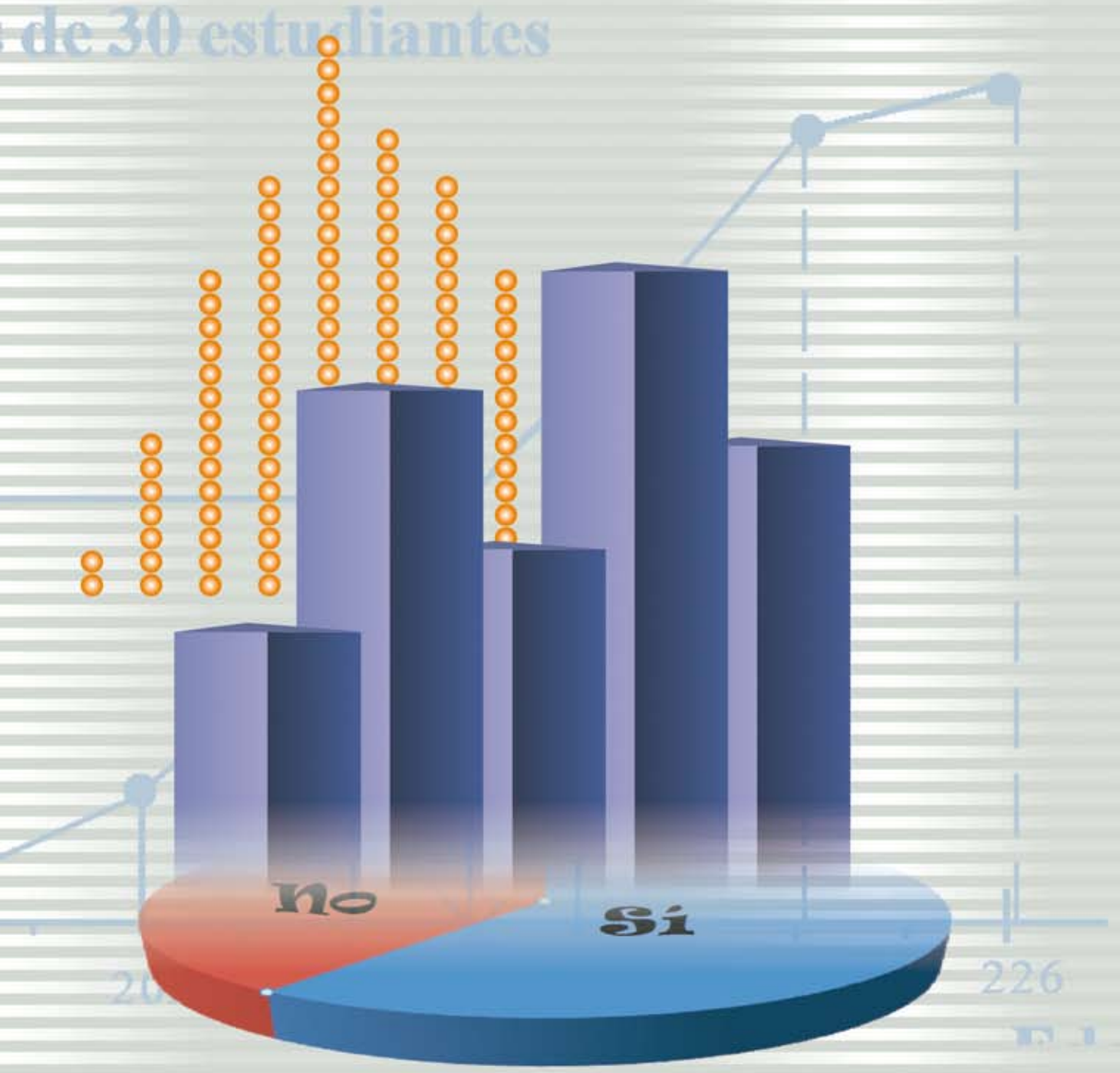
Estadística

EXPLORACIÓN DE DATOS



José Alfredo Juárez Duarte
Arturo Ylé Martínez • Armando Flórez Arco
Santiago Inzunsa Cázares

DGEP

Dirección General de Escuelas Preparatorias

DIRECTORIO

Dr. Víctor Antonio Corrales Burgueño
Rector

DR. José Alfredo Leal Orduño
Secretario General

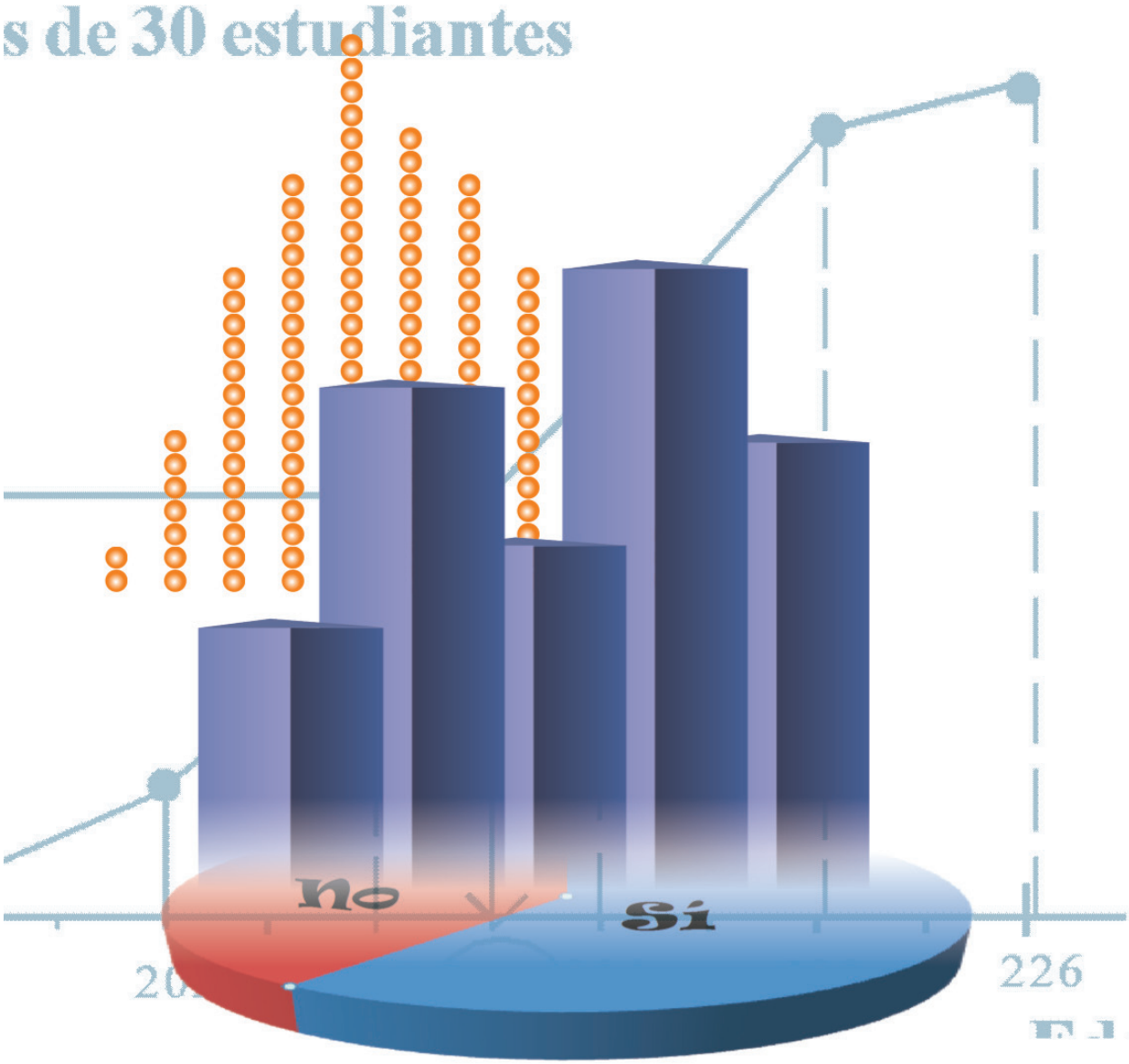
LAE y MA Manuel de Jesús Lara Salazar
Secretario de Administración y Finanzas

Q.F.B. Ofelia Loaiza Flores
Director de Servicios Escolares

Dr. Armando Flórez Arco
Director de DGEP

Estadística

EXPLORACIÓN DE DATOS



José Alfredo Juárez Duarte
Arturo Ylé Martínez
Armando Flórez Arco
Santiago Inzunsa Cázares

Estadística
EXPLORACIÓN DE DATOS

José Alfredo Juárez Duarte
Arturo Ylé MArtínez
Armando Flórez Arco
Santiago Inzunsa Cázarez

Primera edición, junio de 2010
Segunda edición, julio de 2011
Tercera edición, agosto de 2012

Diseño editorial y de portada: Leticia Sánchez Lara

Servicios Editoriales Once Ríos
Río Usumacinta 821 Col. Industrial Bravo, Culiacán, Sin.
Tel-fax: 01(667) 712-2950

Esta edición consta de 9 000 ejemplares

Registro en trámite

Impreso en México
Printed in México

Presentación

Este libro, está destinado a los alumnos de bachillerato de la Universidad Autónoma de Sinaloa, que cursan el quinto semestre del plan 2009. Su principal característica es su focalización en la exploración de datos. De esta manera, se intenta atender una problemática que diversas investigaciones a nivel internacional han puesto de manifiesto: en los cursos tradicionales de introducción a la estadística, los alumnos aprenden a realizar una gran variedad de cálculos, pero, no muestran la más mínima capacidad de razonar estadísticamente, a pesar de aprobar tales cursos con altas calificaciones. Ante esta realidad, el presente trabajo atiende, en la medida de lo posible, algunas de las recomendaciones que organismos involucrados en educación estadística, han planteado para mejorar la enseñanza y aprendizaje de la estadística. Entre estas recomendaciones destacan:

- Enfatizar pensamiento estadístico.
- Más datos y conceptos, menos teoría y pocas reglas tipo recetas.
- Promover actividades de aprendizaje.
- Usar tecnología siempres que sea posible.
- Desarrollar el proceso estadístico como una manera de contestar preguntas cuyas respuestas presentan variabilidad.
- Obtener o generar datos propios.
- Empezar el análisis, graficando los datos.
- Interpretar resúmenes y representaciones gráficas (contestar preguntas/revisar condiciones).

El objetivo que se busca con la enseñanza de la estadística, no es que el alumno sea capaz de realizar cálculos o gráficos tediosos, que pueden ser generados con las nuevas tecnologías, sino que adquieran una cultura estadística que se manifieste en su capacidad para:

- Interpretar resultados en contexto.
- Leer y criticar noticias, reportajes y artículos publicados en distintos medios de comunicación, que incluyan información estadística.
- Comunicar resultados del análisis estadístico.

En la búsqueda de un acercamiento a estas propuestas, en este libro se hace énfasis en el uso de recursos gráficos que, a pesar de tener ya varios años en la literatura, apenas si se les ha dado importancia. Entre ellos tenemos: el gráfico de puntos, el gráfico de caja y el gráfico de tallo y hoja.

La utilización de estos recursos junto con los más tradicionales como el histograma, permite cambiar el foco del trabajo estadístico escolar, que consiste simplemente en la búsqueda de un número como respuesta a un problema, a la búsqueda de un patrón o tendencia de los datos. En este sentido, cobran gran relevancia dos conceptos estadísticos: variabilidad y distribución. Algunas investigaciones plantean que focalizar en variabilidad y distribución, es una alternativa a una enseñanza de la estadística que enfatiza procedimientos de cálculo que pronto se olvidan. En este libro, se intenta convertir a la variabilidad y distribución, en conceptos integradores de todo el curso de estadística. De esta manera, se atiende una de las funciones del nuevo enfoque basado en competencias que consiste en orientar la enseñanza desde una perspectiva transversal e integradora.

El presente libro, es producto de muchos otros. Cada uno de los libros o materiales citados en la bibliografía aportaron algo, desde una idea vaga, hasta una propuesta que sólo requirió de ajuste.

Reiteramos que un libro de texto, es un instrumento de enseñanza para el profesor y un instrumento de aprendizaje para el alumno. El libro de texto debe fomentar el trabajo indpendiente de los alumnos. Para este último objetivo, nos permitimos citar el procedimiento de estudio recomendado por por Robert Johnson y Patricia Kuby:

> *«Pida a su instructor un programa que indique el material (páginas del texto) que será cubierto en cada clase y luego dedique 10 minutos antes de ésta a leer el material de ese día. No tome notas ni subraye nada. Este es un proceso de «calentamiento»; sólo lea «superficialmente» como avance de la clase. Cuando llegue a ésta asistirá a un «inicio ondulante», poque ya conoce los términos y conceptos clave. Cuando estos conceptos se analicen en el aula, usted estará escuchándolos por segunda ocasión, de modo que «oirá» más en clase. A su vez, lo anterior reducirá el tiempo de estudio necesario para aprender bien el material (la lectura antes de clase es sólo superficial; no necesariamente se trata de estudiar en ese momento el material).*
>
> *Tan pronto como pueda después de clase, lea nuevamente el material, a fondo esta vez. haga notas, destaque los puntos importantes y complete los ejercicios asignados. Resuelva las actividades a media que avance en sus clases.*
>
> *Forme un grupo de estudio con dos o tres condiscípulos. Los amigos no siempre son los mejores compañeros para estudiar; busque compañeros de estudio que tengan sus mismos objetivos para este curso y muestren la misma determinación para alcanzarlos. Establezca una sesión semanal para estudiar juntos. Tal vez, deba agregar una sesión de estudio una semana antes de algún examen importante».*

Cualquier comentario o sugerencia para mejorar esta propuesta, que agradecemos de antemano, favor de enviarlo a jjuarez@uas.uasnet.mx. Deseamos a profesores y estudiantes, mucho éxito en su estudio de la estadística, disciplina que sin duda alguna, traerá grandes beneficios en su formación cultural para comprender y retener alguna medida de nuestro mundo cada vez más complejo y cambiante.

ATENTAMENTE
Culiacán Rosales, Sinaloa, julio de 2012
Los autores

Contenido

Introducción a la Estadística

1

UNIDAD

Lección 1.1 ¿Qué es la estadística?

Definición y conceptos básicos

Objetivos: Empezar a apreciar a los datos como números con un contexto.
Aprender a reconocer las variables cualitativas y cuantitativas.
Empezar a apreciar la importancia de la estadística para tratar con situaciones que muestran variabilidad.

Actividad 1

Qué hacer

1) Consulta las **páginas 4 a 6** y al finalizar tu estudio contesta lo indicado:

1.a) Explica con tus propias palabras el significado de:

a) Pregunta estadística
b) Pregunta determinística
c) Variabilidad
d) Variable
e) Unidad observacional o individuo
f) Dato

1.b) En un estudio, se recolectaron datos relativos a las variables siguientes. Clasifica correctamente cada una de las variables en cualitativa o cuantitativa.

a) Edad_____
b) Nivel de escolaridad____
c) Género_____
d) Calificación promedio__
e) Tiempo diario dedicado a estudiar_____
f) Distancia de la casa a la escuela_____
g) Colonia de residencia_____
d) Número de integrantes en la familia._____.

3) Consulta las **páginas 7 a 10** y al finalizar tu estudio explica con tus propias palabras el significado de:

a) Estadígrafo
b) Parámetro
c) Estadística

4) Resuelve el ejercicio 1.1

La palabra estadística tiene dos significados:

1. Escrita en plural, es decir «estadísticas», se refiere a un hecho que implica una presentación numérica.

Ejemplo: Estadísticas de la primera división. Torneo de clausura 2008.
Los 10 mejores equipos.

Equipo	JJ	JG	JE	JP	Puntos
Pachuca	17	11	3	3	36
Toluca	17	10	6	1	36
UNAM	17	8	4	5	28
Monterrey	17	7	5	5	26
Puebla	17	7	5	5	26
UAG	17	6	7	4	25
Cd. Juárez	17	5	8	4	23
América	17	6	5	6	23
Santos	17	5	7	5	22
Morelia	17	5	7	5	22

JJ: Juegos jugados
JG: Juegos ganados
JE: Juegos perdidos

2) El segundo significado de la palabra estadística se refiere a la «ciencia estadística», es decir a un conjunto de conceptos, técnicas y métodos que nos ayudan a contestar preguntas para las que no hay una respuesta única. La estadística como ciencia, cambia números en información.

Por ejemplo, la estadística proporciona las herramientas para contestar preguntas como las siguientes:

Pregunta estadística y pregunta determinística

En el ejemplo de «estadísticas» de futbol, se observan dos tipos de preguntas:

(1) ¿Cuántos juegos jugó un equipo x ? La respuesta es 17. Esta no es una pregunta estadística, es una pregunta determinística.

(2) ¿Cuántos juegos ganó un equipo x? En este caso la respuesta no es única, es por tanto, una pregunta estadística.

