-1,838235294	34,2551328
0,5	3	-1,470588235	14,0309024
1,0	5	-1,102941176	7,39910869
1,5	6	-0,735294118	1,7538628
2,0	7	-0,367647059	0,12788583
2,5	7	0	0
3,0	7	0,367647059	0,12788583
3,5	6	0,735294118	1,7538628
4,0	5	1,102941176	7,39910869
4,5	3	1,470588235	14,0309024
5,0	3	1,838235294	34,2551328
Total	**55**	**0**	**115,133785**

$$Ap = \frac{\sum Z_i^{\,4} f_i}{n \cdot s^4} \Rightarrow Ap = \frac{115.13}{55 \times 1.36^4} = 0.62 < 3$$

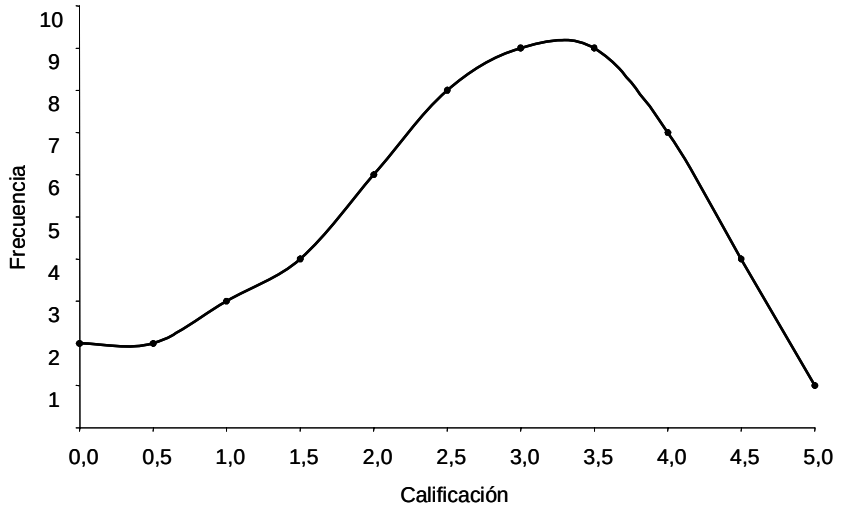
Por lo tanto, la curva es simétrica platicúrtica o achatada.
Estos resultados indican que la mayoría de los estudiantes en Competencias Comunicativas están en el rango de la media del curso, además sus notas son muy homogéneas alrededor de la media.

Cultura Política
Se observa que $Mo > Me > \bar{x}$, lo que indica que la distribución es asimétrica negativa. Para confirmarlo se hace uso de la media de Bowley, pues la distribución tiene dos modas y no permite un resultado seguro con el coeficiente de Pearson.

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{2.0 + 3.5 - 2(3.0)}{3.5 - 2.0} = -0.33 < 0$$

El polígono de frecuencias de las calificaciones de Cultura Política confirma los resultados.

Figura 1.7.
Curva asimétrica negativa
Polígono de frecuencias de calificaciones de Cultura Política



Esto quiere decir que las calificaciones de la mayoría de los estudiantes del curso Cultura Política están por encima de la media.

Se observa que $Mo = Me = \bar{x}$, lo que indica que la distribución es simétrica. Para confirmarlo se hace uso del coeficiente de Pearson y la media de Bowley:

$$As = \frac{\bar{x} - Mo}{s} = \frac{2.53 - 2.5}{0.87} = 0.03 \approx 0$$

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{2.0 + 3.0 - 2(2.5)}{3.0 - 2.0} = 0$$

Para determinar el grado de apuntamiento o curtosis, se debe determinar el puntaje típico o estandarizado de cada clase y luego aplicar la fórmula que lo calcula. En la tabla siguiente tabla se indican estos valores.

Tabla 1.13.

Cálculo de ***Z*** para la distribución de frecuencia de las calificaciones de Estadística Descriptiva

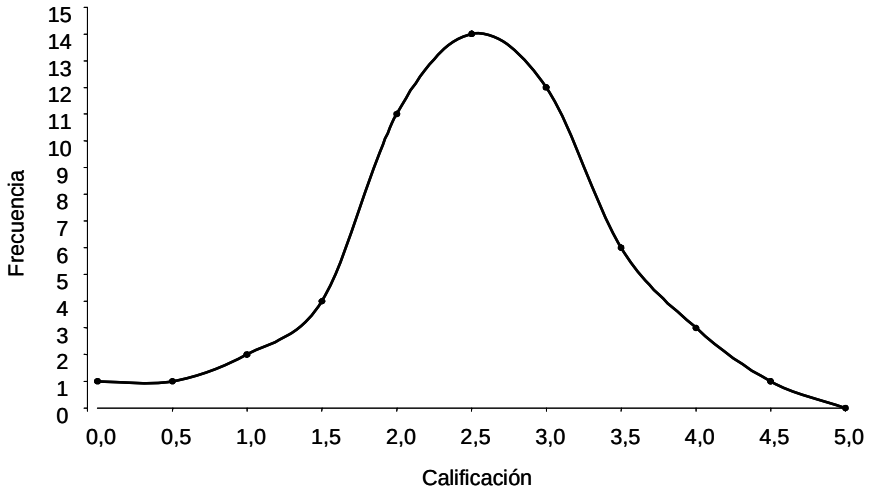
Calificación	f	Z	$Z_i^4 f_i$
0,0	1	-2,908045977	71,516306
0,5	1	-2,333333333	29,6419753
1,0	2	-1,75862069	19,1301647
1,5	4	-1,183908046	7,85835926
2,0	11	-0,609195402	1,51502275
2,5	14	-0,034482759	1,9794E-05
3,0	12	0,540229885	1,02210536
3,5	6	1,114942529	9,27173856
4,0	3	1,689655172	24,4519547
4,5	1	2,264367816	26,289837
5,0	0	-1,352941176	0
Total	**55**	**-4,571331981**	**190,697484**

$$Ap = \frac{\sum Z_i^4 f_i}{n \cdot s^4} \Rightarrow Ap = \frac{190.70}{55 \times 0.87^4} = 6.05 > 3$$

Por lo tanto, la curva es simétrica leptocúrtica o apuntada.

Esto indica que las calificaciones de Estadística Descriptiva de la muestra de 55 estudiantes están muy cerca de la media y que existe además, un pico en 2.5, señalando una alta frecuencia en esta calificación.

Figura 1.8.
Curva simétrica leptocúrtica
Polígono de frecuencias de calificaciones de Estadística Descriptiva



Herramientas Informáticas

Se observa que $Mo = Me = \bar{x}$, lo que indica que la distribución es simétrica. Para confirmarlo se hace uso del coeficiente de Pearson:

$$As = \frac{\bar{x} - Mo}{s} = \frac{2.5 - 2.5}{1.06} = 0$$

El polígono de frecuencias de las calificaciones de Herramientas Informáticas confirma los resultados. La curva es simétrica mesocúrtica o normal. Verifíquelo y construya la gráfica.

EJERCICIOS TEMA 1.3.

1. Determine el tipo de asimetría de las siguientes distribuciones con sus estadígrafos de dispersión:

a. $\bar{x} = 189{,}97 \qquad Me = 189{,}7 \qquad Mo = 189{,}16$

b. $\bar{x} = 5{,}3 \qquad Me = 5 \qquad Mo = 4$

c. $\bar{x} = 17{,}5 \qquad Me = 17{,}9 \qquad Mo = 18{,}1$

2. Tomando una distribución ligeramente simétrica, calcule su moda sabiendo que su media es 3 y que la diferencia entre la media y la mediana es igual a -2.

3. Con los salarios semanales de los empleados de una empresa se tienen los siguientes resultados:

$$\bar{x} = 9725 \qquad Me = 9672 \qquad s = 1217{,}50$$

Calcule el coeficiente de asimetría de Pearson.

4. Calcule los coeficientes de asimetría y los coeficientes de apuntamiento de las siguientes distribuciones correspondientes a la edad de los niños quemados por pólvora reportados en tres centros hospitalarios durante el mes de diciembre:

X	f	f	f
5	3	3	6
7	19	7	8
9	10	8	11
11	8	9	11
13	7	20	8
15	3	3	6
Total	**50**	**50**	**50**

Construya sus respectivos polígonos de frecuencia y haga un análisis comparativo de los resultados obtenidos.