En el contexto escolar

- ¿Qué tipo de música es más popular entre los estudiantes de preparatoria?
- ¿Cuál es la edad de un estudiante típico de tercero de preparatoria?
- ¿Cuáles son los tiempos que utilizan los estudiantes para trasladarse de su casa a la escuela?
- ¿Cuál es la calificación promedio de un estudiante típico de esta preparatoria?
- ¿Cuál es la estatura de un estudiante típico de preparatoria?
- ¿Cuánto tiempo al día pasa un estudiante típico en el chat?

En otros contextos

- ¿Qué partido político prefieren los sinaloenses?
- ¿Cuántos días debe administrarse el medicamento Tempra a un niño con fiebre?
- ¿Cuanto duran los neumáticos para automóvil?
- ¿Cuáles son los rendimientos de maíz de determinada variedad de semilla?
- ¿Cuántos km por cada litro de gasolina recorre un automóvil típico?

¿Qué tienen en común estas preguntas? En todas ellas las respuestas varían al cambiar de individuo; es decir, estamos frente a situaciones cuyas respuestas son inciertas. Para comprobar esta afirmación realiza la siguiente actividad.

Actividad 1.1 a

El grupo debe organizarse para preguntar y registrar las respuestas de todo el grupo a las siguientes preguntas:

¿Qué tipo de música prefieren los estudiantes de preparatoria?

¿Cuánto tiempo (en minutos) utilizan los estudiantes para trasladarse de su casa a la escuela?

Variabilidad, es la propiedad exhibida por los datos que consiste en diferir de individuo a individuo.

Las respuestas que haz registrado son datos. Estos datos ponen de manifiesto el principio fundamental para usar la estadística: la variabilidad.

En general, los datos varían por lo que la variabilidad abunda en la vida diaria. Las personas somos diferentes. De manera natural tenemos estaturas diferentes, aptitudes y habilidades u opiniones diferentes. Los estudiantes de este grupo varían con respecto al género, letra inicial de su nombre, estatura, calificaciones, y en muchas otras características. Entonces, ¿cuál es la respuesta que debe establecerse a preguntas sobre este tipo de características? Consideremos la respuesta a una pregunta delicada:

Si las respuestas a una pregunta fueran todas iguales, por ejemplo, si todos los estudiantes ocuparan el mismo tiempo de traslado de su casa a la escuela, estaríamos frente a una situación llamada determinística, y entonces no necesitaríamos de la estadística.

¿Cuántos días debe administrase el medicamento Tempra a un niño con fiebre?

Esta pregunta ya fué contestada por un grupo de investigadores del corporativo farmacéutico implicado: No debe administrarse por más de cinco días.

Si diferentes personas responden diferente al mismo tratamiento, ¿cómo se llega a esta conclusión? ¿cómo contestar preguntas que tienen muchas respuestas? La respuesta es: utilizando el método estadístico. En las próximas páginas estudiarás aspectos fundamentales de este método. Por el momento podemos ya establecer un primer significado de la palabra estadística.

La estadística, es la ciencia del razonamiento acerca de los datos que presentan variabilidad.

También podemos ya establecer las siguientes definiciones:

Variable: es el aspecto, fenómeno, rasgo o cualidad que se va a estudiar, y que puede tomar dos o más valores posibles.

Cualquier característica de una persona o cosa a la que se le puede asignar un número o una categoría, recibe el nombre de variable.

Individuo o unidad observacional: Es la persona o cosa a quien se le asigna un número o categoría

Ejemplo En las preguntas planteadas anteriormente, aparecen las siguientes variables:

Preferencia musical, edad, tiempo de traslado, calificaciones, estatura, tiempo utilizado en el chat, preferencia electoral, dosis de medicamento, duración de neumáticos, rendimiento de maíz, rendimiento de gasolina.

Actividad 1.1 b

Escribe una posible respuesta para cada una de las variables siguientes:

Preferencia musical_____, edad_____, tiempo de traslado_____, calificación___, estatura_____, tiempo utilizado en el chat_____, preferencia electoral____, dosis de medicamento_____, duración de neumáticos____, rendimiento de maíz_____, rendimiento de gasolina_____, tipo de sangre_____

Debes recordar que cada respuesta es un dato que corresponde a la variable estudiada.

Un dato es el valor de la variable asociada a un individuo.

No dejes que la apariencia de los datos te engañe respecto a su tipo. Las variables cualitativas no siempre son fáciles de reconocer. Algunas veces se presentan como números. Por ejemplo, las colonias de una ciudad, pueden identificarse usando números de códigos postales. La identificación de un estudiante mediante un número de cuenta, también es cualitativa.

En las respuestas que diste en la actividad anterior, se observan cantidades numéricas, palabras o símbolos.

Dato cuantitativo o numérico, es aquel que se obtiene de situaciones que requieren tomar medidas o donde objetos son contados.

Dato cualitativo o categórico, es aquel que se obtiene cuando se registra simplemente una categoría asignada, mediante una palabra o un símbolo.

Variable cuantitativa o numérico, es aquella que produce datos cuantitativos o numéricos.

Variable cualitativa, es aquella que produce datos cualitativos o categóricos.

Actividad 1.1 c

Los individuos para las variables listadas abajo son estudiantes de tu grupo. Para cada variable, indica si es variable cuantitativa o variable cualitativa.

a) Género:

b) Número de letras del primer nombre:

c) Tipo de sangre:

d) Estatura:

e) Deporte preferido:

En la siguiente actividad, utilizaremos una variable cualitativa de fácil estudio. El objetivo es que aprecies aspectos fundamentales del proceso estadístico.

Actividad 1.1 d

Pregunta. ¿Cuál es el género de las personas que pertenecen a tu preparatoria?

a) Contesta:

- ¿Existe una respuesta única?_____
- ¿Cuáles son las respuestas posibles?_____

Con base en tus conocimentos previos, escribe un párrafo explicando qué harías para contestar esta pregunta.

Para contestar preguntas que presentan variabilidad, se siguen los siguientes pasos:

Recolectar datos.

b) En la siguiente tabla, registra el género de las personas integrantes de tu salón de clase. Utiliza la letra F para femenino, y M para masculino.

Número del alumno (a)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Género																											

Número del alumno (a)	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Género																											

Una vez recolectados los datos, se procede a analizarlos. El objetivo del análisis es transformar los datos de tal manera que se destaque información importante.

c) Para analizar estos datos, empieza por realizar lo indicado a continuación:
Construye una tabla con dos columnas. En la primera escribe las posibles respuestas (las posibles respuestas se llaman modalidades); en la segunda columna se hace el conteo, es decir, se escribe el número de datos que aparecen en cada modalidad.

Género	Número de estudiantes
F	
M	
Total	

Actividad 1.1 d (Cont.)

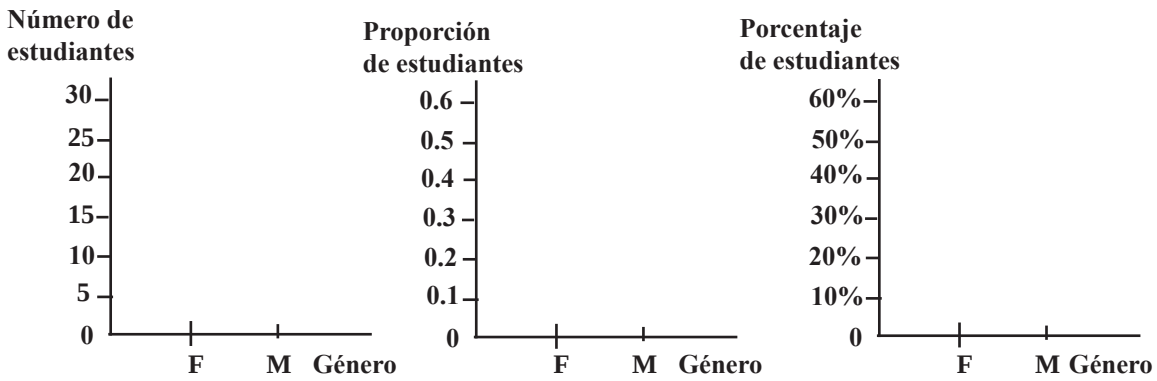
Ahora, puedes calcular las proporciones o porcentajes.

d) Para calcular las proporciones, divide el número de veces que aparece cada modalidad entre el total de datos. Si multiplicas cada proporción por 100, obtienes los porcentajes.

Género	Número de estudiantes	Proporción	Porcentaje
F			
M			
Total			

Parte fundamental del análisis estadístico lo constituyen las gráficas. Una gráfica sencilla es la de barras. Sigue los siguientes pasos para construir una gráfica de barras.

1. Dibuja dos ejes perpendiculares: uno horizontal y el otro vertical. Escribe en el eje horizontal cada una de las modalidades de la variable estudiada (en este caso el género), y en el eje vertical puedes escribir los números correspondientes al conteo, a la proporción o al porcentaje de individuos de cada modalidad. A continuación dibuja para cada modalidad un rectángulo cuya altura coincida con el número de veces, o la proporción, o el porcentaje que aparece cada modalidad. Los rectángulos deben dibujarse separados.



A lo largo del curso estudiarás otros recursos tanto gráficos como numéricos para analizar datos.

Finalmente, se interpretan los resultados.

Conforme avances en este estudio, irás aprendiendo la manera de interpretar datos. Por el momento, para el caso que nos ocupa, nos limitaremos a hacer las siguientes observaciones:

e) La pregunta planteada se refiere a todos los integrantes de la preparatoria, y los datos se obtuvieron únicamente de un salón de clase. El colectivo total implicado en la pregunta de interés se llama población, y el colectivo implicado en los datos obtenidos se llama muestra.

Actividad 1.1 d (Cont.)

En nuestro caso, la población es el total de personas de la preparatoria, y la muestra es el grupo escolar de donde se obtuvieron los datos.

Los resultados muestrales se llaman estadígrafos o estadísticos, y los resultados correspondientes a la población se llaman parámetros.

Escribe los resultados de tu muestra en porcentajes;

Porcentaje de mujeres: _____
Porcentaje de hombres: _____

Estudiamos la muestra pero quien nos interesa es la población. En este caso lo que nos interesa es el porcentaje de mujeres y el de hombres existentes en toda la preparatoria. Es decir, nos interesan los parámetros.

ii) La pregunta más importante es: ¿Los resultados de la muestra son válidos para toda la población? En otras palabras, ¿podemos tomar los estadígrafos obtenidos como los parámetros? ¿Si cambiamos de muestra, es decir si usamos otro grupo escolar, cambiarán los porcentajes? Si la respuesta es afirmativa, ¿Cambiarán mucho o poco? ¿De qué dependerá? Si en vez de usar una muestra obtenemos los datos de toda la población, ¿cambiarán los resultados? Escribe en el siguiente cuadro tu opinión al respecto.

Para obtener más elementos sobre los interrogantes planteados, realiza la siguiente actividad.

Actividad 1.1 e

Organizarse en equipos de 5 integrantes. Cada equipo investigará en grupos escolares distintos la variable género. Compartan los resultados y vuelvan a revizar las interrogantes planteadas.

Anota tus conclusiones

Los resultados de la actividad anterior con toda seguridad validan la siguiente afirmación: si dos o más muestras son obtenidas de la misma población, es casi cierto que los resultados podrían no ser exactamente los mismos. El valor de un estadígrafo variará de muestra a muestra. Esto es llamado variabilidad muestral.

En otros cursos de estadística podrás comprobar que si se usan técnicas de muetreo apropiadas, los resultados muestrales son muy útiles a tal grado que se puede asegurar que diferencias inaceptables entre muestras son prácticamente imposibles. Para lograr esto, es también muy importante el tamaño de la muestra.

Hagamos una recapitulación de todo el proceso realizado:

Se plantea una pregunta que no tiene una sola respuesta; se recolectan y analizan datos, y se interpretan los resultados.

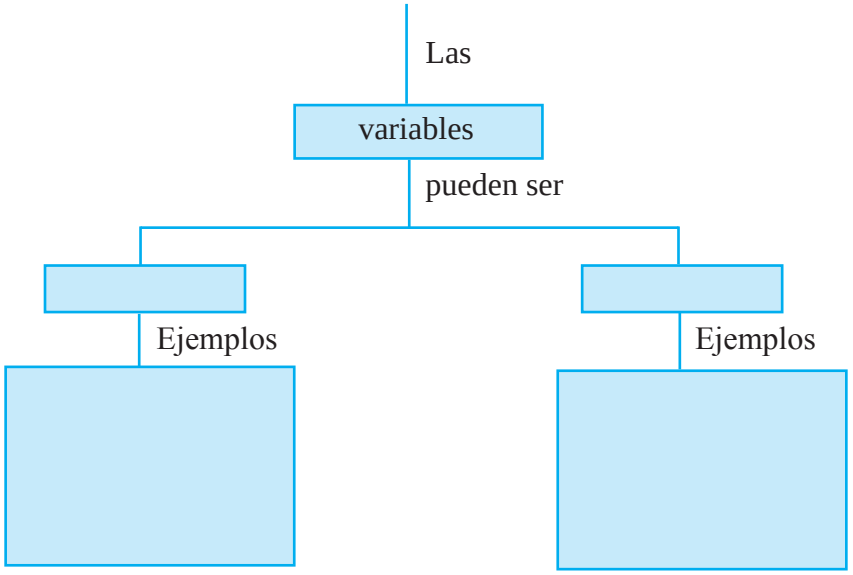
Con base en esto, ya puede establecerse una definición más completa de estadística:

Estadística es la ciencia que para contestar una pregunta cuyas respuestas muestran variabilidad, recolecta y analiza datos, e interpreta resultados.

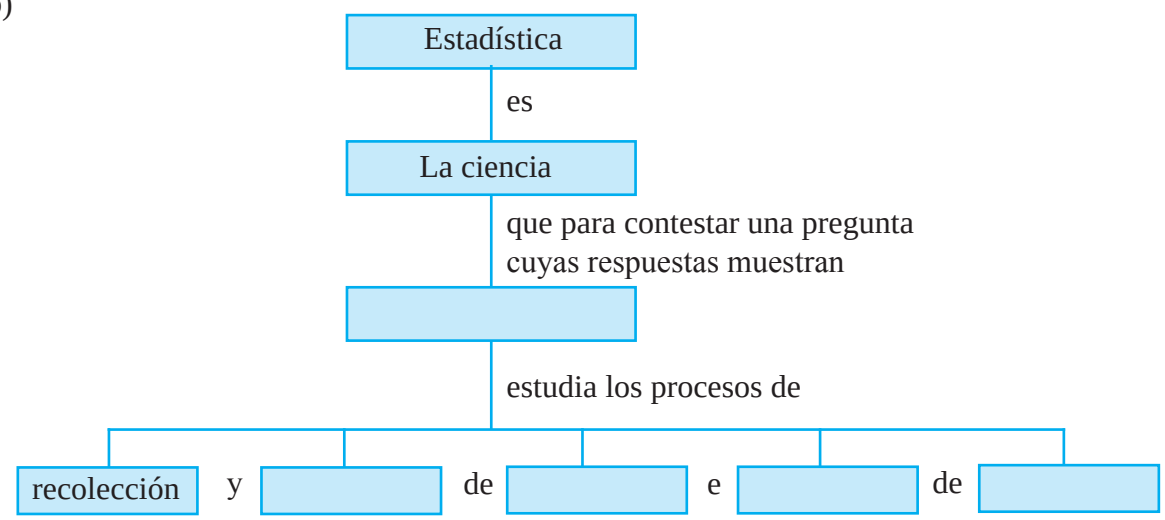
Ejercicio 1.1

Completa los siguientes esquemas:

a)



b)



Lección 1.2 Población y muestra

Objetivo: Comprender los conceptos de población y muestra.

Actividad 2

Qué hacer

Consulta las **páginas 12 a 14** y al finalizar tu estudio contesta:

1) Explica la diferencia entre población y muestra y presenta ejemplos para cada uno de ellos.
2) Escribe dos razones por las cuales es necesario recoger una muestra.
3) Explica el significado de población finita y de población infinita.
4) Resuelve el ejercicio 1.2

Las nociones de población y muestra, son fundamentales en estadística, por lo que las analizaremos con más precisión.

Población: conjunto completo de individuos u objetos a los que se les observa una característica particular que será objeto de estudio.

Tamaño de la población: es el número de individuos que forman la población. Se denota con la letra N.

Por lo general se piensa que una población es una colección de personas, pero en estadística la población puede ser una colección de animales, de objetos manufacturados o de cualquier cosa que se pueda medir, contar o jerarquizar. Por ejemplo, una población de interés agrícola es el conjunto formado por la mosquita blanca en un cultivo de soya.