2. MEDIDAS ESTADÍSTICAS BIVARIANTES

Hasta ahora se ha estudiado el análisis de una sola variable, calculando los estadísticos de muestras que permiten describir e interpretar la distribución de esa variable. En este capítulo se estudiará el comportamiento de dos variables: *distribuciones bivariantes*, con el fin de determinar si existe alguna relación entre las variables, que bien pudieran ser ambas discretas o continuas, o también una de ellas discreta y la otra continua. En este capítulo se desarrolla el tema de la Regresión y Correlación lineal y los Números Índice.

2.1. REGRESIÓN Y CORRELACIÓN

En muchos casos se requiere conocer más que el comportamiento de una sola variable, se requiere conocer la relación entre dos o más variables como la relación entre producción y consumo; salarios y horas de trabajo; oferta y demanda; salarios y productividad; la altura de un árbol y el diámetro de su tronco; el nivel socioeconómico de una persona y su grado de depresión; etc.

Muchos de estos comportamientos tienen una tendencia lineal, aunque hay muchos otros que lo hacen de forma curva, en este curso sólo se trabajará sobre variables con correlación lineal. A continuación se describirá brevemente en qué consiste un diagrama de dispersión y cuáles son los criterios que deben tenerse en cuenta para hallar la mejor línea o línea de tendencia del comportamiento de las variables.

2.1.1. Diagrama de dispersión

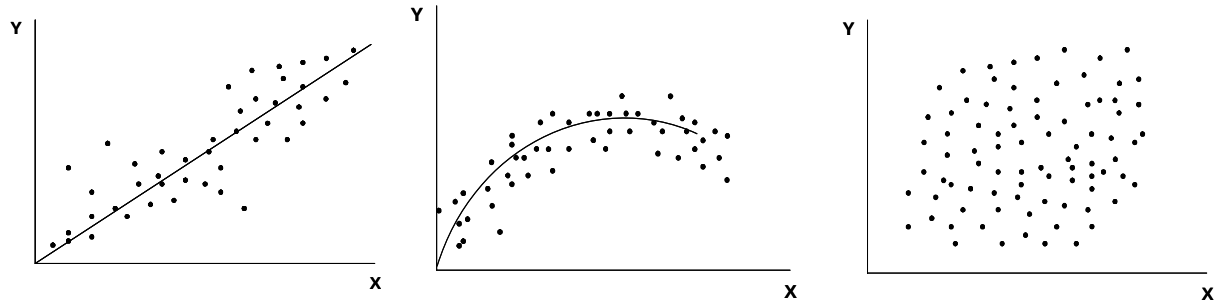
Una distribución bidimensional o bivariante puede representarse gráficamente en un plano cartesiano, ubicando en el eje horizontal o abscisa los valores de la primera variable denominada X y en el eje vertical u ordenada, los valores de la segunda variable, Y. De manera pues que se grafican tantas parejas ordenadas como observaciones hayan de las variables.

A este conjunto de puntos o nube de puntos se le denomina ***diagrama de dispersión***, dado que los puntos se ubican de forma dispersa en el plano cartesiano.

En muchos casos el sólo diagrama de dispersión indica una tendencia de agrupación de los puntos, que puede ser lineal (hacia arriba o hacia abajo), exponencial, curvilínea o poligonal.

Parte del análisis estadístico que hace el investigador es determinar cuál es la mejor línea o curva que representa a ese conjunto de datos. El mejor ajuste se hace cuando se elabora bien la gráfica, se conoce la distribución y se va adquiriendo experiencia en su cálculo y determinación.

Figura 2.1.
Gráficas de dispersión
(a) lineal; (b) curvilínea; (c) sin relación



(a) **(b)** **(c)**

2.1.2. Regresión lineal simple

La regresión examina la relación entre dos variables restringiendo una de ellas respecto a la otra, con el objeto de estudiar las variaciones de la primera cuando la otra permanece constante. La regresión es un método que se emplea para pronosticar o predecir el valor de una variable en función de los valores dados de la otra (o de las otras, cuando se trabaja más de dos variables).

Se trata pues de una dependencia funcional entre las variables. Cuando se trata de dos variables, una (la X) será la ***variable independiente*** mientras que la otra (la Y) será la ***variable dependiente.*** Se habla así de una regresión de Y sobre (o en función de) X.

Cuando se considera, después de una inspección en la gráfica de dispersión, que una línea recta es la mejor curva que se ajusta al conjunto de puntos se procede entonces a emplear el método de la ***regresión lineal simple***. La mejor línea es aquella que hace mínima la suma de los cuadrados de las diferencias entre los puntos dados y los obtenidos mediante la línea ajustada o estimada. Es por eso que a este método también se le conoce como el ***método de los mínimos cuadrados***. La ecuación de la recta estimada está dada por:

$$\hat{Y} = a + bX$$

Donde:

$\hat{Y}$: Variable dependiente (la que se va a predecir)
a: Intercepto de la variable $\boldsymbol{Y}$
X: Variable independiente
b: Pendiente de la recta

En esta ecuación hay dos valores desconocidas: a y b, que deben determinarse aplicando el criterio de los mínimos cuadrados, buscando así la mejor recta que se ajuste a los datos. Se tiene entonces:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} \qquad a = \frac{\sum Y - b\sum X}{n}$$

Donde:

b: Pendiente de la recta
a: Intercepto de la variable $\boldsymbol{Y}$
X: Valores de la variable independiente
Y: Valores de la variable dependiente
n: Tamaño de la muestra

Algunos autores calcular los valores de a y b en términos de las medias de de los conjuntos de datos con las siguientes dos ecuaciones:

$$b = \frac{\sum (X - \bar{x})(Y - \bar{y})}{\sum (X - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

Donde:

X: Valores de la variable independiente
$\bar{x}$: Media del conjunto de datos de la variable $\boldsymbol{X}$
Y: Valores de la variable dependiente
$\bar{y}$: Media del conjunto de datos de la variable $\boldsymbol{Y}$

EJEMPLO 2.1.

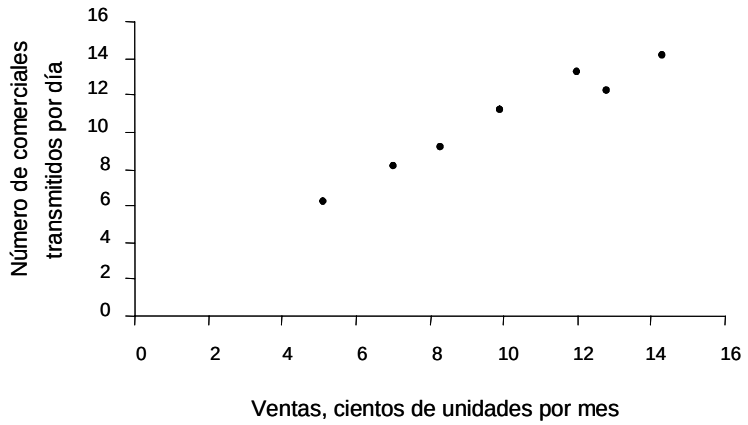
El departamento de publicidad de una industria alimenticia desea saber si existe una relación entre las ventas y el número de comerciales de televisión transmitidos por día. Para ello, toma una muestra aleatoria de siete ciudades. La siguiente tabla muestra los resultados obtenidos.

Tabla 2.1.
Relación de ventas de un producto y la emisión del comercial en televisión

Ventas Cientos de unidades por mes	**Comerciales** Número transmitido por día
8,4	9
5,2	6
7,1	8
10	11
12,9	12
12,1	13
14,4	14

Para conocer el tipo de relación que puede existir entre estas dos variables, el primer paso es determinar es si el diagrama de dispersión efectivamente insinúa una tendencia lineal.

Figura 2.2.
Diagrama de dispersión de ventas de un producto
y la emisión del comercial en televisión



El diagrama confirma la sospecha, se procede ahora a determinar la ecuación de la recta que más se ajusta. Para ello se hace uso del método de los mínimos cuadrados[6].

$$\hat{Y} = a + bX$$

Donde:

[6] Puede usarse cualquiera de las ecuaciones propuestas, la decisión la toma el investigador. En este ejemplo se presenta el cálculo con las dos ecuaciones de modo que el estudiante tenga criterio para decidir cómo hacer sus propios cálculos.

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} \qquad a = \frac{\sum Y - b\sum X}{n}$$

X Ventas	***Y*** Comerciales	***XY***	X^2
8,4	9	75,6	70,56
5,2	6	31,2	27,04
7,1	8	56,8	50,41
10	11	110	100
12,9	12	154,8	166,41
12,1	13	157,3	146,41
14,4	14	201,6	207,36
70,1	**73**	**787,3**	**768,19**

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{7 \times 787{,}3 - (70{,}1)(73)}{7 \times 768{,}19 - (70{,}1)^2} = \frac{393{,}8}{463{,}32} = 0{,}85$$

$$a = \frac{\sum Y - b\sum X}{n} = \frac{73 - (0.85 \times 70{,}1)}{7} = \frac{13{,}415}{7} = 1{,}92$$

De modo que la ecuación de la recta ajustada está dada por:

$$\hat{Y} = 0{,}85X + 1{,}92$$

Si se quisiera hacer el cálculo con la segunda ecuación planteada, se debe determinar primero las medias de cada conjunto de datos.

$$\bar{x} = \frac{\sum X}{n} = \frac{70.1}{7} = 10.01 \qquad \bar{y} = \frac{\sum Y}{n} = \frac{73}{7} = 10.43$$

En la siguiente tabla se resumen todos los cálculos necesarios para determinara la ecuación de la recta ajustada. Se tiene entonces:

$$b = \frac{\sum (X - \bar{x})(Y - \bar{y})}{\sum (X - \bar{x})^2} = \frac{56{,}2571}{66{,}1887} = 0{,}85$$

$$a = \bar{y} - b\bar{x} = 10{,}43 - (0{,}85)(10{,}01) = 1{,}92$$

La ecuación de la recta ajustada está dada por:

$$\hat{Y} = 0{,}85X + 1{,}92$$

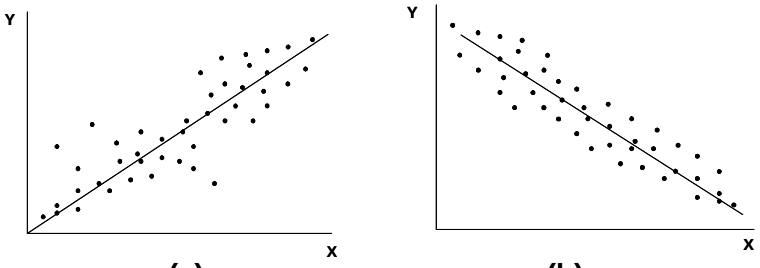
X Ventas	***Y*** Comerciales	$X - \bar{x}$	$Y - \bar{y}$	$(X - \bar{x})(Y - \bar{y})$	$(X - \bar{x})^2$
8,4	9	-1,61	-1,43	2,3023	2,5921
5,2	6	-4,81	-4,43	21,3083	23,1361
7,1	8	-2,91	-2,43	7,0713	8,4681
10	11	-0,01	0,57	-0,0057	0,0001
12,9	12	2,89	1,57	4,5373	8,3521
12,1	13	2,09	2,57	5,3713	4,3681
14,4	14	4,39	3,57	15,6723	19,2721
70.1	**73**	**0,03**	**-0,01**	**56,2571**	**66,1887**

2.1.3. Correlación

La correlación entre dos variables busca determinar el grado de relación que existe entre ellas dos. Ella se calcula con los ***coeficientes de correlación***.

Los coeficientes de correlación son números que varían entre +1 y -1. Su magnitud indica el grado de asociación entre las variables, si es 0 indica que no existe relación alguna y los valores extremos +1 y -1 indican una correlación perfecta positiva o negativa respectivamente.

Figura 2.3.
Gráficas de dispersión lineal
(a) positiva; (b) negativa



(a) **(b)**

Se dice que existe una ***correlación lineal positiva*** entre dos variables, si al aumentar o disminuir los valores de la variable independiente aumentan o

disminuyen los de la variable dependiente. En un gráfico de dispersión, la nube de puntos tiene forma ascendente y por tanto la recta que se ajusta tendrá una pendiente positiva. En cambio, cuando al aumentar los valores de la variable independiente disminuyen los valores de la variable dependiente, o viceversa, se dice que la correlación lineal es ***negativa***. En este caso la nube de puntos descenderá de izquierda a derecha y la pendiente de la recta ajustada será negativa (ver figura 2.3.)

Para determinar el coeficiente de correlación, es necesario conocer primero el ***error estándar del estimado*** de la recta ajustada. Se trata pues de medir el grado de confiabilidad de la ecuación de la recta estimada. El error estándar indicará la dispersión o la variabilidad de los valores observados alrededor de la línea de regresión y se calcula a partir de la siguiente ecuación:

$$Se = \sqrt{\frac{\sum (Y - \widehat{Y})^2}{n-2}}$$

Donde:

Se: Error estándar del estimado
Y: Valores de la variable dependiente
$\widehat{Y}$: Valores estimados de la ecuación
n: Tamaño de la muestra

Esta ecuación implica demasiadas operaciones, por lo que suele utilizarse un método más breve:

$$Se = \sqrt{\frac{\sum Y^2 - a\sum X - b\sum XY}{n-2}}$$

Una vez obtenido el error estándar del estimado, es necesario medir qué porcentaje de la información es recogida o explicada por el modelo de regresión escogido. Se trata pues, de determinar las variaciones de la variable dependiente mediante el ***coeficiente de determinación (R^2)***.

$$R^2 = 1 - \frac{Se^2}{s_y^2}$$

Donde:

R^2: Coeficiente de determinación, $0 \le R^2 \le 1$
Se^2: Varianza del error estimado.

S_y^2: Varianza de la variable dependiente $\boldsymbol{Y}$.

Cuando el R^2 es cercano a 1, se dice que el modelo de regresión lineal ajustado tiene un alto grado de confiabilidad, si al contrario este se acerca a 0 su grado de confiabilidad es muy bajo y se recomienda no utilizar el modelo de regresión estimado.

En la práctica es más frecuente usar r, denominado el ***coeficiente de correlación lineal***. Siendo $r = \sqrt{R^2}$. El coeficiente de correlación lineal *r*, es también conocido como ***coeficiente de Pearson***. Ya se mencionaba que el coeficiente de correlación lineal oscila entre +1 y -1, se puede entonces interpretar el grado de correlación partiendo de los siguientes límites de referencia:

Tabla 2.2.
Grado de correlación lineal

Interpretación	Valores de *r* (+)	Valores de *r* (-)
Correlación perfecta	= 1	= -1
Correlación excelente	0.90 < r < 1	-1 < r < -0.90
Correlación aceptable	0.80 < r < 0.90	-0.90 < r < -0.80
Correlación regular	0.60 < r < 0.80	-0.80 < r < -0.60
Correlación mínima	0.30 < r < 0.60	-0.60 < r < -0.30
No hay correlación	0 < r < 0.30	-0.30 < r 0

Tomado de "Estadística Básica Aplicada"; Ciro Martínez Bencardino.

EJEMPLO 2.2.

Determinar el error estándar de la recta ajustada en el ejemplo 2.1.

$$Se = \sqrt{\frac{\sum Y^2 - a\sum X - b\sum XY}{n-2}} = \sqrt{\frac{811-(1.92)(70.1)-(0.85)(787.3)}{7-2}} = 1.2$$

Se calcula así, el coeficiente de determinación y el coeficiente de correlación lineal. Para ello se determina S_y^2, la varianza de la variable dependiente $\boldsymbol{Y}$.

$$s_y^2 = \frac{\sum Y^2}{n} - \bar{y}^2 = \frac{811}{7} - 10.43^2 = 7.07$$

$$R^2 = 1 - \frac{Se^2}{s_y^2} = 1 - \frac{1.44}{7.07} = 0.8 \quad \Rightarrow \qquad r = \sqrt{R^2} = 0.89$$

Con los resultados obtenidos se puede asegurar que la ecuación de la recta es una muy buena estimación de la relación entre las dos variables. El $\boldsymbol{R^2}$ afirma además que el modelo explica el 80% de la información. Y el valor de r confirma además el grado de relación entre las variables: el número de ventas del producto está directamente relacionado (en un 89%) con los comerciales de televisión que se emiten diariamente.

Ahora, si el gerente de ventas de la empresa quisiera aumentar el número de ventas del producto a 2000 mensuales, ¿Cuántos comerciales estima el departamento de publicidad de la empresa que debe emitir diariamente?

Se trata simplemente de reemplazar en la ecuación estimada, la variable independiente por el valor que se pretende y así obtener el valor de la variable dependiente (número de comerciales). Así:

$$\widehat{Y} = 0.85X + 1.92 \qquad \Rightarrow \qquad \widehat{Y} = (0.85)(20) + 1.92 = 18.92 \approx 19$$

El departamento de publicidad requerirá de 19 comerciales de televisión diariamente para que el número de ventas ascienda a 2000 unidades mensuales.