Ejemplo 1

Consideremos la población constituida por los alumnos de primer año inscritos en la Preparatoria Allende. Podemos estar interesados en estudiar las siguientes características (o variables) poblacionales:

-Estatura (en cm) de alumnos (as):

Después de medir las estaturas de cada alumno (a), obtendríamos un conjunto de datos parecidos a los siguientes:

155, 161, 148, 166, 156, ... 150, 149, 172, 180

-Calificaciones en matemáticas I:

10, 8, 7, 6, 4,8, 10,...9, 10, 8, 8

Matemáticas I

Individuo	Calificación
María	10
José	8
Ana	7
Antonio	6
Alma	4
Rira	8
Ariana	10

Ejemplo 2

Conjunto de temperaturas (º C) en un determinado día, a las 11 horas, en todas las ciudades de Sinaloa:

36, 30, 28, 36, 21, 22, ... 38, 36

A veces, se identifica a la población con la variable o característica poblacional que se pretende estudiar.

Refiriéndonos al ejemplo1, hablamos de:

- Población de estaturas de los alumnos...
- Población de calificaciones en Matemáticas I....

Refiriéndonos al ejemplo 2, hablamos de:

- Población de temperaturas a las 11 horas.

Hay dos tipos de poblaciones: finitas e infinitas. Una población es finita, cuando se puede enumerar físicamente a todos los elementos que componen la población. Cuando los elementos son ilimitados, se dice que la población es infinita. Las poblaciones infinitas generalmente provienen de procesos productivos, en el sentido de que se siguen produciendo indefinidamente.

Ejemplo 3

Poblaciones finitas:

-Estudiantes de una preparatoria.
-Municipios del estado de Sinaloa.

Poblaciones infinitas:

-Todas las personas que podrían contagiarse de influenza.
-Todas las toneladas de trigo que se producirán en México
-Posibles lanzamientos de una moneda al aire.

No siempre es posible estudiar a todos los elementos de la población. ¿Por qué?

- Porque la población puede ser infinita.

  Ejemplo: Población constituida por las presiones atmosféricas, en diferentes puntos de una ciudad.

- Porque el estudio de la población conlleva la destrucción de ésta.

  Ejemplo: Población de fósforos de una caja.

- Porque el estudio de la población es muy caro.

Ejemplo: Población constituida por todos los ciudadanos de un país que prefieren a cierto candidato.

Cuando no es posible estudiar exhaustivamente, todos los elementos de una población, estudiamos sólo algunos elementos, a los que damos el nombre de muestra.

El tamaño de la población se denota con la letra N, y el de la muestra con n.

Una muestra es una parte de la población que se estudiará para conocer las características de dicha población.

Al igual que la población, a veces, se identifica a la muestra con la variable o característica poblacional que se pretende estudiar. En este caso, una muestra, se define como el conjunto de datos u observaciones recolectados a partir de un subconjunto de la población, que se estudia.

Actividad 1.2 a

Estudia el siguiente ejemplo que trata de la aplicación de términos básicos

Ejemplo

Supongamos que nos interesa determinar el número de libros comprados por cada alumno, el deporte preferido y el tiempo semanal dedicado al estudio de los estudiantes de la universidad.

Identificaremos algunos de los términos estudiados:

1. La población es la colección de todos los alumnos (as) que estudian en la universidad.
2. Una muestra es cualquier subconjunto de esa población. Por ejemplo, los alumnos del bachillerato.
3. Las variables de interés son: “número de libros comprados”, “deporte preferido” y “tiempo semanal dedicado al estudio”.
4. La variable “número de libros comprados” es cuantitativa, la variable “deporte preferido” es cualitativa y la variable “tiempo semanal dedicado al estudio” es cuantitativa.
5. Un dato para cada variable podría ser: 5 libros comprados, beisbol y 6 horas.
6. El estadígrafo es el valor que se obtiene de todos los integrantes de la muestra, es decir alumnos de preparatoria.
7. El parámetro es el valor para toda la población, en este caso alumnos de toda la universidad.

Ejercicio 1.2

I. Contesta correctamente:

1. Un fabricante de medicamentos está interesado en la proporción de personas que padecen hipertensión (presión arterial elevada), cuya condición pueda ser controlada por un nuevo producto desarrollado por la empresa. Se condujo un estudio en el que participaron 5,000 personas que padecen de hipertensión, y se encontró que un 80% de las personas pueden controlar su hipertensión con el medicamento. Suponiendo que las cinco mil personas son representativas del grupo con hipertensión, conteste las siguientes preguntas:
   a) ¿Cuál es la población?
   b) ¿Cuál es la muestra?
   c) Identifique el parámetro de interés
   d) Identifique el estadígrafo y proporcione su valor
   e) ¿Se conoce el valor del parámetro?

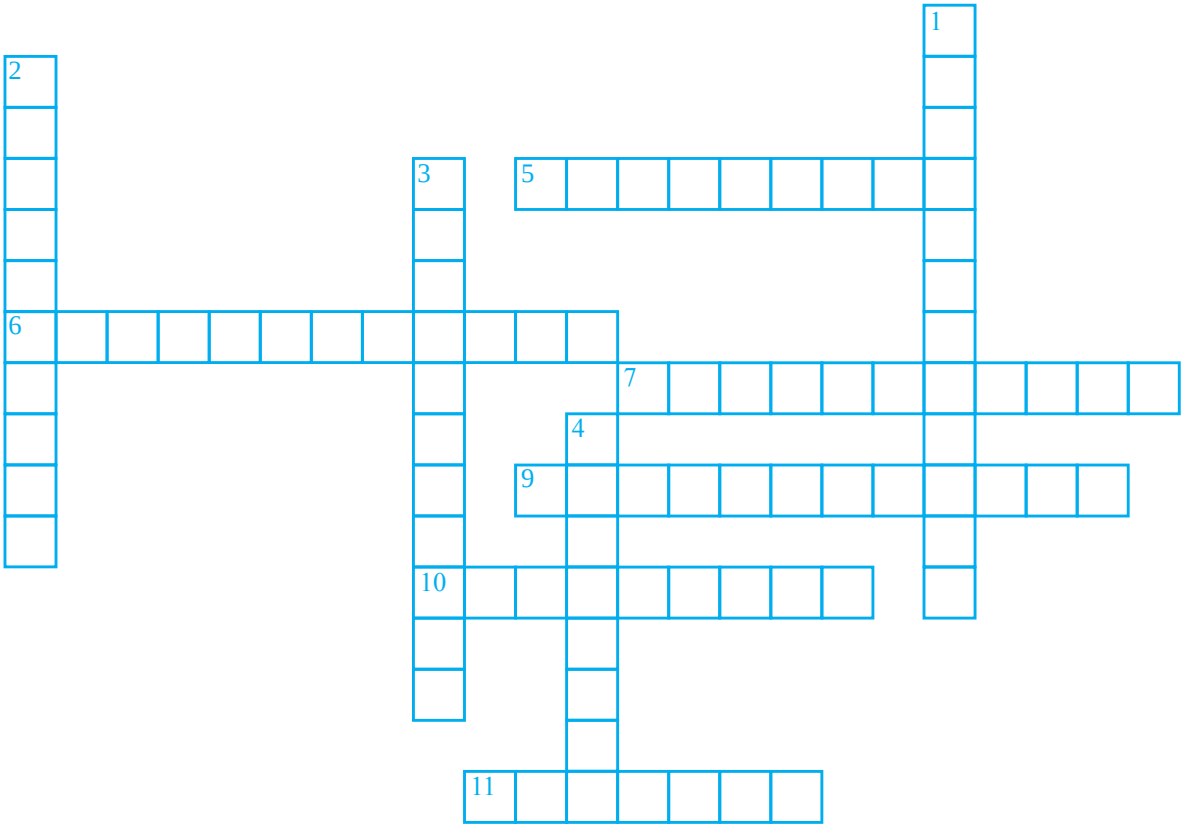
2. Un campesino posee 127 gallinas. Para probar la eficacia de la alimentación, las pesa a todas antes y después de los 20 días que dura el tratamiento. ¿El conjunto de las 127 gallinas es población o muestra?

3. De las siguientes variables, indique cuáles son cualitativas y cuáles cuantitativas:
   a) Preferencia electoral en la última elección.
   b) Temperatura de una persona.
   c) Número de periódicos leídos por una persona diariamente.
   d) Número de votos obtenidos por un candidato.

4. Suponga que un niño de 12 años le pide que le explique la diferencia entre una muestra y una población.
   a) ¿Qué información debe incluir en su respuesta?
   b) ¿Qué razones proporcionaría al niño sobre por qué debe tomarse una muestra en vez de encuestar a todos los elementos de la población?

5. Suponga que un niño de 12 años le pide explicarle la diferencia entre el valor de un estadígrafo y un parámetro?
   a) ¿Qué información debe incluir en su respuesta?
   b) ¿Qué razones proporcionaría al niño sobre por qué debe reportarse el valor de un estadígrafo en vez de un parámetro?

II. Resuelve el siguiente crucigrama



Horizontales

5. Conjunto completo de individuos u objetos a los que se les observa una característica particular que será objeto de estudio.
6. Resultados obtenidos de una muestra
7. Variable que simplemente registra una categoría asignada mediante una palabra o símbolo.
9. Principio fundamental que pone de manifiesto la necesidad de usar la estadística.
10. Es la persona o cosa a la que se le asigna un número o categoría.
11. Es una parte de una población que se estudiará para conocer las características de dicha población.

Verticales

1. Variable que mide una característica numérica.
2. Resultados que corresponden a toda la población.
3. Ciencia que para contestar preguntas cuyas respuestas presentan variabilidad, recolecta y analiza datos, e interpreta resultados.
4. Es cualquier característica de interés de una persona o cosa a la que se le puede asignar un número o una categoría.

Lección 1.3 El método estadístico

Objetivo: **Comprender que la estadística es un proceso de investigación que consiste en: formular preguntas, recolectar y analizar datos, e interpretar resultados.**

Actividad 3

Qué hacer

1. Consulta la **página 17** y al finalizar tu estudio explica la diferencia entre censo y encuesta.
2. Consulta la **página 18** y al finalizar tu estudio explica la importancia de que una muestra sea representativa de la población.
3. Consulta las **páginas 18 a 21** y al finalizar tu estudio contesta lo indicado:
   3.1) Explica en qué consiste el proceso de investigación estadístico.
   3.2) En las siguientes situaciones indica cuáles constituyen ejemplos de estadística descriptiva y cuáles de estadística inferencial:
      a) Un lote de 100 aparatos de televisión se considera en buen estado para su venta si al ser probados, 10 de ellos no presentan defectos.
      b) En una empresa 150 empleados ganan un promedio mensual de $3100.00.
      c) Según una encuesta entre 1000 ciudadanos, el candidato a diputado de determinado partido por el segundo distrito, tendrá en las eleciciones una preferencia de 59%.
4. Consulta las **páginas 23 y 24** y al finalizar tu estudio contesta lo indicado:
   4.1) Proporciona un ejemplo de aplicación de la estadística diferente a los presentados.

Un censo, es una encuesta al 100 %

Las encuestas de opinión que con bastante frecuencia aparecen en medios de comunicación, son los ejemplos más visibles de la aplicación de la estadística en la vida diaria.

A partir de la definición de estadística, y ante la necesidad de usar muestras, puede delinearse el proceso a seguir en toda investigación estadística.

A continuación se precisará tal proceso.

Antes de iniciar el proceso estadístico, se debe definir la población, y a continuación contestar la siguiente pregunta: ¿Debemos recolectar datos correspondientes a los individuos de toda la población o es suficiente estudiar una muestra? Cuando se recolectan los datos de toda la población, se efectúa un **censo**, y cuando se recolectan de una muestra se efectúa un **sondeo o encuesta** muestral.

Como ya se ha señalado, lo más usual es que estudiemos una muestra.

La teoría estadística, ha demostrado que para obtener resultados confiables, no es necesario estudiar a toda la población; es decir, es suficiente estudiar una muestra extraída en forma adecuada de dicha población.

¿Qué significa extraer en forma adecuada una muestra? La respuesta no es sencilla y corresponde contestarla a una rama de la estadística denominada teoría del muestreo. Para los propósitos de este curso introdcuctorio, es suficiente entender que para que una muestra sea adecuada, debe ser representativa de toda la población.

Una muestra es representativa, si contiene en sí misma de manera aproximada, todas las características fundamentales de la pòblación.

Por ejemplo, si en la población hay un 52% de mujeres, una muestra representativa deberá contener aproximadamente 52% de mujeres.

Imagina un pueblo pequeño que tenga la misma proporción que México tiene como un todo: misma proporción de médicos, de escuelas, de pobres, de personas que votan por el PRI, PRD o PAN, de personas que están a favor del aborto, etcétera. Esto permitiría que sólo se tendría que entrevistar a las personas de ese pueblo para saber, por ejemplo, cuál es la opinión pública en México con respecto a temas como la preferencia partidista. Una muestra perfecta sería como ese pueblo: una versión a escala de la población, que reflejaría cada una de las características de toda la población. Por supuesto, una muestra perfecta como ésta, no puede existir para poblaciones complejas, pero una buena muestra reproduce en forma aproximada las características de interés que existen en la población. Esta buena muestra es la llamada muestra representativa de la población. Debemos tomar conciencia que esta buena muestra, en general, tendrá un error. Es decir, la muestra sólo proporciona resultados que son aproximados a los resultados correspondientes a toda la población. El gran mérito de la estadística es lograr que ese error sea pequeño. Para plantear ese error, la estadística se auxilia de otra ciencia denominada probabilidad, la cual estudiarás en el próximo semestre.

La muestra, sólo permitirá estimar las características de la población

En resumen, el proceso de investigación estadística conlleva la siguiente secuencia de eventos:

1. Planteamiento del problema

La situación a investigar se define cuidadosamente. Básicamente, esta etapa consiste en:

-Formular una (o más) preguntas que pueden ser contestadas con datos.
- Definir la población.
-Identificar la variable.

2. Recolectar datos

- Diseñar un plan para recolectar de manera adecuada los datos.
- Utilizar el plan para recolectar los datos.

3. **Analizar los datos muestrales.**
   - Los datos se organizan en tablas.
   - Los datos se presentan en gráficas.
   - Se calculan medidas de resumen y de variabilidad.

4. **Interpretar los resultados muestrales.**
   - Interpretar el análisis.
   - Relacionar la interpretación con la pregunta original.

5. **Generalizar de manera adecuada, los resultados muestrales a toda la población**

Para efecto de enseñanza, todo este proceso se ha dividido tradicionalmente en dos partes:

La primera, que incluye los primeros cuatro subprocesos, se denomina estadística descriptiva, y, la segunda que se encarga del quinto subproceso, se llama estadística inferencial.

La estadística descriptiva incluye la recolección, análisis e interpretación de resultados muestrales.

Al aplicar las herramientas de la estadística descriptiva se obtienen los estadígrafos,

La estadística inferencial incluye las herramientas para obtener conclusiones acerca de la población a partir de los resultados muestrales.

El objetivo de la estadística inferencial es inferir las características de la población a partir de las características de la muestra. En otras palabras, a partir de los estadígrafos, se infieren los parámetros.

En este proceso inferencial, siempre es de esperar un error, de tal manera que:

$$Parámetro = estadígrafo \pm un\ pequeño\ error$$

Es decir, la inferencia estadística sólo hace estimaciones sobre la población total, a partir de la información de la muestra.

En este texto, para efecto de organización de los temas, se considerarán las siguientes divisiones:

Estadística descriptiva
1. Planteamiento del problema.
2. Recolección de datos.
3. Exploración de datos:
   3.1 Análisis de datos.
   3,2 Interpretación de resultados.

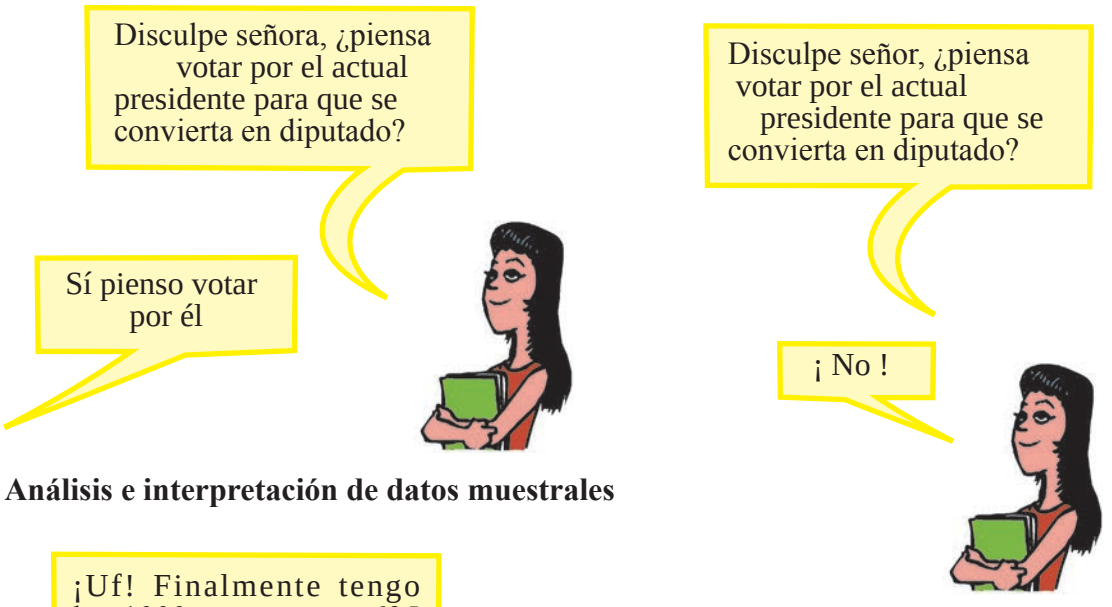
Estadística inferencial
4. Generalización a toda la población.