2.1.4. Regresión múltiple

Cuando se emplea más de una variable independiente para evaluar una variable dependiente es conveniente utilizar un método de ***regresión múltiple***, que consiste en el mismo procedimiento de una regresión lineal simple: describir la ecuación de regresión, determinar el error de estimación y analizar la correlación entre las variables.

A continuación se desarrollarán estos conceptos suponiendo dos variables independientes. Para más variables independientes, sólo basta con seguir los mismos pasos.

La ecuación de regresión está dada por:

$$\widehat{Y} = a + b_1 X_1 + b_2 X_2$$

Donde:

$\hat{Y}$: Variable dependiente.
a: Intercepto de la variable Y.
X_1, X_2: Valores de las dos variables independientes.
b_1, b_2: Pendientes asociadas con cada variable independiente, respectivamente.

Los valores de las tres constantes numéricas se obtienen resolviendo el siguiente sistema de ecuaciones:

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$
$$\sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2$$
$$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$$

Una vez obtenida la ecuación de regresión, se determina el ***error estándar de la estimación de regresión múltiple***:

$$Se = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-3}} \quad \Leftrightarrow \quad Se = \sqrt{\frac{\sum Y^2 - a\sum Y - b_1 \sum X_1 Y - b_2 \sum X_2 Y}{n-3}}$$

Y el ***coeficiente de determinación múltiple***, estará dado por:

$$R^2 = \frac{a\sum Y + b_1 \sum X_1 Y + b_2 \sum X_2 Y - n\bar{y}^2}{\sum Y^2 - n\bar{y}^2}$$

Donde:
Y: Valores de la variable dependiente.
a: Intercepto de la variable Y.
X_1, X_2: Valores de las dos variables independientes.
b_1, b_2: Pendientes asociadas con cada variable independiente, respectivamente.
$\bar{y}$: Media de los valores de la variable dependiente.

EJEMPLO 2.3.

El jefe de producción de una empresa manufacturera desea estimar los gastos indirectos de producción con base en el número de horas de trabajo y en el número de horas máquina. En la siguiente tabla se relaciona la información correspondiente al primer semestre del año.

El jefe de producción define:
X_1 : Horas de trabajo (cientos).
X_2 : Horas de máquina (cientos)
Y : Gastos indirectos de producción (cientos de miles de pesos)

Tabla 2.3.
Gastos indirectos de producción

Mes	X_1	X_2	Y	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2	Y^2
Enero	45	16	29	1305	464	720	2025	256	841
Febrero	42	14	24	1008	336	588	1764	196	576
Marzo	44	15	27	1188	405	660	1936	225	729
Abril	45	13	25	1125	325	585	2025	169	625
Mayo	43	13	26	1118	338	559	1849	169	676
Junio	46	14	28	1288	392	644	2116	196	784
TOTAL	**265**	**85**	**159**	**7032**	**2260**	**3756**	**11715**	**1211**	**4231**

$$\sum Y = na + b_1\sum X_1 + b_2\sum X_2 \qquad 159 = 6a + 265b_1 + 85b_2 \qquad (1)$$
$$\sum X_1Y = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1X_2 \quad \Rightarrow \quad 7032 = 265a + 11715b_1 + 3756b_2 \qquad (2)$$
$$\sum X_2Y = a\sum X_2 + b_1\sum X_1X_2 + b_2\sum X_2^2 \qquad 2260 = 85a + 3756b_1 + 1211b_2 \qquad (3)$$

Se resuelve el sistema de ecuaciones:

Ecuación (1) multiplicada por 85/6 y restada por la ecuación (3):

$$\begin{aligned} 2252.5 &= 85a + 13754.17b_1 + 1204.17b_2 \\ -2260 &= -85a - 3756b_1 - 1211b_2 \\ \hline -7.5 &= -1.83b_1 - 6.83b_2 \end{aligned} \qquad (4)$$

Se despeja la variable b_1 de la ecuación (4):

$$b_1 = \frac{7.5 - 6.83b_2}{1.83}$$

Ecuación (1) multiplicada por 265/6 y restada por ecuación (2):

$$\begin{aligned} 7022.5 &= 265a + 11704.17b_1 + 3754.17b_2 \\ -7032 &= -265a - 11715b_1 - 3756b_2 \\ \hline -9.5 &= -10.83b_1 - 1.83b_2 \end{aligned} \qquad (5)$$

Variable b_1 reemplazada en la ecuación (5):

$$9.5 = 10.83\left(\frac{7.5 - 6.83b_2}{1.83}\right) + 1.83b_2 \quad \Rightarrow \qquad b_2 = \frac{67.19}{73.97} = 0.91$$

b_2 reemplazada en la ecuación (4):

$$b_1 = \frac{7.5 - 6.83b_2}{1.83} = 0.7$$

b_1 y b_2 reemplazada en la ecuación (1):

$$a = \frac{159 - 265b_1 - 85b_2}{6} = -17.31$$

Se obtiene así la ecuación de regresión múltiple:

$$\hat{Y} = a + b_1X_1 + b_2X_2 \quad \Rightarrow \qquad \hat{Y} = -17.31 + 0.7X_1 + 0.91X_2$$

EJERCICIOS TEMA 2.1.

1. La cantidad de leche producida por una vaca decrece después de que esta da a luz. Un veterinario desea expresar esta relación por medio de una ecuación. Los datos reunidos producen los siguientes resultados:

Litros por día	12	11	8	9	8	7
Número de días	10	30	40	50	55	65

Determine la mejor ecuación que se ajusta a los datos dados. Y verifique si la ecuación obtenida se ajusta correctamente.

2. Se desea conocer la relación que pueda existir entre las alturas en una muestra de 12 padres y sus hijos. La siguiente tabla refleja los datos obtenidos (en pulgadas):

Altura del padre	65	63	67	68	62	70	66	68	67	69	71
Altura del hijo	68	66	68	69	66	68	65	71	67	68	70

Determine la mejor ecuación que se ajusta a los datos dados. Y verifique si la ecuación obtenida se ajusta correctamente.

3. Ajustar a una recta los datos de la siguiente tabla tomando:

a. X como variable independiente

b. X como variable dependiente.

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

4. Con los siguientes datos, correspondientes a la producción **X** (miles de unidades) y **Y** el costo de la producción de esas unidades (millones de pesos), se pide:

X (miles de unidades)	2	5	8	10	12	15	17	20
Y (millones de pesos)	4	8	10	11	12	14	15	16

a. Dibuje un diagrama de dispersión de la información suministrada. ¿sugiere la gráfica una asociación lineal?

b Aplique el criterio de los mínimos cuadrados para buscar la mejor recta que se ajuste a los datos graficados en el diagrama de dispersión.

c. Determine el error estándar, el coeficiente de determinación y el coeficiente de correlación lineal.

d. Según este modelo de regresión, ¿cuál seria el costo de producción para 25.000 unidades?

5. Una compañía de ahorro y crédito, desea saber cómo son afectadas las ventas de viviendas por diferentes tasas de interés. Durante ocho meses se recopiló la información y se obtuvo el siguiente resultado:

Tasa de interés (%)	7	6.5	5.5	6	8	8.5	6	6.5
Ventas de viviendas	23	38	45	36	16	18	39	41

a. Estimar las ventas en función de la tasa de interés.
b. ¿Cuántas viviendas se pueden vender si el interés es del 7.5%?
c. Determinar el error estándar del estimado.
d. ¿Es confiable el modelo?
e. Calcule el tipo de asociación entre las variables.

6. Una empresa transportadora de frutas, está interesada en precisar la relación que existe entre la distancia a la cual se transporta una carga de fruta, la temperatura a la cual se mantiene y el porcentaje del despacho que se daña al llegar a su destino. Se realizó un muestreo para ocho destinos diferentes y estos son los resultados:

Distancia (decenas de km)	**Temperatura** (ºC)	**Carga dañada** (%)
39	8	7
52	6	6
48	7	7
46	12	10
61	9	9
34	6	4
25	10	3
55	4	4

a. Estimar el porcentaje de carga dañada en función de la distancia y de la temperatura.
b. Estime el porcentaje de fruta que se podría dañar en un viaje de 480 km a una temperatura de 9ºC.
c. Determine el error estimado.
d. ¿Es confiable el modelo?

2.2. NÚMEROS ÍNDICE

Los ***números índice*** son cifras relativas expresadas en términos porcentuales, que sirven para indicar las variaciones que sufre una serie de valores respecto a una de ellas, tomada como punto de referencia y a la cual se le denomina ***base***.

Los números índices no son una medida cuantificable, se trata de un indicador de variación en la variable observada. Son indicadores muy utilizados en el sector económico por ejemplo, la variación en los precios de un producto respecto al año anterior, la cantidad de unidades vendidas de un producto respecto al mes anterior, el costo de producción por unidad de este trimestre comparado con el inmediatamente anterior, etc.

Si se trata de una serie corta, el período base seleccionado será el primer valor de la serie; pero si la serie es extensa se debe seleccionar cono período base aquel que haya sido más estable, es decir, que no presente cambios muy bruscos debido a factores internos y/o externos. Sin embargo, la selección de la serie base dependerá de los análisis que el investigador requiera hacer para sus variables.

Los números índice se pueden construir para una sola observación o para un conjunto de ellas; en el primer caso, se hablará de ***índices simples*** y para un conjunto de datos dados, se hablará de ***índices compuestos***. Estos últimos se clasifican a su vez en ***agregativos*** y ***de promedios***. Los promedios se clasifican en aritméticos, geométricos, medianos, etc., pero en la práctica los más utilizados son los ***aritméticos***.

2.2.1. Construcción de números índice

Para calcular un número índice se toma un valor de la serie como base y se establece un cociente entre el valor de la variable a estudiar y el valor de la variable base. Este cociente debe expresarse en porcentaje, determinando así el número índice respecto a la base definida.

Se pueden obtener bases fijas y bases variables para establecer comparaciones. La ***base fija*** es aquella que representa el mismo período de referencia o de comparación para toda la serie.

$$I_0^t = \frac{X_t}{X_0} \times 100\%$$

Donde:

I_0^t: Índice.

0: Período base.

t: Período que se analiza.

X_t: Precio, cantidad o valor del período que se investiga.

X_0: Precio, cantidad o valor del período considerado como base.

Los índices son de ***base variable*** cuando a cada observación se le divide por el valor de la observación inmediatamente anterior.

$$I_{t-1}^t = \frac{X_t}{X_{t-1}} \times 100\%$$

Donde:

I_{t-1}^t: Índice.

t-1: Período base.

t: Período que se analiza.

X_t: Precio, cantidad o valor del período que se investiga.

X_0: Precio, cantidad o valor del período considerado como base.

2.2.2. Tipos de números índices

El índice de mayor aplicación es el ***índice de precios***, que mide los cambios de precios en uno o más artículos en un período determinado respecto a un período base. El más conocido es el *índice de precios al consumidor*, que mide el cambio de todos los precios respecto a una variedad de artículos que se consumen; este índice se emplea para definir el costo de vida.

Un ***índice de cantidad*** mide la variación de las cantidades de uno o más bienes en un período dado respecto al período base.

El ***índice de valor*** mide los cambios en valor monetario total, es decir, combina los cambios de precios y cantidad para presentar un índice más informativo.

2.2.3. Índices simples

Se construyen para una sola observación y su base puede ser fija o variable. Cuando se trata de medir la variación de un fenómeno observado a

través de una serie de períodos, los índices simples son los más adecuados.

EJEMPLO 2.4.

Un almacén vende cinco referencias diferentes de un artículo determinado. Los datos siguientes indican las ventas de ellos en los meses de febrero y marzo.

Mes	A	B	C	D	E
Febrero	86	395	1308	430	113
Marzo	95	380	1466	469	108

Se desea analizar la variación del artículo con referencia C en el inventario de marzo respecto al mes de febrero.

$$I_{marzo}^{febrero} = \frac{1466}{1308} \times 100\% = 112\%$$

Se considera que el aumento en ventas del artículo con referencia C es del 12% en el mes de marzo respecto al mes de febrero.

Si se quisiera comparar el total de artículos vendidos correspondiente al período de estudio respecto al total de artículos vendidos del período base, se suman todos los elementos correspondientes al período de estudio y se divide entre la suma de los mismos elementos del período base.

$$I_{marzo}^{febrero} = \frac{95+380+1466+469+108}{86+395+1308+430+113} \times 100 = 108\%$$

Se concluye pues, que las ventas del producto aumentaron en marzo en un 8% (108-100) respecto a las ventas del mismo en febrero.

Este último índice calculado en el ejemplo 2.3., se denomina ***índice agregativo*** (o agregado) ***simple*** y se define como:

$$I_0^t = \frac{\sum X_t}{\sum X_{t-1}} \times 100\%$$

2.2.4 Índices compuestos

Se construyen a partir de un grupo de series de tiempo, concernientes a varios artículos. Se trata de examinar el valor no de un artículo, sino de un grupo de ellos respecto a otro considerado de más importancia. Los índices compuestos determinan una condición particular, por ejemplo el costo de vida relativo a transporte, vivienda, alimentación, etc. Se habla entonces de calcular un ***índice agregado ponderado***.

Son muchas las fórmulas para calcular índices ponderados, los más conocidos son los de *Laspeyres, Paashe, Fisher, Keynes, Marshall, Edgeworth, Walsh, Drobisch* y *Sidgwick*. Generalmente en ellos las ponderaciones son las cantidades o precios. Cuando se van a calcular los índices de precios en un grupo de artículos, las ponderaciones son las cantidades, y en el cálculo de los índices de cantidad las ponderaciones son los precios.

El ***índice de Laspeyres de precios*** es la relación que existe al comparar los precios actuales de un grupo de artículos con los precios de esos mismos artículos considerados en el período base, manteniéndose constante como ponderación las cantidades del período base.

$$L^{I_0^t} = \frac{\sum P_t Q_0}{\sum P_0 Q_0} \times 100\%$$

Donde:

L: Índice de Laspeyres.

I_{t-1}^{t}: Índice de precios.

P_t: Precio de los artículos en el período que se investiga.

P_0: Precio de los artículos en el período base.

Q_0: Cantidad de los artículos en el período base.

De igual manera se puede representar así el ***índice de Laspeyres de cantidad***:

$$L^{J_0^t} = \frac{\sum P_0 Q_t}{\sum P_0 Q_0} \times 100\%$$

Donde:

L: Índice de Laspeyres.

J_{t-1}^{t}: Índice de cantidad.

P_0: Precio de los artículos en el período base.

Q_0: Cantidad de los artículos en el período base.

Q_t: Cantidad de los artículos en el período que se investiga.

El ***índice de precios de Paashe*** es la relación que existe entre los precios actuales de un grupo de artículos, con los precios de ellos en el período base, manteniéndose constante las ponderaciones que corresponden a las cantidades de dichos artículos para el período que se investiga.

$$P^{I_0^t} = \frac{\sum P_t Q_t}{\sum P_0 Q_t} \times 100\%$$

Donde:

P: Índice de Paashe.

I_{t-1}^t: Índice de precios.

P_t: Precio de los artículos en el período que se investiga.

P_0: Precio de los artículos en el período base.

Q_t: Cantidad de los artículos en el período que se investiga.

Para indicar las variaciones en las cantidades, el ***índice de cantidad de Paashe*** señala:

$$P^{J_0^t} = \frac{\sum P_t Q_t}{\sum P_t Q_0} \times 100\%$$

Donde:

P: Índice de Paashe.

J_{t-1}^t: Índice de cantidad.

P_t: Precio de los artículos en el período que se investiga.

Q_0: Cantidad de los artículos en el período base.

Q_t: Cantidad de los artículos en el período que se investiga.

El ***índice de precios de Fisher*** es un promedio geométrico, que se define como la raíz cuadrada del producto del índice de Laspeyres por el de Paashe.

$$F^{I_0^t} = \sqrt{\frac{\sum P_t Q_0}{\sum P_0 Q_0} \cdot \frac{\sum P_t Q_t}{\sum P_0 Q_t}} \times 100\%$$

Donde:

F : Índice de Fisher.

I_{t-1}^{t} : Índice de precios.

P_0 : Precio de los artículos en el período base.

P_t : Precio de los artículos en el período que se investiga.

Q_0 : Cantidad de los artículos en el período base.

Q_t : Cantidad de los artículos en el período que se investiga.

Así mismo, se tiene el ***índice de cantidad de Fisher***:

$$F^{J_0^t} = \sqrt{\frac{\sum P_0 Q_t}{\sum P_0 Q_0} \cdot \frac{\sum P_t Q_t}{\sum P_t Q_0}} \times 100$$

Donde:

F : Índice de Fisher.

J_{t-1}^{t} : Índice de precios.