Este es un curso de estadística descriptiva con énfasis en la exploración de datos.

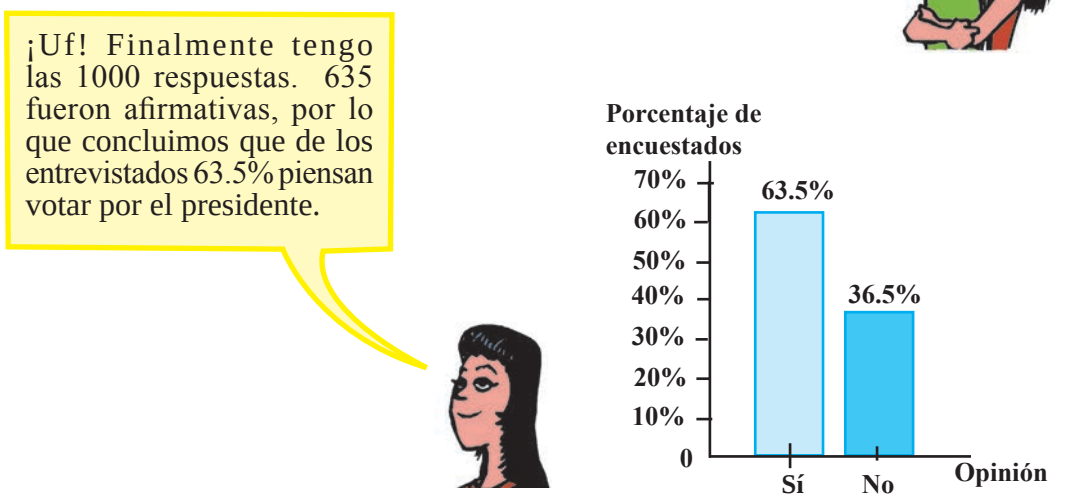
Ejemplo

Planteamiento del problema

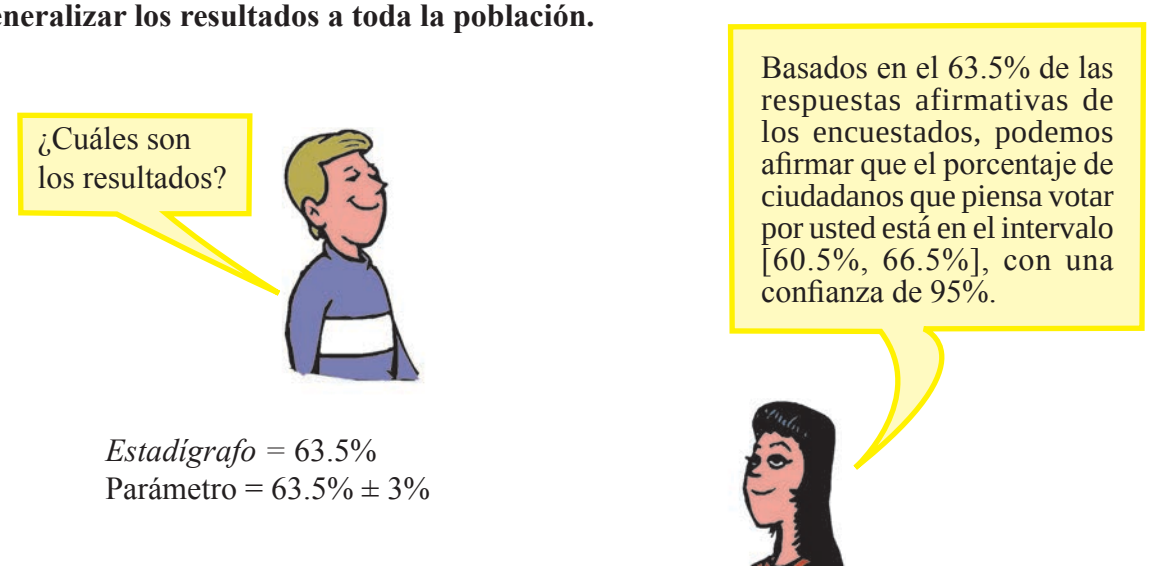
Recolectar los datos



Análisis e interpretación de datos muestrales

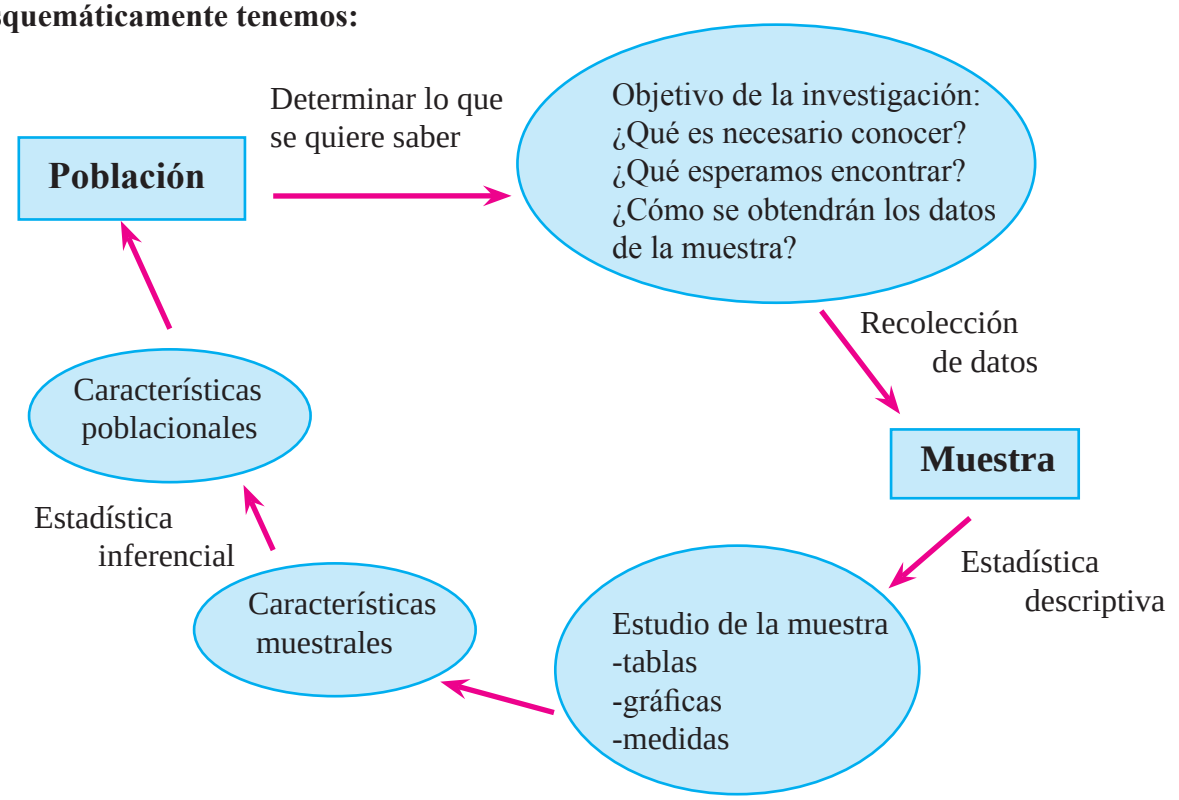


Generalizar los resultados a toda la población.



Estadígrafo = 63.5%
Parámetro = 63.5% ± 3%

Esquemáticamente tenemos:



Resumiendo: un análisis estadístico incluye dos fases fundamentales:

1ª Fase	Estadística Descriptiva	Su objetivo es describir una muestra, presentando evidencias de sus características y propiedades principales.
2ª Fase	Estadística Inferencial	Conociendo las propiedades obtenidas a partir del análisis descriptivo de la muestra, estima propiedades para toda la población.

Actividad 1.3 b

Estudia el siguiente ejemplo:

Ejemplo

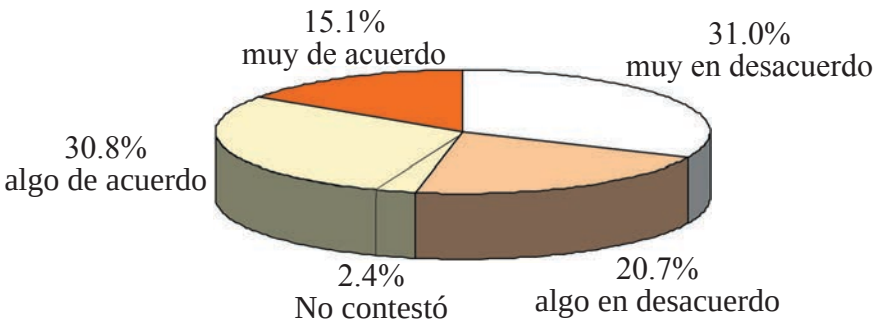
1. Planteamiento del problema.
   El horario de verano es un programa de ahorro de energía implementado por el Gobierno Federal. ¿Qué opina la población sinaloense acerca de este programa?
   - Formular la pregunta. ¿Cuál es la opinión de un sinaloense con respecto al horario de verano?
   - Definir la población. Población del estado de Sinaloa.
   - Identificar la variable. Opinión con respecto al horario de verano.

2. Recolección de datos.
   Para contestar esta pregunta, el periódico el DEBATE coordinó una encuesta a nivel estatal. Se entrevistará a 899 ciudadanos en sus domicilios en los municipios de Ahome, Guasave, Salvador Alvarado y Culiacán.

3. Análisis de los datos muestrales. Los resultados a nivel muestral fueron presentados a través del siguiente gráfico circular.

RECHAZAN

El horario de verano



4. Interpretar los resultados muestrales.
   Si observamos que el 31% está muy en desacuerdo, y el 15.1% muy de acuerdo, puede concluirse que el horario de verano se rechaza. Sin embargo, existe ambigüedad en el 30.8% que está algo de acuerdo y en el 20.7% que está algo en desacuerdo. Sumando estos dos porcentajes obtenemos que 51.5% de las personas, de alguna manera no rechazan el cambio de horario. En conclusión, la opinión parece dividida.

Actividad 1.3a (Cont.)

Con base en la información proporcionada en el reporte sobre el horario de verano, contesta:
Determina cuál de las siguientes afirmaciones es de naturaleza descriptiva y cuál inferencial.

a) El 31% de las personas encuestadas, está muy en desacuerdo con el horario de verano_____
b) El 31% de las personas está muy en desacuerdo con el horario de verano_____
c) Según el reporte, ¿Cuál de las dos afirmaciones anteriores debe hacerse?_____
d) ¿Qué grupo de personas fue encuestada?_____
e) ¿Cuántas personas fueron encuestadas?_____
f) ¿Qué información se obtuvo de cada persona?_____
g) Determina la cantidad de personas que:
Están muy de acuerdo con el horario de verano_____
Están algo de acuerdo con el horario de verano_____
Están muy en desacuerdo con el horario de verano_____

Actividad 1.3 b

Los campos de aplicación de la Estadística son muchos y muy variados. Estudia la siguiente actividad que muestra ejemplos de dichas aplicaciones.

Estudios de mercado

Un gerente de una fábrica de detergentes pretende lanzar un nuevo producto para lavar ropa, por lo que encarga a una empresa especializada en estudios de mercado «estimar» el porcentaje de compradores potenciales de ese producto.

Problema: Se pretende, a partir de un porcentaje de respuestas afirmativas, de entre los encuestados sobre la compra de un nuevo producto, obtener una estimación del número de compradores de la población.

Pregunta: ¿Cuál es la opinión de las personas acerca del nuevo producto para lavar ropa?

Variable: Opinión de las personas.

Población: Conjunto de todas las familias del país.

Muestra: Conjunto de algunas familias, encuestadas por la empresa.

Medicina

Se pretende estudiar el efecto de un nuevo medicamento para curar determinada enfermedad. Se selecciona un grupo de 20 pacientes. A 10 de estos pacientes se les administró el nuevo medicamento y a los 10 pacientes restantes se les suministró el medicamento habitual.

Problema: Se pretende, a partir de los resultados obtenidos, tomar una decisión sobre cuál de los dos medicamentos es mejor.

Pregunta: ¿Es efectivo el nuevo medicamento?

Variable: Efectividad del medicamento.

Población: Conjunto de todos los pacientes con una enfermedad que el medicamento estudiado pretende curar.

Muestra: Conjunto de 20 pacientes seleccionado.

Actividad 1.3 b (Cont.)

Control de calidad Un administrador de una fábrica de tornillos pretende asegurarse de que un porcentaje de piezas defectuosas no exceda un determinado valor, a partir del cual determinado pedido pudiera ser rechazado.

Problema: Se pretende, a partir de un porcentaje de tornillos defectuosos presentes en una muestra, «estimar» el porcentaje de defectuosos en toda la población.

Pregunta: ¿Se encuentran dentro del rango permitido los tornillos fabricados?

Variable: Calidad de tornillos.

Población: Conjunto de todos los tornillos fabricados o por fabricar por una fábrica. utilizando un mismo proceso. Muestra: Conjunto de tornillos escogidos de entre los producidos.

Pedagogía Un conjunto de pedagogos desarrolla una nueva técnica para el aprendizaje de lectura, en escuelas primarias. Se pretende probar que el tiempo de aprendizaje con la nueva técnica es menor que con el método tradicional.

Problema: Se pretende decidir cuál técnica de aprendizaje es mejor.

Pregunta: ¿La nueva técnica requiere menos tiempo de aprendizaje que la anterior?

Variable: Tiempo de aprendizaje.

Población: Conjunto de todos los alumnos de escuelas primarias que no saben leer.

Muestra: Conjunto de alumnos de algunas escuelas seleccionadas para este estudio. Los alumnos fueron separados en dos grupos para aplicar las dos técnicas en confrontación.

Ejercicio 1.3

Proyecto 1: ¿Cómo son los alumnos(as) del grupo?

Objetivos

Se trata de elaborar, conforme se avanza en el curso, un perfil de los alumnos, identificando al alumno típico y analizando si hay diferencias entre el alumno y la alumna típicos. Para trabajar, pueden hacerlo en equipos de 5 integrantes. En este primer avance del proyecto, deberán reportar los siguientes aspectos:

Planteamiento del problema:

-Formular preguntas sobre las características que se quieren incluir en el estudio (por ejemplo: sexo, deporte preferido, color de ojos, color de pelo, número de calzado etc.).

- Definir la población. En este caso, la población será todo el grupo.

-Identificar las variables.

Recolectar datos

Deberán analizarse las diferentes formas en que se podrían obtener los datos:

-Por simple observación: como el sexo, color de pelo y ojos.

-Si se requiere una medición: como el peso y estatura.

-Mediante pregunta directa; es decir realizar una pequeña encuesta. Por ejemplo, preguntar sobre número de hermanos, sobre qué medio utiliza para venir a la escuela.

Los datos deberán reportarse en una tabla parecida a la siguiente:

Plantea aquí tus variables

Estu-diante	Sexo	Color de ojos	Color de pelo	Número de calzado				
1 2 3 4 5 . . .								

Lección 1.4 Nociones de muestreo

Objetivos: Empezar a conocer que una muestra bien elegida representará con más seguridad a la población y que hay maneras de elegir una muestra que las hace no representativas de la población.

Comprender aspectos básicos del muestreo.

Actividad 4

Qué hacer

Consulta las páginas indicadas en cada caso, y al finalizar tu estudio resuelve lo solicitado:

1) **Páginas 26 a 28.** Explica las diferencia entre una investigación basada en un experimento y una basada en estudios observacionales; el significado de marco muestral, unidades de muestreo, población homogénea y resultados con igual probabilidad de ocurrir.
2) **Página 28 a 30.** Explica cómo formar una tabla de números aleatorios. Explica la diferencia entre muestreo probabilístico y no probabilístico.
3) **Páginas 31 a 32.** Explica brevemente las características principales del muestreo aleatorio simple.
4) **Págs. 33 a 34.** Explica brevemente las características principales del muestreo sistemático.
5) **Págs. 34 a 35.** Explica brevemente las características principales del muestreo estratificado.
6) **Página 36.** Explica brevemente las características principales del muestreo por conglomerados.

Un ejemplo de experimento: Pruebas realizadas sobre cultivos de cítricos infestados con Diaphorina Citri, en el norte de Sinaloa, revelan que el insecticida sistémico Thiametoxam, a dosis de 50 ml por 100 litros de agua, logra 85% de mortandad del insecto plaga.

En la lección anterior, se estableció que la teoría estadística ha demostrado que para obtener resultados confiables, no es necesario estudiar a toda la población; es decir, es suficiente estudiar una muestra *extraída en forma adecuada* de dicha población.

Bajo esta perspectiva, el objetivo fundamental de la estadística es conocer determinadas características de toda una población por medio de la selección y estudio de sólo algunos elementos pertenecientes a dicha población.

Para el logro de este objetivo, se requieren **métodos para seleccionar muestras** que produzcan datos que sean representativos de la población.

Antes de estudiar estos métodos, necesitas conocer algunos términos que te facilitarán la comprensión de dichos métodos.

1. **Tipo de investigación.** Para recolectar datos, existen dos tipos de investigación o estudio: **Experimento y estudios observacionales.**

En un **experimento**, el investigador controla o modifica el entorno y observa el efecto sobre la variable bajo estudio.

Ejemplo 1 Un investigador agrícola que busca probar la cantidad óptima de cierto insecticida, aplica diversas dosis y analiza los efectos sobre el control de plagas.

En un **estudio observacional,** el investigador no modifica el entorno, simplemente observa y recolecta los datos.

Ejemplo 2 La encuestas son estudios observacionales.

2. Marco muestral. Es una lista, o conjunto, de los elementos que pertenecen a la población de la cual se toma la muestra.

De manera ideal, el marco muestral debe ser idéntico a la población con cada uno de los elementos de la población incluido una vez y sólo una vez.

Ejemplo 3 Posibles Marcos muestrales:

-Directorio telefónico.

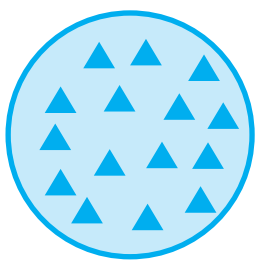
-Lista de electores registrados.

-Lista de asistencia de estudiantes de una escuela.

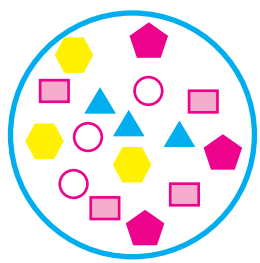
Dependiendo de la naturaleza de la información que se busque, la lista de electores o el directorio telefónico pueden o no servir, como marco muestral. Debido a que sólo los elementos del marco tienen la oportunidad de ser seleccionados como parte de la muestra, es importante que el marco muestral sea representativo de la población.

3. **Unidades de muestreo.** Son divisiones posibles de la población para propósitos de selección de la muestra. Por ejemplo, Una ciudad, está compuesta por colonias, manzanas, familias o personas.

4. **Población homogénea y población heterogénea.** Observa las siguientes ilustraciones para que obtengas una idea intuitiva de la homogeneidad y heterogeneidad de un grupo.



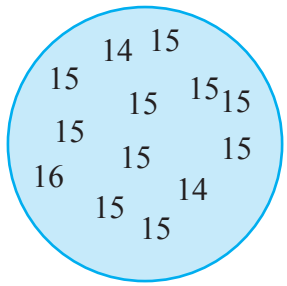
Grupo homogéneo



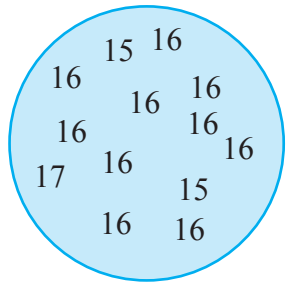
Grupo heterogéneo

En un contexto estadístico, una población homogénea presenta valores correspondientes a alguna variable, con poca diferencia entre ellos. Entre más disparidad presenten los valores, la población se considera más heterogénea.

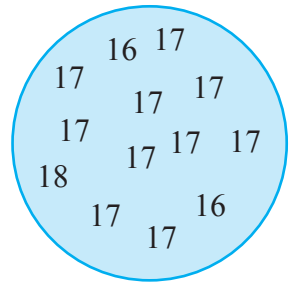
Por ejemplo, si consideramos la población de edades de estudiantes en cada grado de una preparatoria, tendremos poblaciones *más o menos homogéneas.*



Edades de estudiantes de 1º de preparatoria

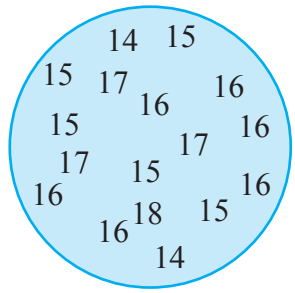


Edades de estudiantes de 2º de preparatoria

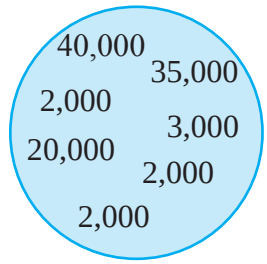


Edades de estudiantes de 3º de preparatoria

En cambio, si consideramos edades de estudiantes de preparatoria en general, sin separar grados de estudio, tendremos una población más heterogénea.



Otro ejemplo de población heterogénea, lo constituyen los salarios mensuales de los trabajadores de una gran empresa. En esta empresa trabajan ejecutivos, técnicos especializados y obreros, todos ellos con sueldos muy diversos..



5. Resultados con igual probabilidad de ocurrir. Para obtener una idea intuitiva sobre lo que significa igual probabilidad, analicemos los siguientes ejemplos.

Ejemplo 1

Imaginemos un dado normal de seis caras, cada una marcada con 1, 2, 3, 4, 5 y 6 puntos.

Si el dado es «*honesto*», las seis caras están en igualdad de condiciones de salir. Se afirma entonces, que al lanzar el dado cada resultado tiene la misma probabilidad de ocurrir.

¿Cuál es la probabilidad de obtener un punto?_____

Ejemplo 2

Un experimento consiste en girar la flecha indicada en la figura:

¿Existen las mismas condiciones para que la flecha termine apuntando tanto en *zona blanca* como en *zona azul*? No, la zona azul presenta mayor área que la blanca. En este caso, los resultados no tienen la misma probabilidad de ocurrir.

La probabilidad de que la flecha termine apuntando en blanco es _____; y en azul _____

6. **Tabla de números aleatorios.** Está formada por números que se pueden generar de la siguiente manera:

1. Marcar diez papelitos o bolas con los números del *0* al *9*.
2. Colocar los papelitos dentro de un recipiente
3. Mezclar los papelitos lo mejor posible y seleccionar uno de ellos. Anotar el número. Ejemplo, sale el *2*.
4. Regresar el papelito al recipiente, volver a mezclar y sacar otro papelito. Ejemplo, sale el *7*. El número aleatorio de dos dígitos es *27*.
5. La operación se repite de manera sucesiva tanto como sea necesario. Por ejemplo, para formar el primer número de la tabla al margen, el proceso se repitió cinco veces: primero salió el *2*, después el *7*, otro *7*, un *6* y finalmente un *7*.

Para cada número de esta tabla en particular, debemos mezclar, seleccionar, anotar y regresar el papelito, cinco veces.

Parte de una tabla de números aleatorios de cinco dígitos

27767	43584
13025	14338
80217	36292
10875	62004
54127	57326
60311	42824
49739	71484
78626	51594
66692	13986
44071	28091

Actividad 1.4 a

Organizarse en equipos para formar una tabla de 200 números aleatorios de cuatro dígitos. La tabla final debe ser única para todo el grupo. Es decir deberá contener las mismas columnas y renglones y la misma posición de los números.

Escribe los números de la tabla en el cuadro siguiente.

Tabla de números aleatorios formada por todo el grupo

Una vez comprendido los conceptos anteriores, procederemos a estudiar las cuestiones básicas de los procedimientos de muestreo.

Hay muchos tipos de muestreo, sin embargo, todos pueden clasificarse en dos categorías: **muestreo no probabilístico** y **muestreo probabilístico.**

Muestreo no probabilístico. En este muestreo, el investigador escoge los elementos de acuerdo con su juicio, necesidades o conveniencia. Otras muestras que son de este tipo, son las muestras voluntarias, en las que los receptores de lo solicitado, deciden si deben contestar o no.Ninguno de estos procedimientos son estadísticamente aceptables.

Ejemplos

- Si queremos estimar cuánto gasta una persona que va de compras a un centro comercial y extraemos una muestra entre los compradores que parecen haber gastado cierta cantidad, habremos elegido de manera deliberada una muestra para confirmar nuestra opinión anterior. Este tipo de muestra se llama **muestra de juicio** porque el investigador emplea su propio juicio para elegir los individuos que debe incluir en la muestra.

Ejemplos (Cont.)

- Se desea introducir en el mercado un jabón que limpia las impurezas de la cara. Con base en su experiencia el investigador decide aplicar directamente el jabón al público y pedir su opinión en forma escrita. Tal actividad se realiza en centros comerciales. Todo ello es a conveniencia del investigador, por lo que es una **muestra de conveniencia.**
- En algunos restaurantes, se muestra una libreta abierta para que los clientes escriban de manera voluntaria, su opinión acerca del servicio brindado. Esta es una **muestra voluntaria.**
- Son **muestras voluntarias**, las producidas por las encuestas realizadas por diversos medios de comunicación, que piden a las personas den su opinión por teléfono u otro medio, sobre ciertos temas.

En mayo de 2009, en radio y televisión se anunció que después de una amplia investigación con muestras de sangre de 3 000 mexicanos voluntarios, se logró decifrar el genoma humano común a todos los mexicanos.

Las muestras *no probabilísticas*, no son aceptables desde el punto de vista estadístico, porque la selección arbitraria y no estructurada, impide controlar el error muestral. No se conoce ningún método objetivo para medir la confianza que debe tenerse en los resultados.

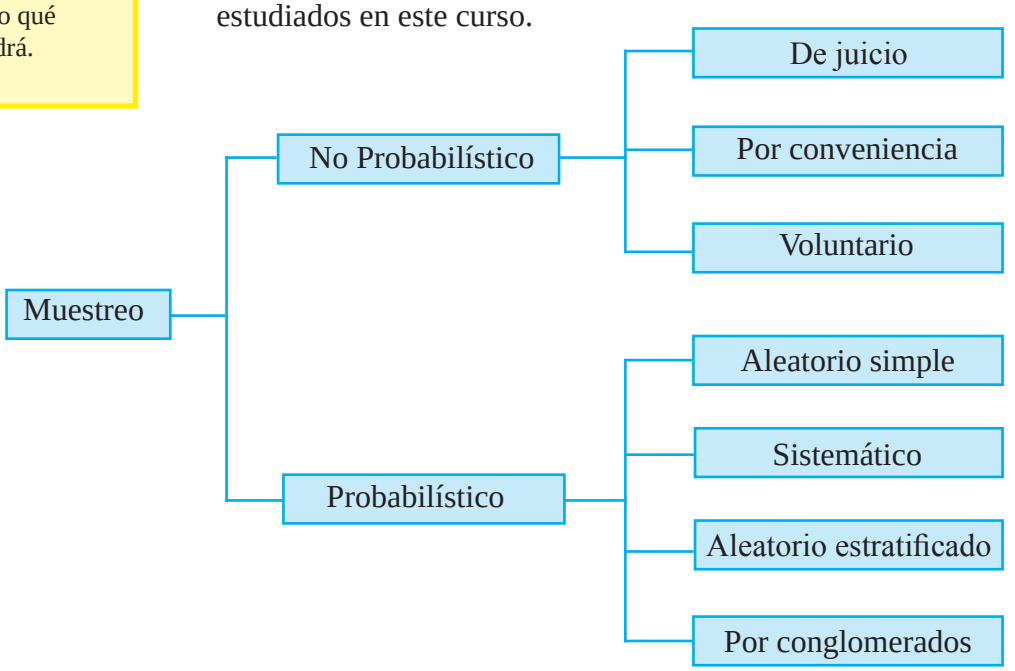
Sin embargo, existen algunas situaciones en las que el muestreo no probabilístico se vuelve una alternativa útil. Por ejemplo, en la investigación médica, a menudo se utilizan grupos de voluntarios.

Muestreo probabilístico. Son procedimientos de muestreo en los que, con la ayuda de la probabilidad, puede determinarse tanto el tamaño de muestra requerida como el grado de error de la muestra.

En un muestreo probabilístico, no es la persona quien elije la muestra, sino el azar; como en el caso de un dado que al lanzarlo no se sabe de antemano qué número saldrá.

Entre los principales procedimientos de muestreo probabilístico se encuentran: *El muestreo aleatorio simple, el muestreo sistemático, el muestreo estratificado* y *el muestreo por conglomerados.*

El siguiente esquema permite apreciar los distinto tipos de muestreo estudiados en este curso.



Muestreo aleatorio simple

Uno de los métodos más comúnmente usados para recolectar datos es el muestreo aleatorio simple.

Muestreo aleatorio simple: es aquel en el que todos los elementos de la población tienen la misma probabilidad de ser elegidos.

Las muestras aleatorias simples pueden ser **sin reemplazo** y **con reemplazo**. En el primer caso, al extraer un individuo de la población, no se devuelve a ésta. En el segundo caso, una vez registrado el individuo seleccionado, se devuelve a la población pudiendo ser elegido de nuevo.

Así pues, en un muestreo con reemplazo la muestra puede tener duplicados de la población. Esto es conveniente en poblaciones infinitas, puesto que en ellas, existen «muchísimas» unidades que ofrecen la misma respuesta para la variable, y, por tanto, al registrar un valor determinado quedan aún «infinidad» de unidades que darán la misma respuesta. Sin embargo, en el muestreo de poblaciones finitas, el muestreo de una persona que se repite dos veces no proporciona más información. Por lo general se prefiere el muestreo sin reemplazo, de modo que la muestra no contenga duplicados.

El muestreo con reemplazo, también se llama **con repetición** y el sin reemplazo, **sin repetición.**

Para elegir una muestra aleatoria simple, primero se asigna un número a cada elemento del marco muestral. Esto suele hacerse de manera secuencial usando la misma cantidad de dígitos para cada elemento. Luego, se consulta una tabla de números aleatorios y se eligen tantos números con esa cantidad de dígitos como sea necesario para obtener el tamaño de muestra deseado. Cada elemento numerado en el marco muestral que corresponda a un número aleatorio seleccionado, se elige para la muestra.

Ejemplo 1

Sugerir una forma para seleccionar una muestra aleatoria de 10 alumnos:

a. De tu grupo
b. De toda la escuela
c. Describir alguna manera incorrecta de seleccionar la muestra y decir por qué.

Solución

a. Basta con asignar números a los alumnos de tu grupo empezando con 00, 01, 02, 03,... Enseguida se extraen pares de dígitos aleatorios de alguna tabla, hasta completar los 10 alumnos de la muestra. (En caso de que un número aleatorio, por ejemplo 98, no corresponda a algún compañero, no se toma en cuenta).

b. De la misma forma que en el inciso (a) sólo sería necesario asignar números a toda la escuela.

c. Una manera incorrecta sería elegir únicamente de entre los que se encuentran en la biblioteca. Estos alumnos serán por lo general más estudiosos que los demás y también podrían diferir en otras características.

¿Cómo llamarías a este tipo de selección de muestra?_____
_____

Ejemplo 2

Supóngase que en un salón de clase determinado hay 50 butacas y se desea obtener una muestra de 4 elementos para conocer su deterioro.

Solución

27767	43584
13025	14338
80217	36292
10875	62004
54127	57326
60311	42824
49739	71484
78626	51594
66692	13986
44071	28091

Paso 1. Se enumeran las butacas de 01 a 50. (Atención: puesto que el tamaño de la población es de dos dígitos (50), la numeración de los elementos debe ser de dos dígitos (por eso se escribe 01).

Paso 2. De manera aleatoria se selecciona una porción de la tabla, una columna y un renglón. Supongamos que en la tabla anexa, el comienzo fue a partir del segundo bloque y tercer renglón. Es decir a partir del número 36292.

Paso 3. Ahora, puesto que el tamaño de la población es un número de dos dígitos (50), los dígitos se escogen de dos en dos. Entonces los números escogidos (siguiendo hacia abajo) son:

36, ~~62~~, ~~57~~, 42, ~~71~~, ~~51~~, 13, 28.

Obsérvese que se descartan el 62, 57, 71 y 51, porque no pertenecen a la población.

Actividad 1.4 b

Selecciona de manera aleatoria a 10 compañeros (as) de tu grupo y pregúntales cuántas letras tiene su primer nombre. Usa la tabla de números aleatorios construida en la actividad (1.4 a).

Registra aquí tus datos:

El muestreo aleatorio simple, es el método fundamental para seleccionar una muestra. Sin embargo, existen variantes del muestreo aleatorio simple que, bajo ciertas condiciones, pueden ser más eficaces. Entre los procedimientos de muestreo, alternativos al muestreo aleatorio simple están: el sistemático, el estratificado y el de conglomerados.