P_0 : Precio de los artículos en el período base.

P_t : Precio de los artículos en el período que se investiga.

Q_0 : Cantidad de los artículos en el período base.

Q_t : Cantidad de los artículos en el período que se investiga.

EJEMPLO 2.5.

Una farmacia reporta la siguiente tabla referente a los precios (en cientos de miles de pesos) y cantidades vendidas (por empaque) de cinco fármacos comunes en los dos últimos años. Calcular los índices de precios y de cantidades por los métodos de Laspeyres, Paashe y Fisher.

Tabla 2.4.

Precios y cantidades vendidas en una farmacia en 2003 y 2004

ARTÍCULO	2003		2004	
	Precio	Cantidad	Precio	Cantidad
A	30	20	25	32
B	18	10	38	5
C	45	12	47	15
D	26	7	40	3
E	35	11	36	12

Para el cálculo de cada índice, se deben determinar todos los valores que interviene en ellos, en la siguiente tabla se resumen todos los cálculos:

Artículo	P_{2003}	Q_{2003}	P_{2004}	Q_{2004}	$P_{2003}.Q_{2003}$	$P_{2004}.Q_{2004}$	$P_{2004}.Q_{2003}$	$P_{2003}.Q_{2004}$
A	30	20	25	32	600	800	500	960
B	18	10	38	5	180	190	380	90
C	45	12	47	15	540	705	564	675
D	26	7	40	3	182	120	280	78
E	35	11	36	12	385	432	396	420
TOTAL					**1887**	**2247**	**2120**	**2223**

Cálculo de índices de precios:

$$L^{I_{2003}^{2004}} = \frac{\sum P_{2004}Q_{2003}}{\sum P_{2003}Q_{2003}} \times 100\% = \frac{2120}{1887} \times 100\% = 112.35\%$$

$$P^{I_{2003}^{2004}} = \frac{\sum P_{2004}Q_{2004}}{\sum P_{2003}Q_{2004}} \times 100\% = \frac{2247}{2223} \times 100\% = 101.08\%$$

$$F^{I_{2003}^{2004}} = \sqrt{\frac{\sum P_{2004}Q_{2003}}{\sum P_{2003}Q_{2003}} \cdot \frac{\sum P_{2004}Q_{2004}}{\sum P_{2003}Q_{2004}}} \times 100\% = \sqrt{\frac{2120}{1887} \cdot \frac{2247}{2223}} \times 100\% = 106.57\%$$

Interpretación: los precios de los productos A, B, C, D y E de la farmacia aumentaron en un 2.35%, 1.08% y 6.75% según le método de Laspeyres, Paashe y Fisher, respectivamente, durante el año 2004 respecto al 2003.

Cálculo de índices de cantidad:

$$L^{J_{2003}^{2004}} = \frac{\sum P_{2003}Q_{2004}}{\sum P_{2003}Q_{2003}} \times 100\% = \frac{2223}{1887} \times 100\% = 117.81\%$$

$$P^{J_{2003}^{2004}} = \frac{\sum P_{2004}Q_{2004}}{\sum P_{2004}Q_{2003}} \times 100\% = \frac{2247}{2120} \times 100\% = 106\%$$

$$F^{J_{2003}^{2004}} = \sqrt{\frac{\sum P_{2003}Q_{2004}}{\sum P_{2003}Q_{2003}} \cdot \frac{\sum P_{2004}Q_{2004}}{\sum P_{2004}Q_{2003}}} \times 100\% = \sqrt{\frac{2223}{1887} \cdot \frac{2247}{2120}} \times 100\% = 111.74\%$$

Interpretación: las cantidades vendidas de los productos A, B, C, D y E de la farmacia aumentaron en un 17.81%, 6% y 11.74% según le método de Laspeyres, Paashe y Fisher, respectivamente, durante el año 2004 respecto al 2003.

2.2.5 Usos de los números índices

Los números índices tienen especial importancia en la vida económica de un país, es común escuchar términos como índice de precios al consumidor (IPC), índice de pérdida de poder adquisitivo, índice de importación o exportación, etc. A continuación se ampliará un poco sobre los más importantes números índices.

- ***Calculo del salario y del ingreso***

$$\text{Salario Real} \quad \Rightarrow \quad \text{SR} = \frac{\text{Salario nominal (\$)}}{IPC} \times 100$$

$$\text{Ingreso Real} \quad \Rightarrow \quad \text{IR} = \frac{\text{Ingreso nominal (\$)}}{IPC} \times 100$$

EJEMPLO 2.6.

Un empleado ganaba en diciembre de 2004 $780.000 y en el mes de junio de 2005, aumentaron su salario en $110.000 más. Los IPC para los mismos meses y años fueron: 2532.4 y 3105.2 respectivamente. Se quiere saber si con el reajuste que le hicieron su salario mejoró con relación al que tenía anteriormente.

Se calcula primero el IPC de cada año respecto al 2004.

$$IPC_{2004}^{2004} = \frac{2532.4}{2532.4} \times 100 = 100 \qquad IPC_{2004}^{2005} = \frac{3105.2}{2532.4} \times 100 = 122.6$$

Esto quiere decir que los artículos de primera necesidad aumentaron en un 22.6% para el período diciembre de 2004 y junio de 2005. De manera que debe haber un porcentaje igual o mayor de incremento en el salario nominal para que las condiciones económicas sean iguales o mejores para el empleado.

El salario real para junio de 2005 es:

$$\text{SR} = \frac{890.000}{122.6} \times 100 = 725938$$

Esto quiere decir que el empleado sólo está recibiendo el equivalente a $725.938 de los $780.000 que recibía. Aunque gane más salario, el aumento es injusto. Su aumento debería de ser mínimo el 22.6% de lo que ganaba en diciembre de 2004, es decir: $176.280 más para un salario de $956.280.

- ***Poder de compra o poder adquisitivo o valor del dinero***

$$\text{Poder de compra} \quad \Rightarrow \quad \text{PA} = \frac{1}{\text{IPC}} \times 100$$

$$\text{Índice de poder adquisitivo} \quad \Rightarrow \quad \text{IPA} = \frac{\text{I}_0}{\text{I}_t} \times 100$$

Donde:

I_0: Índice de precios al consumidor, considerado como período de referencia.

I_t: Índice de precios al consumidor, considerado como período que se investiga.

EJEMPLO 2.7.

Determinar el poder de compra y el índice de poder de compra para junio de 2005 respecto a diciembre de 2004.

$$\text{PA} = \frac{1}{122.6} \times 100 = 0.8156$$

Esto quiere decir que un peso en diciembre de 2004 equivale a 82 centavos en

junio de 2005. Su valor se ha reducido durante ese período en 18 centavos.

$$\text{IPA} = \frac{2532.4}{3105.2} \times 100 = 81.56\%$$

- ***Porcentaje de desvalorización***

$$\%\text{ de desvalorización} = 100\left[1 - \frac{I_0}{I_t}\right]$$

EJEMPLO 2.8.

Determinar el porcentaje de desvalorización para los datos del ejemplo 2.7.

$$\%\text{ de desvalorización} = 100\left[1 - \frac{2532.4}{3105.2}\right] = 18.45\%$$

De diciembre de 2004 a junio de 2005, la moneda ha perdido un 18.45% de su poder de compra. Esto quiere decir que en junio de 2005 se necesita más dinero para comprar el mismo artículo en diciembre de 2004.

- ***Porcentaje de variación y de devaluación***

$$\%\text{ de variación} = 100\left[\frac{T_1}{T_0} - 1\right]$$

$$\%\text{ de devaluación} = 100\left[1 - \frac{T_0}{T_1}\right]$$

Donde:
T_0: Valor de la moneda de referencia.
T_t: Valor de la moneda que se quiere cambiar.

La devaluación es entendida como la pérdida de valor de una moneda en relación a las monedas extranjeras.

EJERCICIOS TEMA 2.2.

1. Para los siguientes datos, calcular los índices de precios y de cantidades por los métodos de Laspeyres, Paashe y Fisher.