Muestreo aleatorio sistemático

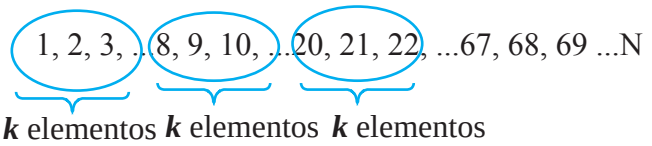
Recomendable para poblaciones grandes, heterogéneas y ordenadas aleatoriamente.

Para obtener una muestra sistemática, se procede como se indica a continuación:

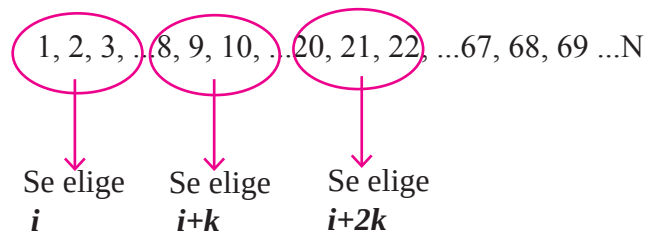
Paso 1. Conocido ***N*** (el tamaño de la población) y ***n*** (el tamaño de muestra), se particionan las ***N*** unidades ordenadas de la población en ***n*** grupos de tamaño ***k;*** es decir, se calcula el cociente:

$$k = \frac{N}{n}$$

(En caso de resultar decimal, redondear al entero próximo)



Paso 2. Ahora, del primer grupo seleccionamos aleatoriamente una unidad, digamos que es la identificada con ***i***; entonces del segundo se tomará la identificada con el número ***i + k***; del tercero ***i + 2k***; del cuarto ***i + 3k;*** del quinto ***i + 4k*** y así sucesivamente hasta completar ***n.***



Ejemplo

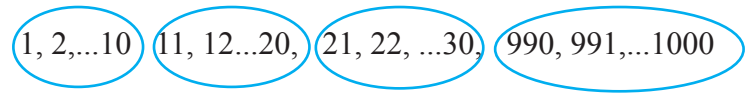
De una preparatoria con 1000 estudiantes, se desea obtener una muestra sistemática de tamaño 100. ¿Qué elementos deben seleccionarse?

Solución

Paso 1. Para particionar la población, calculamos $k = \frac{N}{n}$.

$$N = 1000, \quad n = 100 \longrightarrow k = \frac{1000}{100} = 10$$

El primer grupo lo forman las unidades 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; el segundo las unidades 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 y así sucesivamente.



10 elementos en cada grupo

Ejemplo
(Cont.)

Ahora, del primer grupo seleccionamos aleatoriamente una unidad, digamos que es la identificada con el número 3; entonces del segundo se tomará la identificada con el número 3 + 10; del tercero el 13 + 10; del cuarto el 23 + 10.....y del último 983 + 10.

El 3 se elige *aleatoriamente*

$3 + \boldsymbol{k} = 3 + 10$ $3 + \boldsymbol{2k} = 3 + 20$ $3 + \boldsymbol{3k} = 3 + 30$

Los estudiantes que deben seleccionarse son los que estén numerados con:

3, 13, 23, 33, 43, 53, 63, ...993.

Muestreo estratificado

Con frecuencia, tenemos información adicional que nos ayuda a diseñar nuestra muestra. Por ejemplo, antes de realizar una encuesta sobre el valor catastral de una vivienda, sabemos de antemano que en colonias residenciales dicho valor será muy superior a las de colonias populares, o que los sueldos de personas especializadas en una gran empresa, difieren de los de los obreros.

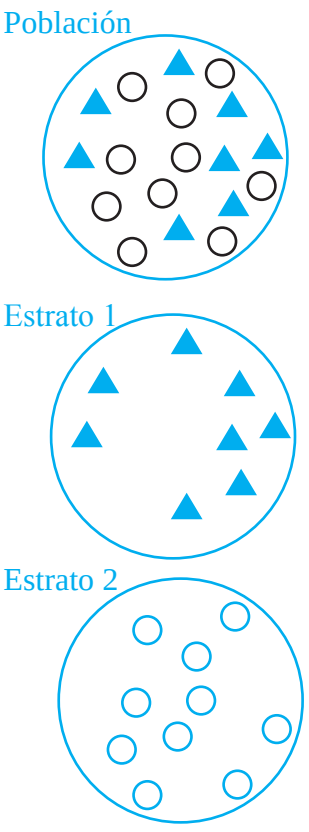
Las dos poblaciones señaladas en el párrafo anterior, son muy heterogéneas en las variables implicadas, a saber: valor catastral y sueldos.

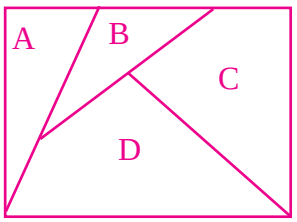
Si la variable que nos interesa asume valores muy distintos en diferentes subpoblaciones, se podrían obtener estimaciones más precisas de las cantidades de la población al tomar una **muestra aleatoria estratificada.**

Así pues, el muestreo estratificado puede ser más efectivo si se trata de poblaciones heterogéneas. En este tipo de poblaciones, los individuos deben agruparse según características homogéneas, por ejemplo clase social, edad, sexo, etc. Al hacer la estratificación, los grupos se establecen de modo que las unidades de muestreo tiendan a ser homogéneas dentro de cada estrato, y los estratos tiendan a ser diferentes entre sí.

Para obtener una muestra aleatoria estratificada, se pocede como sigue:

Primero, se divide la población en subpoblaciones relativamente **homogéneas,** llamadas estratos; los estratos no se traslapan y deben conformar la población completa, de modo que cada unidad de muestreo pertenece exactamente a un estrato.





No hay traslape entre A, B, C, D

Segundo, se hace un muestreo aleatorio en cada estrato y, posteriormente reunimos la información para obtener las esimaciones globales de la población. Existen dos criterios para hacer esta selección:

Selección simple. Es el más sencillo, pero el menos recomendable; consiste en repartir la muestra total en partes iguales para cada estrato

Selección proporcional. Se obtiene dividiendo la muestra total en partes proporcionales a la población de cada estrato.

Ejemplo

Se quieren conocer los ingresos semestrales por concepto de exámenes extraordinarios en las 37 preparatorias de la Universidad Autónoma de Sinaloa. Para ello, se estudiarán 10 preparatorias. ¿Qué procedimiento de muestreo es más adecuado?

Solución

Puesto que se sabe de antemano que algunas preparatorias tienen más de 1000 alumnos y otras menos de 500, se puede dividir la población formada por todas las preparatorias de la UAS, en tres estratos: preparatorias grandes (con más de 1200 alumnos), medianas (entre 500 y 1200 alumnos) y chicas (con menos de 500 alumnos). El número de preparatorias en cada una de estas categorías (estratos), es el siguiente:

Estrato	Número de preparatorias
Chica	12
Mediana	11
Grande	14
Total	37

Una vez establecidos los estratos, el siguiente paso será un plan de muestreo de manera que cada grupo quede representado proporcionalmente. La tabla siguiente muestra el porcentaje correspondiente a cada estrato:

Estrato	Número de preparatorias	Proporción	Unidades por estrato
Chica	12	$12/37 = 0.324$	3
Mediana	11	$11/37 = 0.297$	3
Grande	14	$14/37 = 0.378$	4
Total	37	0.999	10

Entonces, se estudiarán 3 preparatorias chicas, 3 medianas y 4 grandes.

Las unidades por estrato se determinan multiplicando cada proporción por el tamaño de muestra.

$0.324 \times 10 = 3.32 \approx 3$
$0.297 \times 10 = 2.97 \approx 3$
$0.378 \times 10 = 3.78 \approx 4$

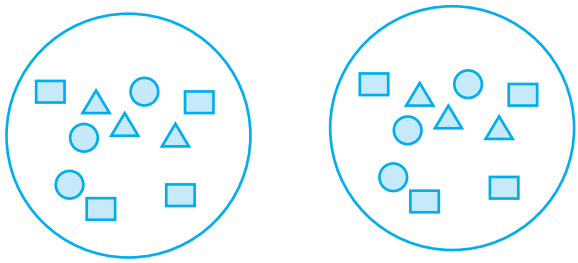
Muestreo por conglomerados

En los tres procedimientos de muestreo estudiados hasta el momento, se parte de que es fácil la enumeración de la población. Esto permite tomar de toda la población o de todos los estratos, la muestra de unidades que serán investigadas directamente. Sin embargo, existen situaciones en las que no existe ninguna lista utilizable de unidades para ser enumeradas y de la cual seleccionar la muestra. De aquí, surge la necesidad de seleccionar grandes unidades o conglomerados en vez de seleccionar elementos directamente de la población.

El procedimiento a seguir, es el siguiente:

1. **Identificar subdivisiones posibles de la población.** Estas subdivisiones se denominan conglomerados, y a menudo ocurren de manera natural. Para lograr los mejores resultados, las diferencias entre los conglomerados se hacen tan pequeñas como sea posible, en tanto que las diferencias entre los elementos individuales dentro de cada conglomerado se hacen tan grandes como sea posible. Lo ideal sería que cada conglomerado fuera una miniatura de toda la población.

En el muestreo por conglomerado, cobran importancia las llamadas unidades de muestreo. Por ejemplo, si se estudia a una escuela grande, la primer unidad de muestreo puede ser cada módulo, la segunda cada salón de clase y la tercera cada alumno.



Heterogeneidad dentro del conglomerado

Homogeneidad entre conglomerados

2. **Tomar una muestra aleatoria de conglomerados, y analizar a cada individuo perteneciente a los conglomerados seleccionados.** Obsérvese que en este procedimiento, los elementos individuales de la población, sólo participarán en la muestra, si pertenecen a un conglomerado incluido en la muestra.

Ejemplo

El director de cierta preparatoria, quiere estimar el número de butacas en mal estado de su escuela. Puesto que no existe una lista de todas las butacas que le permitan realizar un muestreo aleatorio simple, y puesto que él sabe que en dicha preparatoria hay 30 aulas cada una con aproximadamente 50 butacas, decide aplicar un muestreo por conglomerados. Para ello, elige 5 aulas al azar y procede a revizar cada una de las butacas de dichas aulas.

Actividad 1.4 c

Estudia atentamente la siguiente comparación entre los muestreos probabilísticos.

Comparaciones entre los muestreos probabilísticos

Muestreo aleatorio simple

Se extrae una muestra aleatoria de toda la población

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72

Muestreo aleatorio sistemático

A partir del primer elemento elegido aleatoriamente del intervalo [1, k], donde ***k*** es el entero más próximo a $k = \frac{N}{n}$, se eligen los elementos de ***k*** en ***k***.

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72

Muestreo estratificado

La población se divide en estratos. Se extrae una muestra aleatoria de cada estrato.

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72

Muestreo por conglomerados

La población se divide en conglomerados. Se extrae una muestra aleatoria simple de conglomerados. Se seleccionan todos los elementos incluídos en los conglomerados elegidos.

			1	2	3					1	2
			4	5	6					3	4
			7	8	9					5	6
			10	11	12						
								1	2		
								3	4		

AUTOEVALUACIÓN (UNIDAD I)

1. Debes tener una idea clara sobre el significado de los siguientes términos estadísticos:

Estadística	Dato
Proceso estadístico	Estudio observacional
Estadística descriptiva	Experimento
Estadística inferencial	Variable cualitativa
Población	Variable cuantitativa
Individuo	Marco muestral
Muestra	Muestreo de juicio
Muestra representativa	Muestreo probabilístico
Parámetro	Muestreo aleatorio simple
Estadígrafo	Muestreo sistemático
Variable	Muestreo por conglomerado.

Forma un diccionario con cada uno de estos términos. Inventa ejemplos.

2. Responde "verdadero" si la proposición siempre es verdadera. Si la proposición no siempre es verdadera, sustituya las palabras en negritas por las que hagan siempre verdadera la proposición.

   a) La estadística **inferencial** es el estudio y descripción de datos que resultan de un experimento.

   b) Una **población** típicamente es una colección muy grande de individuos u objetos sobre los que se desea tener información.

   c) Un parámetro es la medida de alguna característica de una **muestra**.

   d) El objetivo fundamental de la **estadística** es obtener una muestra, analizarla y luego hacer inferencias sobre las características desconocidas de la población de la cual se obtuvo la muestra.

3. Clasificar cada una de las siguientes variables como A) cualitativa, B) cuantitativa

   a) Método de pago para hacer las compras (en efectivo, con tarjeta de crédito o con cheque).

   b) Código postal del hogar del cliente.

   c) Cantidad de impuesto por venta en la compra.

   d) Número de artículos comprados.

4. La calificación promedio de todos los alumnos de la Preparatoria, debe estimarse utilizando la calificación promedio de 200 alumnos elegidos aleatoriamente. Haga corresponder las expresiones de la columna dos con los términos estadísticos de la columna uno.

_____ Dato	a) Los 200 alumnos.
_____ Datos (conjunto)	b) La calificación promedio de todos los alumnos.
_____ Experimento	c) 8, la calificación de un alumno.
_____ Parámetro	d) La calificación promedio de los 200 alumnos.
_____ Población	e) Todos los alumnos de la preparatoria
_____ Muestra	f) La calificación de un alumno.
_____ Estadígrafo	g) Las 200 calificaciones.
_____ Variable	h) El proceso utilizado para seleccionar a los 200 alumnos y obtener sus calificaciones.

5. Analiza el siguiente reporte de investigación médica, y contesta lo que se indica:

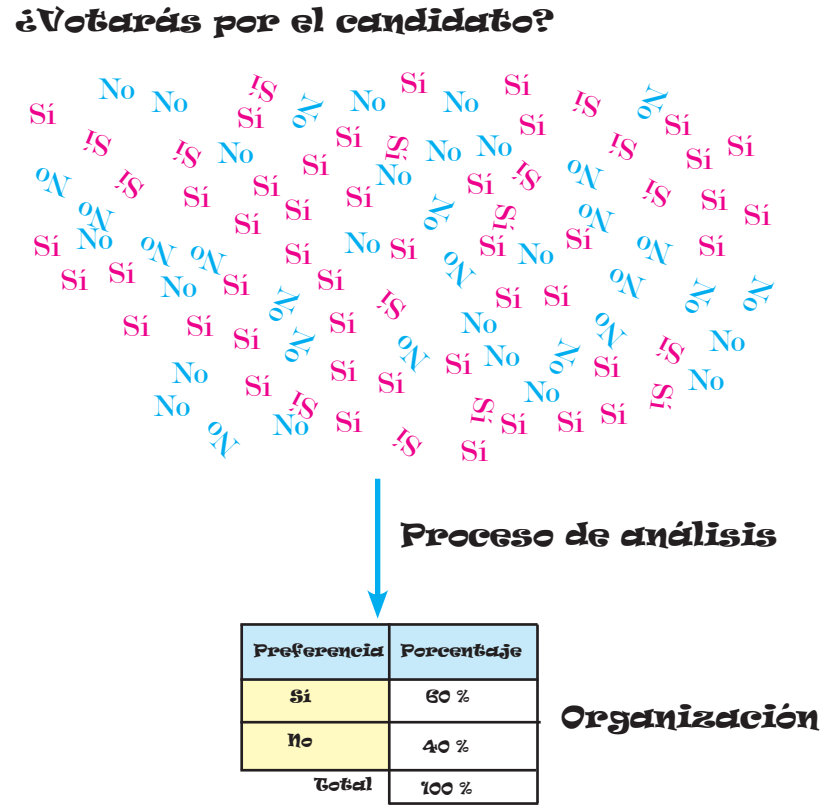
« *PLAVIX, agregado a aspirinas y tus medicinas actuales, ayuda a subir tu protección contra ataques del corazón o apoplejía. Sin embargo, resultados de un experimento clínico que incluyó a 12, 000 sujetos, indican que el riesgo de sangrado puede incrementarse con PLAVIX. Entre los 6259 individuos que tomaron PLAVIX + aspirina, 3.7 % mostró problemas de sangrado, mientras sólo 2.7 % de los 6303 que tomó placebo tuvo mayor sangrado.*

a. ¿Cuál es la población de interés en esta investigación?
b. ¿Cuál es la muestra?
c. Los resultados son estadígrafos o parámetros?
d. ¿La información acerca de efectos secundarios de PLAVIX indica que el tratamiento es confiable?

Exploración de datos cualitativos

2

UNIDAD



Lección 2.1 Clasificación de variables

Objetivos: Aprender a reconocer variables nominales, ordinales, discretas y continuas.

Actividad 5

Qué hacer

1. Consulta las **páginas 43 a 45** y al finalizar tu estudio, clasifica las variables siguientes en nominal, ordinal, discreta o continua.

   a) Número de hijos de una familia.
   b) Total de puntos de un equipo al finalizar un torneo.
   c) Número de periódicos que se venden en una ciudad en un día.
   d) Litros de precipitación en una ciudad en cuatro meses.
   e) Tiempo que se necesita para resolver un problema.
   f) Dinero total gastado en libros en el año escolar.
   g) Tiempo diario que se pasa en el internet.
   h) Grado de confianza en el ejército.

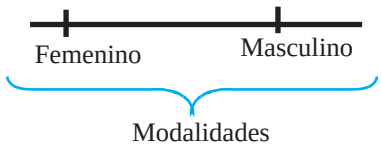
Antes de abordar la exploración de datos, precisaremos algunas cuestiones relacionadas con las variables.

1. Variables y sus valores, o categorías posibles

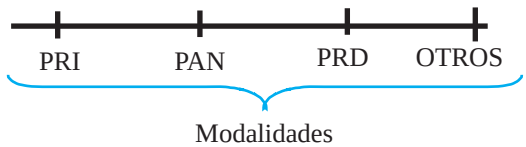
Modalidades. Para una variable cualitativa, las modalidades son las *posibles categorías distintas* que pueden ser asignadas a los individuos, y, para una variable cuantitativa, las modalidades son los *posibles valores distintos* que puede tomar los individuos.

Ejemplos

(a) Variable: Género

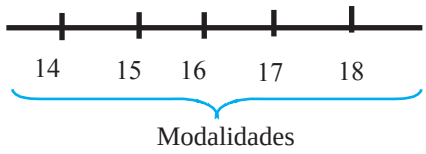


(b) Variable: Preferencia electoral.

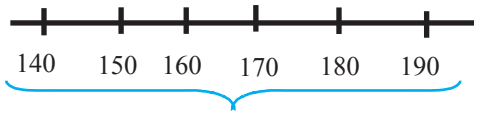


Ejemplos (Cont.)

(c) Variable: Edad (años)

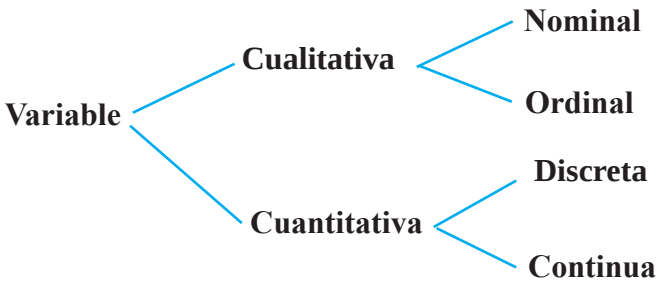


(d) Variable: Estaturas (cm).



2. Recordemos que hay dos clases de variables: cualitativas y cuantitativas.

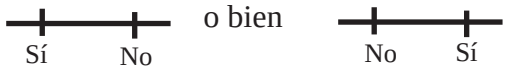
Cada uno de estos tipos de variables se subdivide aún más, tal y como se muestra a continuación.



Variable nominal. Es una variable cualitativa que produce datos que simplemente se clasifican en distintas categorías que no implican orden.

Ejemplo

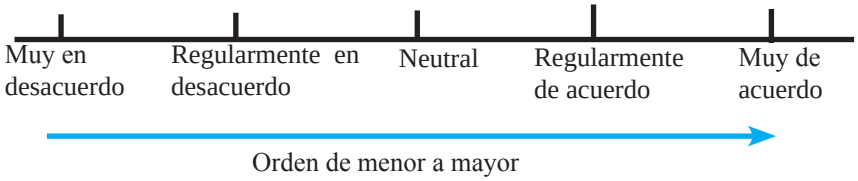
(a) Las posibles respuestas a la pregunta «¿está usted suscrito a algún periódico?», son: *sí* o *no*. Estas respuestas se pueden escribir:



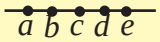
Es decir, no importa el orden.

Variable ordinal. Es una variable cualitativa que produce datos que se pueden clasificar en categorías distintas en las que existe algún orden o jerarquía.

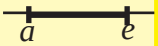
Ejemplo (a) Las posibles respuestas a la pregunta «*¿está usted de acuerdo con el horario de verano*?», pueden ser:



Valores de una variable discreta:
a, b, c, d, e, f



Valores de una variable continua:
$[a, e]$



Variable discreta. Es una variable cuantitativa que produce datos cuyos posibles valores son contables, o aislados en un intervalo. Es decir, entre dos valores cualesquiera siempre hay un hueco.

Ejemplo El número de hijos, el número de carros en un estacionamiento, son variables cuantitativas discretas.

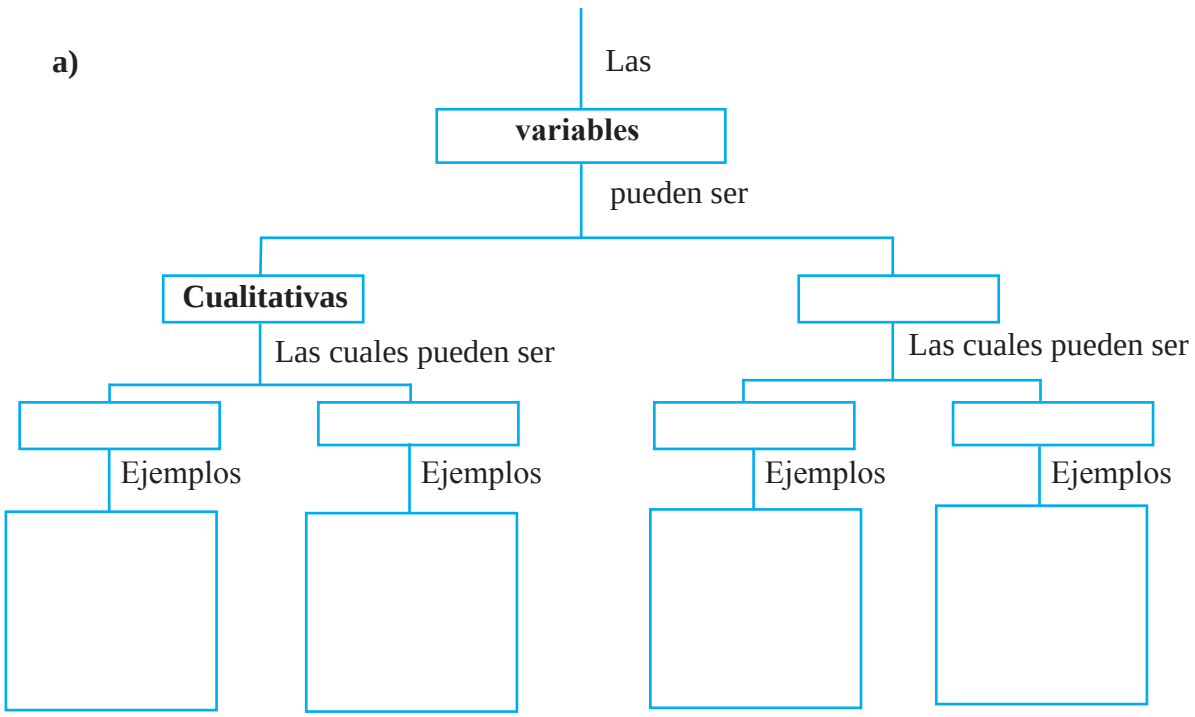
Variable continua. Es una variable cuantitativa que produce datos cuyos posibles valores son todos los valores de un intervalo y por lo tanto valores incontables.

Ejemplo El peso, la estatura de una persona, la temperatura. En general, todas las magnitudes relacionadas con el espacio, el tiempo, la masa o bien combinación de ellas son variables cuantitativas continuas.

Ejercicio 2.1

Completa el siguientes esquema:

a)



Lección 2.2 Exploración de datos cualitativos

Objetivos: Aprender los métodos de exploración de datos cualitativos.

Actividad 6

Qué hacer

1) Consulta las **páginas 47 a 51** y al finalizar tu estudio resuelve lo indicado.

Los datos siguientes, muestran las preferencias partidistas de 200 ciudadanos para la elección de diputado federal del primer distrito de Sinaloa. Estos datos, fueron adaptados de una encuesta publicada en el periódico El Debate de Culiacán el 2 de junio de 2009. Con respecto a la obtención de la muestra, el reporte informa lo siguiente: se hicieron 300 encuestas; el muestreo fue aleatorio con base a las secciones electorales de cada distrito previamente estratificado por el criterio urbano y rural. Se seleccionaron 20 puntos de levantamiento y se aplicaron 15 encuestas por punto muestral. Las entrevistas se realizaron de manera personal en viviena; tanto la vivienda como el entrevistado se seleccionaron de forma sistemática en cada punto.

PRI, PRI, PRD, PRI, PAN, PAN, PRI, PRI, PAN. PAN. PRD, PRI, PAN, PRD, PAN, PAN, PAN, PAN, PRI, PAN, PAN, PRI, PAN, PVEM, PAN, PRI, PRD, N. Alianza, S. a México, Ns/Nc. PRI, PRI, PRI, PAN, PAN, PRD, PRI, PAN, PAN, PRI, PRI, PAN, PAN, PAN, PRI, Ns/Nc, N. Alianza, PRI, PAN, PAN, PRI, Ns/Nc, Ns/Nc, N. Alianza, PRI, PAN, PAN, PAN, PAN, PRI, PAN, PRI, PRI, PRI, Ns/Nc, PAN, PRI, PAN, PAN, PRD, PAN, PRI, PRD, Ns/Nc, PAN, PAN, PAN, PRD, PRI,, PAN, PAN, PRI, PRI, PRI, PRI, PAN, PAN, PRI, Ns/Nc, PRI, PRI, PAN, PAN, PRI, PRI, PRI, PRI, PRI, PRI, PAN, PAN, PRI, N. Alianza, Ns/Nc, N. Alianza, PRI, PRI, PRI, PAN, PAN, PAN, PAN, PRD, PVEM, Ns/Nc, PAN, PAN, PRI, PAN, PRI, PRI, PRI, PRI, PAN, PAN, PRI, PRI, PRI, Ns/Nc, PRI, PRI, PAN, PAN, PAN, PRI, PRI, PRI, PRI, Ns/Nc, PRI, PRI, PAN, PRI, PRI, PRI, Ns/Nc, PRI, PRI, PRI, PAN, PRI, PAN, PAN, PAN, PAN, PRI, PRI, PAN, Ns/Nc, PRI, PRI, PRI, PAN, PRI, PRI, PAN, PAN, PAN, Ns/Nc, PRI, PRI, PAN, PRI, PRI, PAN, PRI, Ns/Nc, PRI, PAN, PRI, PRI, PAN, PRI, PRI, PAN, Ns/Nc, PVEM, PAN, PAN, PAN,, PRI, PRI, PAN, PAN, PRI, PRI, PRI, PRI, PAN, PAN.

Haz una exploración de los datos, siguiendo las fases:

a) Organiza los datos
b) Representa los datos de manera conveniente.
c) Identifica la categoría modal.
d) Interpreta los resultados.
e) Según el procedimiento seguido para obtener la muestra, ¿qué se puede concluir acerca de su representatividad?

Fases de la exploración de datos:

Etapa de análisis

- Organización.
- Representación gráfica.
- Cálculo de medidas de resumen.

Etapa de interpretación

- Valorar la representatividad de la muestra.
- Establecer conclusiones teniendo en cuenta el contexto.

Una vez recolectados los datos, debemos buscar la información que ellos contienen. Por lo general, los datos obtenidos estarán desorganizados y serán difíciles de comprender. De alguna manera necesitamos presentar los datos de una forma adecuada que nos permita extraer información útil para los propósitos deseados. A este proceso de transformación de los datos se le llama **análisis de datos.**

El **análisis de datos** se divide en tres fases:

1. Organización de los datos.
2. Representaciones gráficas de los datos.
3. Cálculo de medidas de resumen.

El objetivo del análisis es transformar los datos de tal manera que se destaque información importante y se descubran patrones y tendencias.

El foco principal de este curso, está en el desarrollo de herramientas estadísticas del análisis de datos. Aprenderás a organizar datos en tablas, a hacer gráficas de los datos y a calcular algunas medidas estadísticas. A este proceso de análisis, junto con el proceso de interpretación de resultados, le llamaremos **exploración de datos.** En lo que resta de este texto, podrás estudiar la exploración de datos, empezando con los datos cualitativos.

Fase 1. Organización de datos

¿Cómo organizar los datos?

El siguiente ejemplo ilustra el proceso de organización de datos cualitativos.

Ejemplo

Con objeto de determinar la preferencia de fase especializada de 30 estudiantes, en el ciclo 2006-2007, se realizó una encuesta que produjo los siguientes resultados :

CS, FM, CS, CS, QB, CS, CS, FM, QB, CS, CS, CS, CS, QB, CS, CS, QB, FM, CS, CS, QB, QB, FM, CS, CS, CS, CS, QB, QB, CS.

Abreviaturas:
FM: *Físico-matemáticas*
CS: Ciencias sociales y humanidades
QB: *Químico-biológicas*

Organizar los datos en *una tabla de frecuencias*

Frecuencia absoluta o simplemente **frecuencia**, denotada como ***f***, es el número de veces que aparece cada categoría o valor.

Procedimiento:

Para variables cualitativas, formamos una tabla provisional con tres columnas y procedemos como sigue:

a) Las distintas categorías que toma la variable se colocan en la primera columna.

b) Se realiza el conteo de las veces que aparece cada categoría. Para ello, se recorre la lista original en el orden en que está dada y se hacen marcas en la segunda columna enfrente de la categoría correspondiente (conteo).

c) En la tercera columna se indica el número de veces que cada modalidad aparece, este número se llama frecuencia absoluta (f) o simplemente frecuencia.

Fase especializada	Conteo	Frecuencia (*f*)
Ciencias sociales	𝍸 𝍸 𝍸 ///	18
Químico-biológicas	𝍸 ///	8
Físico-matemáticas	////	4
Total		30

Esta tabla se llama ***distribución de frecuecias absolutas*** *o simplemente* ***distribución de frecuencias.***

Una **distribución de frecuencias** de una variable cualitativa, es una descripción del número de veces (es decir de las fecuencias) con que se presentan cada una de las diversas modalidades o categorías que corresponden a esa variable.

Además de las frecuencias absolutas, existen las frecuencias relativas que se definen como sigue:

$$\textit{Secuencia relativa} = f_r \quad \frac{\textit{frecuencia absoluta}}{\textit{tamaño de la muestra}} = \frac{f}{n}$$

Llamaremos **tabla de frecuencias** para datos cualitativos, a aquella que muestre las frecuencias absolutas, frecuencias relativas y porcentajes. La tabla de frecuencias del ejemplo anterior se muestra a continuación. Los porcentajes se obtienen multiplicando las frecuencias relativas por 100.

Fase especializada	Frec. abs.	Frec. rel.	Porcentaje
Ciencias sociales	18	0.60	60%
Químico-biológicas	8	0.27	27%
Físico-matemáticas	4	0.13	13%
Total	30	1.00	100%

El cálculo de los **porcentajes,** también puede hacerse mediante proporciones:

$$\textit{Si } 30 \longrightarrow 100\%$$
$$18 \longrightarrow x$$
$$x = \frac{18 \times 100\%}{30} = 60\%$$

$$\textit{Si } 30 \longrightarrow 100\%$$
$$8 \longrightarrow x$$
$$x = \frac{8 \times 100\%}{30} = 27\%$$

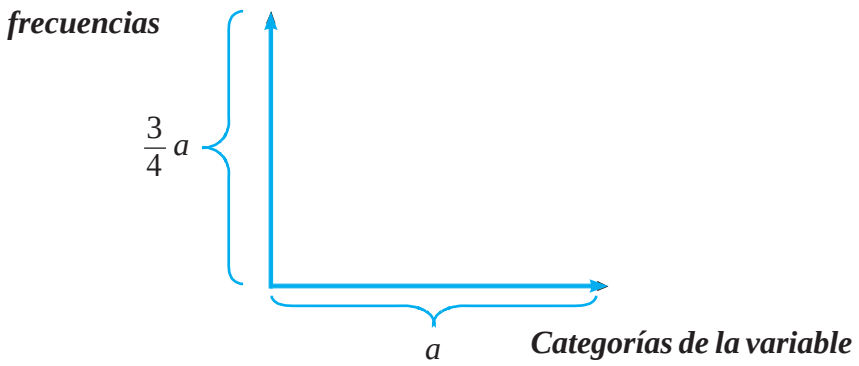
Fase 2. Representación gráfica

Una de las etapas más importantes de la exploración de datos, es la representación gráfica de éstos. Las representaciones gráficas deben conseguir que un simple análisis visual ofrezca la mayor información posible. El valor de las gráficas radica en el hecho de que permiten apreciar la situación de un grupo o un individuo con mayor rapidez y de forma más intuitiva que las representaciones numéricas. Las gráficas facilitan la búsqueda de patrones particulares en los datos.