ARTÍCULO	2002		2004	
	Precio	Cantidad	Precio	Cantidad
A	320	15	320	26
B	140	18	200	35
C	80	35	600	54
D	560	28	520	25
E	1120	14	1200	18

2. Una marroquinería produce bolsos para dama en tres líneas diferentes. Los datos sobre producción de bolsos y de tiempo por empleado (en horas) ocupados en la empresa durante el período 2001-2003 son los siguientes:

ARTÍCULO	Producción en miles de docenas			Horas-hombre por docena		
	2001	2002	2003	2001	2002	2003
A	5	8.3	9.4	6.3	6.2	6.3
B	7.5	10.2	11.8	4.7	4.9	4.6
C	5.5	5.6	4.2	3.2	3.2	3.2

a. Calcule un índice de cantidad para el año 2003, empleando como base el año 2001 y utilizando como ponderación los datos sobre horas-hombre empleadas por docena de bolsos en 2001.

b. Calcule un índice de producción de 2002 con base en 2001.

3. Un empleado ganaba $722.000 mensuales en 2003. Hoy día recibe $912.000 mensuales, con lo cual mejora su ingreso real en un 37%. Si el actual índice de precios es 560, ¿cuál era el de 2003?

4. Tomadas las cosechas de ciertos productos agrícolas (en cientos de toneladas), determinar el índice agregativo simple para 2004 con base en 2002.

Productos	2002	2004
A	11.158	13.044
B	1.196	1.357
C	1.111	1.326
D	1.460	1.840
E	859	997
F	1.106	870
G	41	59
H	6.686	7.978
I	204	202

5. Con los siguientes datos:

AÑOS	**SALARIOS** (miles de millones de pesos)	**OBREROS** Nº	**IPC** 1991=100
1998	18.0	320	140
1999	20.6	380	148
2000	23.0	400	152
2001	38.0	700	160
2002	51.0	1.000	166
2003	58.0	1.050	168
2004	60.0	1.100	170

Se pide

a. Salarios reales con respecto a 1998.
b. Salarios nominales por obrero.
c. Índices de los salarios reales con base 1998.
d. Índices de los salarios nominales con base 1998.
e. Salarios reales por obrero, con base 1998.
f. Índices de salarios reales por obrero, con base 1998.

INFORMACIÓN DE RETORNO DE LA UNIDAD

EJERCICIOS TEMA 1.1.

1. **a.** $\bar{x} = 10525$
b. $Me = 8775$
c. $Mo = 5000$
d. La media.
e. 35.000

2. **a.** $\bar{x} = 5{,}92$ $Me = 6$ $Mo = 6$
b. $\bar{x} = 6{,}95$ $Me = 7$ $Mo = 7$
c. $\bar{x} = 72{,}2$ $Me = 71{,}2$ $Mo = 64$

3. Salario promedio turno del día: $424.000
Salario promedio turno de la noche: $452.000

4. 84,5

5. 4,45

6. 6,102 horas

7. 68,57 km/h

8. Factor de crecimiento promedio: 1,25

9. $Q_1 = 58{,}4$ $Q_2 = 71{,}2$ $Q_3 = 85{,}9$

10. **a.** $P_5 = 33{,}4$
b. $P_{15} = 33{,}8$
c. $P_{95} = 36{,}1$
d. $P_{25} = 34{,}15$
e. $P_{50} = 34{,}5$
f. $P_{10} = 33{,}7$
g. $P_{75} = 35{,}1$
h. $P_{30} = 34{,}2$

EJERCICIOS TEMA 1.2.

1. 45 minutos

2. $Q_D = 27{,}5 \qquad Q_{D2} = 13{,}75$

3.
a. $R = 8 \quad s^2 = 7{,}81 \qquad s = 2{,}79 \qquad DM = 2{,}5 \qquad CV = 72{,}85\%$
b. $R = 4{,}12 \qquad s^2 = 2{,}16 \qquad s = 1{,}47 \qquad DM = 1{,}27 \qquad CV = 46{,}59\%$
c. $R = 3 \quad s^2 = 1{,}33 \qquad s = 1{,}15 \qquad DM = 0{,}6 \qquad CV = 84{,}46\%$
d. $R = 4{,}12 \qquad s^2 = 2{,}16 \qquad s = 1{,}47 \qquad DM = 1{,}14 \qquad CV = 23{,}9\%$

4. $s^2 = 306{,}76 \qquad s = 17{,}51$

5. $s^2 = 137{,}68 \qquad s = 11{,}73 \qquad DM = 9{,}28$

6. Johan es el mejor

7.
a. El tubo B tiene mayor dispersión absoluta.
b. El tubo A tiene mayor dispersión relativa.
c. El tubo B tiene menor posición relativa.

8.
$Z_1 = -5{,}75 \qquad Z_2 = -4{,}6 \qquad Z_4 = -2{,}3 \qquad Z_5 = -1{,}15$
$Z_6 = 0 \qquad Z_9 = 3{,}44 \qquad Z_{10} = 4{,}6 \qquad Z_{11} = 5{,}75$

9.
a. $Z = -2{,}71$
b. $Z = 0{,}86$
c. $Z = -1$

10.
a. Estatura: $Z = 0{,}74$ Peso: $Z = 0{,}15$
b. El peso tiene mayor dispersión absoluta.
c. El peso tiene menor dispersión relativa.

EJERCICIOS TEMA 1.3.

1.
a. Asimétrica positiva.
b. Asimétrica positiva.
c. Asimétrica negativa.

2. $Mo = 9$

3. $As = 0{,}13$

EJERCICIOS TEMA 2.1.

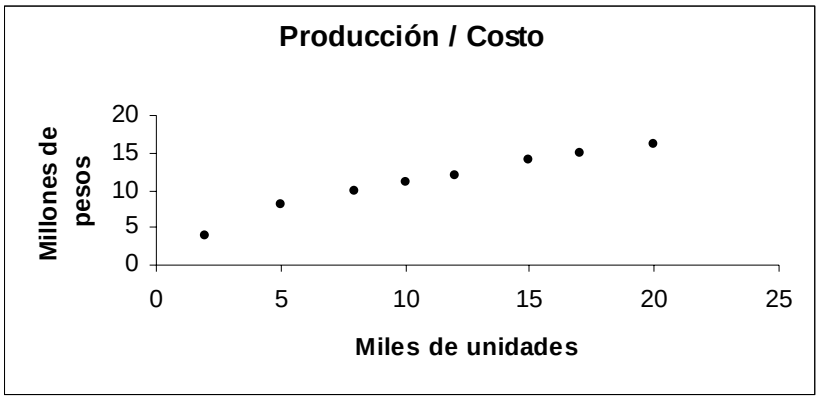
1. $\hat{Y} = 127{,}65 - 9{,}38X$ Correlación excelente.

2. $\hat{Y} = 39{,}27 + 0{,}43X$ Correlación regular.

3. **a.** $\hat{Y} = -0{,}33 + 0{,}71X$

b. $\hat{X} = 1 + 1{,}29Y$

4. **a.**



b. $\hat{Y} = 4{,}18 + 0{,}64X$

c. $Se = 0{,}72 \qquad R^2 = 0{,}96 \qquad r = 0{,}98$

d. Para producir 25.000 unidades se estima un costo de 20.1 millones de pesos.

5. **a.** $\hat{Y} = 99{,}5 - 10X$

b. 25 viviendas

c. $Se = 58{,}07$

d. El modelo es confiable, explica el 84,8% de la información.

e. La tasa de interés está relacionada en un 92,1% con las ventas de viviendas.

BIBLIOGRAFÍA DE LA UNIDAD

BEJARANO BARRERA, Hernán (1995). *Estadística Descriptiva.* Santa fe de Bogotá: UNISUR.

CHRISTENSEN, Howard B. (1999). *Estadística Paso a Paso.* México: Editorial Trillas.

MARTÍNEZ BENCARDINO, Ciro (2004). *Estadística Básica Aplicada.* Santa fe de Bogotá: ECOE Ediciones.

MARTÍNEZ BENCARDINO, Ciro (2003). *Estadística y Muestreo.* Santa fe de Bogotá: ECOE Ediciones.

MILTON, J. Susan (1999). *Estadística para biología y ciencias de la salud.* Madrid: McGraw Hill — Interamericana.

PORTUS GOVINDEN, Lincoyán (2001). *Introducción a la Estadística.* Segunda edición. Santa fe de Bogotá. McGraw Hill.

PORTILLA CHIMAL, Enrique (1980). *Estadística, Primer Curso.* Bogotá: Nueva Editorial Interamericana.

SPIEGEL, Murria R. (1991). *Estadística. Serie de compendios Schaum*. México: McGraw Hill.

SMITH, A. Stanley. (1992). *Curso de Estadística Elemental para las ciencias aplicadas.* Primera edición. Santa fe de Bogotá. Editorial Addison – Wesley Iberoamericana.

TRIOLA, MARIO F. (2004). *Probabilidad y Estadística.* Novena edición. México. Pearson Educación.

http://www.aulafacil.com/CursoEstadistica/CursoEstadistica.htm

http://www.elosiodelosantos.com/regresionlineal.html

http://www.universidadabierta.edu.mx/SerEst/MAP/METODOS%20CUANTITATIV

OS/Pye/tema_12.htm

http://server2.southlink.com.ar/vap/MEDIDAS.htm

http://cosmech.tripod.com/Estadistica/medidas1.htm

http://eris.unalmed.edu.co/~cescobar/Bioestadistica/bioestadistica.htm

http://ftp.medprev.uma.es/libro/node42.htm

http://www.eumed.net/cursecon/medir/

Anexo A

Sumatorias y Productorias

A lo largo de los trabajos en estadística se encontrarán muchas veces con la suma de un gran número de términos. Con el fin de utilizar el lenguaje algebraico que permita realizar simplificaciones, se requiere el uso del símbolo ***sumatoria*** el cual representa la operación de adición algebraica sobre una determinada cantidad de elementos numéricos.

Considere las siguientes cantidades: 7, 9, 15, 14, 8, 3, 5, 16. Estos términos pueden sumarse de la forma más común:

$$S = 7 + 9 + 15 + 14 + 8 + 3 + 5 + 16 = 77$$

Si cada uno de estos términos numéricos es representado por X_i, donde el subíndice i indica la cantidad relativa de elementos considerados, se puede expresar la anterior operación como:

$$S = X_1 + X_2 + X_3 + ... + X_8 = 77$$

Ahora, esta operación puede expresarse de la siguiente forma:

$$\sum_{i=1}^{8} X_i = X_1 + X_2 + X_3 + ... + X_8 = 77$$

El símbolo griego sigma ($\sum$), que se lee sumatoria representa, para el caso más general, la suma de n términos cualquiera. Se tiene entonces que:

$$\sum_{i=1}^{n} X_i$$

es la suma de n términos, donde n es el límite superior de la sumatoria; i es el elemento genérico de la sumatoria; i = 1 es el límite inferior de la sumatoria.

La sumatoria tiene algunas propiedades importantes que deben tenerse en cuenta:

- La sumatoria de una constante C de 1 a n es igual a n veces la constante.

$$\sum_{i=1}^{n} C = C + C + ... + C = nC$$

- La sumatoria del producto de una constante por una variable es igual a la constante por la sumatoria de la variable.

$$\sum_{i=1}^{n} CX_i = CX_1 + CX_2 + ... + CX_n = C\sum_{i=1}^{n} X_i$$

- La sumatoria de los valores de una variable más una constante es igual a la sumatoria de la variable más *n* veces la constante.

$$\sum_{i=1}^{n} (X_i + C) = (X_1 + C) + (X_2 + C) + ... + (X_n + C) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} C$$

- La sumatoria de una constante con límite inferior diferente a 1 es:

$$\sum_{i=m}^{n} C = (n - m + 1)C$$

EJEMPLO A.1.

Dados los valores: $X_1 = 2$, $X_2 = 6$, $X_3 = 1$, $X_4 = 0$, $X_5 = 7$, $X_6 = 7$, $X_7 = 8$, hallar:

a. $\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 2 + 6 + 1 + 0 + 7 = 16$

b. $\sum_{i=3}^{7} X_i = X_3 + X_4 + X_5 + X_6 + X_7 = 1 + 0 + 7 + 7 + 8 = 23$

c. $\sum_{i=1}^{5} X_i^{\,2} = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 = 2^2 + 6^2 + 1^2 + 0^2 + 7^2 = 90$

d. $\sum_{i=3}^{7} 7X_i = 7\sum_{i=3}^{7} X_i = (7)(X_3 + X_4 + X_5 + X_6 + X_7) = (7)(1 + 0 + 7 + 7 + 8) = 161$

e. $\sum_{i=2}^{6} 5 = (6 - 2 + 1)(5) = 25$

f. $\sum_{i=2}^{5} (X_i - 2) = \sum_{i=2}^{5} X_i - \sum_{i=2}^{5} 2 = \left[X_2 + X_3 + X_4 + X_5\right] - \left[(5 - 2 + 1)(2)\right] = 14 - 8 = 6$

La ***productoria*** indica el producto de determinada cantidad de elementos numéricos. Se utiliza la letra griega pi (Π), que se lee *producto de*. De este modo,

el producto de n términos cualquiera está dado por:

$$\prod_{i=1}^{n} X_i = X_1 \cdot X_2 \cdot ... \cdot X_n$$

donde n es el límite superior de la productoria; i es el elemento genérico; i = 1 es el límite inferior. Este símbolo es usado para calcular la media geométrica.

Igual que la sumatoria, la productoria tiene propiedades importantes:

- La productoria de una constante C es igual a una potencia, donde la base es la constante y el exponente es el límite superior del producto.

$$\prod_{i=1}^{n} C = C \cdot C \cdot ... \cdot C = C^n$$

- El producto de una constante C por una variable es igual a la constante elevada al límite superior por la productoria de la variable.

$$\prod_{i=1}^{n} CX_i = C^n \left[\prod_{i=1}^{n} X_i \right]$$

- La productoria de una constante con límite inferior diferente a 1 es:

$$\prod_{i=m}^{n} C = C^{n-m+1}$$

EJEMPLO A.2.

Dados los valores: X_1 = 2, X_2 = 6, X_3 = 1, X_4 = 0, X_5 = 5, X_6 = 3, hallar:

a. $\prod_{i=2}^{5} X_i = 6 \times 1 \times 0 \times 5 = 0$

b. $\prod_{i=1}^{4} 2 = 2 \times 2 \times 2 \times 2 = 2^4 = 32$

c. $\prod_{i=1}^{6} 3X_i = 3^6 \left[\prod_{i=1}^{6} X_i \right] = 729 \times \left[2 \times 6 \times 1 \times 0 \times 5 \times 3 \right] = 0$

EJERCICIOS ANEXO A

1. Si $X_1 = 2$, $X_2 = 4$, $X_3 = 5$, $X_4 = 6$ y $X_5 = 1$, calcular:

a. $\sum_{i=1}^{5} X$ **b.** $\sum_{i=1}^{3} X^2$ **c.** $\sum_{i=2}^{4} 5X$

d. $\sum_{i=1}^{5} (X+2)$ **e.** $\sum_{i=3}^{5} (X-4)^2$ **f.** $\sum_{i=1}^{4} X^2 - 4$

2. Complete el siguiente cuadro.

OPERADOR	DESARROLLO	RESULTADO
$\sum_{i=1}^{10} i$		
$\sum_{i=1}^{5} i^2$		
$\sum_{i=1}^{5} 10$		
$\sum_{i=1}^{5} (2i+1)$		
$\sum_{x=1}^{3} (2x^2+2x+1)$		
$\sum_{x=2}^{5} (2x-1)^2$		
$\sum_{x=4}^{7} (x^2-2)$		
$\sum_{x=1}^{4} x^x$		
$\left[\sum_{x=1}^{5} x\right]^2$		
$\prod_{i=1}^{4} (2i-4)$		
$\prod_{i=1}^{5} i$		
$\prod_{i=1}^{5} 8$		

$\prod_{i=1}^{4} 3i$		
$\frac{2\sum_{n=2}^{5} 2n+1}{3\prod_{n=2}^{5} 2n-1}$		

INFORMACIÓN DE RETORNO DEL ANEXO A

1. **a.** 18 **b.** 45 **c.** 64
d. 28 **e.** 14 **f.** 77

2.

OPERADOR	RESULTADO	OPERADOR	RESULTADO
$\sum_{i=1}^{10} i$	55	$\sum_{x=1}^{4} x^x$	288
$\sum_{i=1}^{5} i^2$	55	$\left[\sum_{x=1}^{5} x\right]^2$	225
$\sum_{i=1}^{5} 10$	50	$\prod_{i=1}^{4} (2i-4)$	0
$\sum_{i=1}^{5} (2i+1)$	35	$\prod_{i=1}^{5} i$	120
$\sum_{x=1}^{3} (2x^2+2x+1)$	43	$\prod_{i=1}^{5} 8$	32768
$\sum_{x=2}^{5} (2x-1)^2$	164	$\prod_{i=1}^{4} 3i$	1944
$\sum_{x=4}^{7} (x^2-2)$	118	$\frac{2\sum_{n=2}^{5} 2n+1}{3\prod_{n=2}^{5} 2n-1}$	$\frac{57}{5759}$