Para representar datos cualitativos, existen muchos tipos de gráficas. Las más usuales son el **gráfico de barras** y el **gráfico circular** o **de sectores.**

Gráfico de barras

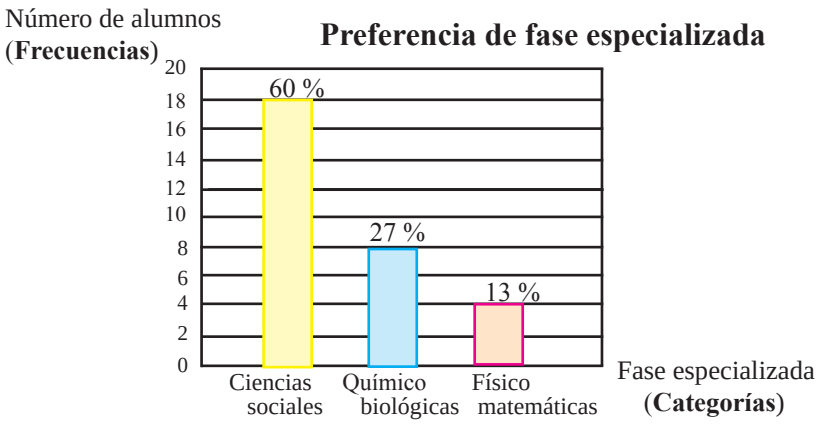
En la lección (1.1), ya estudiaste la manera de construir gráficos de barras. Sobre los ejes cartesianos distribuimos en el eje de las abscisas las modalidades de la variable. Sobre cada categoría, se levantan barras o rectángulos de igual base (que no se traslapen) cuya altura sea igual a la frecuencia absoluta, frecuencia relativa o porcentaje, que corresponde a cada categoría. Se recomienda, siempre que sea posible, que el eje de ordenadas sea aproximadamente una cuarta parte más pequeño que el de las abscisas.



Para efecto de una división apropiada del eje vertical, se aproxima la frecuencia más alta de la distribución, al número inmediatamente mayor que sea divisible entre dos. Es recomendable escribir en la parte superior de las barras sus porcentajes.

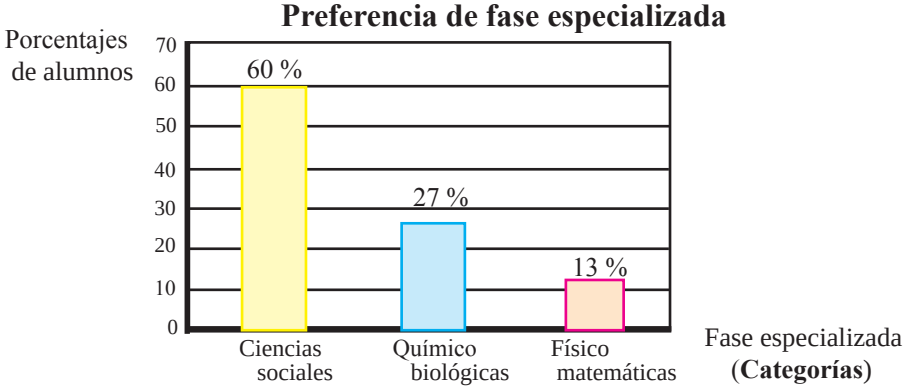
Todas las representaciones gráficas deben ser completamente autosuficientes. Esto incluye un título descriptivo significativo, identificación de las escalas vertical y horizontal, y la fuente.

Ejemplo Construiremos ahora, el diagrama de barras con base en la tabla de frecuencias correspondientes a los datos de elección de fase especializada.



Fuente: Datos obtenidos en el ciclo 2006-2007

Recordemos que en el eje vertical, también podemos escribir las frecuencias relativas o porcentajes.



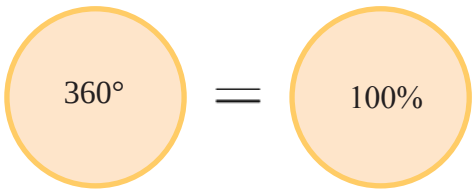
Fuente: Datos obtenidos en el ciclo 2006-2007

Gráfico circular o de sectores

Son gráficos en los que a cada categoría se le asigna un sector circular de área proporcional a la frecuencia que representan. Los gráficos circulares son buenos para mostrar los tamaños relativos de grupos. Varios grupos diferentes pueden ser representados y comparados con este tipo de diagramas. Los gráficos circulares son particularmente adecuados para variables cualitativas.

Para el dibujo, necesitamos la equivalencia existente entre porcentajes y grados, para ello, partimos de lo siguiente:

El ángulo que describe una circunferencia mide 360°, o sea, el círculo es el sector circular cuyo ángulo mide 360°; y por otra parte la suma de todos los datos de una distribución determinada equivale al 100% . En consecuencia todo el círculo equivale al 100% y esto nos permite establecer una relación entre grados y porcientos.



Por lo tanto: $360^{o} = 100\%$

O bien: $100\% = 360^{o}$

$100 \times 1\ \% = 360^{o}$

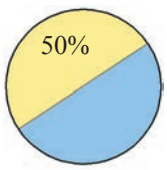
$1\% = 360^{o}/100$

$1\% = 3.60$ GRADOS

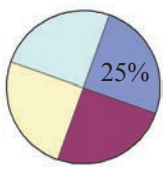
Entonces, para convertir porcentajes a grados, multiplicamos por 3.6

Ejemplos

$50\% = 50 \times 1\%$
$= 50 \times 3.6° = 180°$



$25\% = 25 \times 1\%$
$= 25 \times 3.6° = 90°$



Procedimiento para dibujar un diagrama circular.

1. Multiplicar los porcentajes por 3.6; esto da la medida angular del sector representativo de cada porcentaje.
2. Se traza una circunferencia de radio arbitrario, en función del espacio disponible.
3. Se traza un radio y a partir de él se miden con un transportador los grados correspondientes a cada sector yendo del mayor al menor.
4. Terminado el punto anterior se escriben en cada sector los datos porcentuales correspondientes; luego se anexa el título y la fuente y demás indicaciones necesarias para hacer comprensible la gráfica.

Ejemplo Construiremos ahora, el gráfico circular con base en la tabla de frecuencias correspondientes a los datos de elección de fase especializada.

Fases especializadas	Frec. abs.	Frec. rel.	Porcentaje
Ciencias sociales	18	0.60	60%
Químico-biológicas	8	0.27	27%
Físico-matemáticas	4	0.13	13%
Total	30	1.00	100%

Anexar una columna a la tabla anterior y anotar en ella, la equivalencia de cada porcentaje a grado.

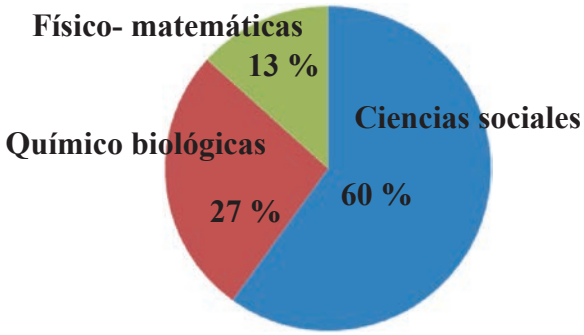
Fases especializadas	Frec. abs.	Frec. rel.	Porcentaje	Grados
Ciencias sociales	18	0.60	60%	$60 \times 3.6^\circ = 216^\circ$
Químico-biológicas	8	0.27	27%	$27 \times 3.6^\circ = 97.2^\circ$
Físico-matemáticas	4	0.13	13%	$13 \times 3.6^\circ = 46.8^\circ$
Total	30	1.00	100%	360°

El cálculo de los **grados** para cada sector, también puede hacerse mediante proporciones:

$$Si\ 100\% \longrightarrow 360^\circ$$
$$60\% \longrightarrow x$$
$$x = \frac{60\% \times 360\%}{100\%} = 216^\circ$$

$$Si\ 100\% \longrightarrow 360^\circ$$
$$27\% \longrightarrow x$$
$$x = \frac{27\% \times 360^\circ}{100\%} = 97\%$$

Finalmente, con la ayuda de un transportador dibujamos el gráfico circular:



Fase 3. Cálculo de medidas de resumen.

En el caso de datos cualitativos, la única medida de resumen, es la ***moda*** o ***categoría modal,***

La moda o categoría modal, es la categoría que más repite, es decir, es la categoría que aparece con mayor frecuencia.

En nuestro ejemplo, la moda es: *Ciencias sociales.*

Fase 4. Interpretación de resultados.

El estudiante típico de esta muestra, prefiere la especialidad en ciencias sociales. La siguiente especialidad que más se solicita es la químico biológicas, y la menos requerida es físico matemáticas. Estos resultados coinciden con las preferencias vocacionales que generalmente se manejan en distintos medios. Sin embargo, debemos ser concientes que los porcentajes obtenidos en este ejemplo, podrían diferir con varios puntos con respecto a la población total, puesto que la muestra estudiada, no fue elegida bajo las condiciones que debe cumplir una muestra representativa.

Ejercicio 2.2

1. Considera los siguientes datos sobre el tipo de problemas de salud (J = articulación hinchada, F = fatiga, B = dolor de espalda, M = debilidad muscular, T = tos, N = nariz con flujo o irritación, O = otro) presentado por tres personas. Realiza la exploración de estos datos.

O	O	N	J	T	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	T
J	F	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	T	O	O	O	M	B	F
J	O	F	N										

2. Representa mediante un gráfico de barras y un gráfico circular la siguiente información.

   a) Las principales religiones en el mundo, en número de fieles, es el siguiente:

RELIGIÓN	NÚMERO
CRISTIANISMO	1 929 957 000
ISLAMISMO	1 147 494 000
HINDUISMO	747 797 000
BUDISMO	353 141 000
RELIGIONES TRIBALES	231 614 000
JUDAÍSMO	14 890 000
CONFUCIANISMO	6 112 000
NO RELIGIOSOS Y ATEOS	906 995 000

FUENTE: Organización de Naciones Unidas. 1997 (Paulo Afonso Lopes)

   b) Actividades realizadas durante el domingo por los católicos de Sinaloa. Febrero 2006.

ACTIVIDAD	NÚMERO DE PERSONAS
DESCANSAR	307
VER TELEVISIÓN	134
ASISTIR A MISA	198
VISITAR FAMILIARES	207
IR AL CINE	64
PASEAR	99
TRABAJAR	62
OTRAS	9
NS/NC	1

FUENTE: ***El Debate de Culiacán.***

3. Después de estudiar esta unidad, estás en posibilidades de explorar las variables cualitativas que planteaste en tu proyecto. Realiza esta exploración y presenta un reporte parcial.

Lección 2.3 Comparación de grupos: uso del gráfico de barras múltiples

Objetivo: Aplicar el conocimiento estadístico, en la comparación de grupos que involucran datos cualitativos.

Actividad 7

Qué hacer

1) Consulta las **páginas 54** a **56** y al finalizar tu estudio resuelve lo indicado.

Las tablas siguientes muestran los porcentajes de las preferencias partidistas por nivel académico, para la elección de diputados federales por cada uno de los ocho distritos de Sinaloa. Estos datos corresponden a la encuesta del periódico El Debate ya señalada en la lección anterior.

Distrito 1

	S/E	Univ. o más
PAN	36.4 %	34.3 %
PRI	50.0 %	34.3 %
PRD	0 %	8.6 %
OTROS	13.6 %	22.8 %

Distrito 2

	S/E	Univ. o más
PAN	12.5 %	25.0 %
PRI	75.0 %	46.9 %
PRD	6.2 %	7.8 %
OTROS	6.3 %	20.3 %

Distrito 3

	S/E	Univ. o más
PAN	25.0 %	49.1 %
PRI	50.0 %	32.7 %
PRD	8.3 %	5.5 %
OTROS	16.6 %	12.7 %

Distrito 4

	S/E	Univ. o más
PAN	47.1 %	52.5 %
PRI	41.2 %	29.5 %
PRD	11.7 %	4.9 %
OTROS	0 %	13.1 %

Distrito 5

	S/E	Univ. o más
PAN	27.3 %	35.9 %
PRI	59.1 %	44.7 %
PRD	0 %	5.8 %
OTROS	13.6 %	12.7 %

Distrito 6

	S/E	Univ. o más
PAN	42.9 %	27.7 %
PRI	19 %	40 %
PRD	0 %	1.5 %
OTROS	38.1 %	30.7 %

Distrito 7

	S/E	Univ. o más
PAN	9.1 %	26.6 %
PRI	18.2 %	48.1 %
PRD	9.1 %	3.8 %
OTROS	63.6 %	21.5 %

Distrito 8

	S/E	Univ. o más
PAN	13.3 %	27.3 %
PRI	40 %	31.8 %
PRD	6.7 %	5.7 %
OTROS	40 %	35.2 %

Abreviaturas
S/E Sin estudios
Univ. Universidad

En la categoría denominada «**OTROS**», están incluidas las preferencias por PVEM, N. Alianza, S. a México, y también quienes No Contestaron o manifestaron No saber.

Haz una comparación global en los ocho distritos, de las preferencias partidistas por nivel académico, tratando de dar respuesta a las siguientes preguntas:

a) ¿Por cuál partido votan, en general, los ciudadanos «*sin estudios*»?
b) ¿Por cuál partido votan, en general, los ciudadanos con «*estudios universitarios*»?
c) ¿Puede afirmarse que entre menos instrucción tenga el ciudadano, más se inclina por algún partido? _____

Utiliza los recursos gráficos que consideres convenientes para apoyar tus respuestas.

La comparación de dos o más grupos distintos con respecto a una misma característica, es con frecuencia uno de los objetivos de los estudios estadísticos. Para variables cualitativas, uno de los instrumentos gráficos que facilitan estas comparaciones, es el *gráfico de barras múltiples.*

Gráfico de barras múltiples

Estos gráficos se utilizan cuando las categorías de la variable estudiada, presentan a su vez, otras categorías o modalidades y quiere destacarse esa división.

En el diagrama de barras múltiples, se dibuja una barra para cada una de las categorías en que se divide la modalidad.

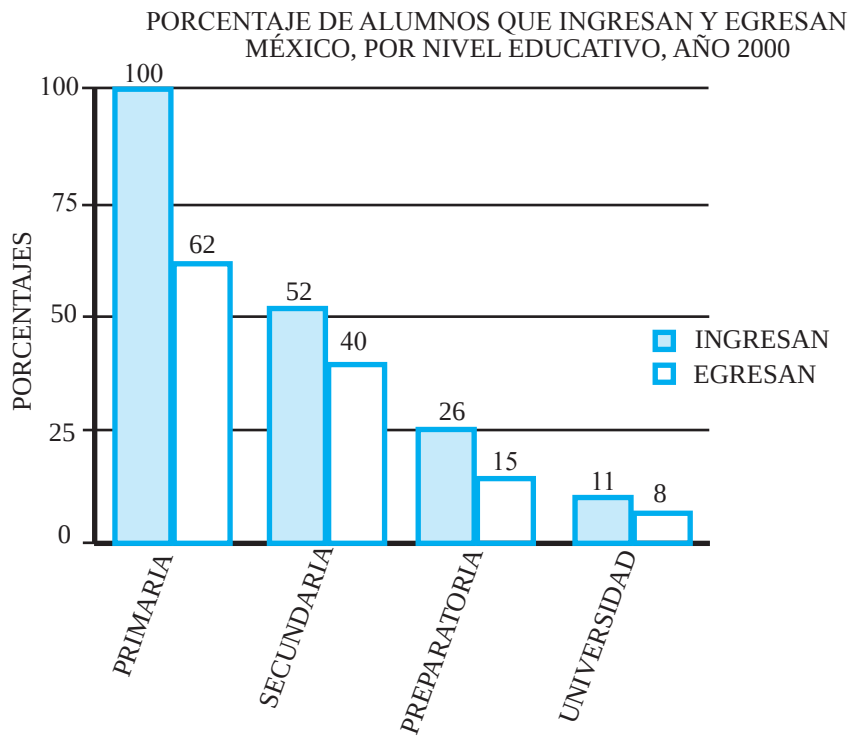
Ejemplo Construyamos el diagrama de barras múltiples para ilustrar los siguientes datos:

En México, de cada 100 niños que ingresan a primaria egresan 62, de éstos, 52 pasan a la secundaria y egresan 40, de los cuales 26 cursan el nivel bachillerato y egresan 15 para que a su vez sólo 11 de éstos llegan al nivel superior y egresan 8, de los cuales sólo 2 se titulan (datos del año 2000). Construya una tabla estadística. Fuente: *El Debate de Culiacán.*

A partir de estos datos, se construye la siguiente tabla:

NIVEL	INGRESO	EGRESO
PRIMARIA	100	62
SECUNDARIA	52	40
PREPARATORIA	26	15
UNIVERSIDAD	11	8

El gráfico de barras múltiples es:



Ejemplo Considera la siguiente información:

Según encuesta realizada en 2007 por el Centro de Ciencias de Sinaloa, sobre la variable «elección profesional» entre estudiantes de tercer año de preparatoria, las preferencias de carreras según área de especialidad es la siguiente:

Fases especializadas	Porcentaje
Ciencias sociales	50 %
Químico-biológicas	19 %
Físico-matemáticas	31 %
Total	100%

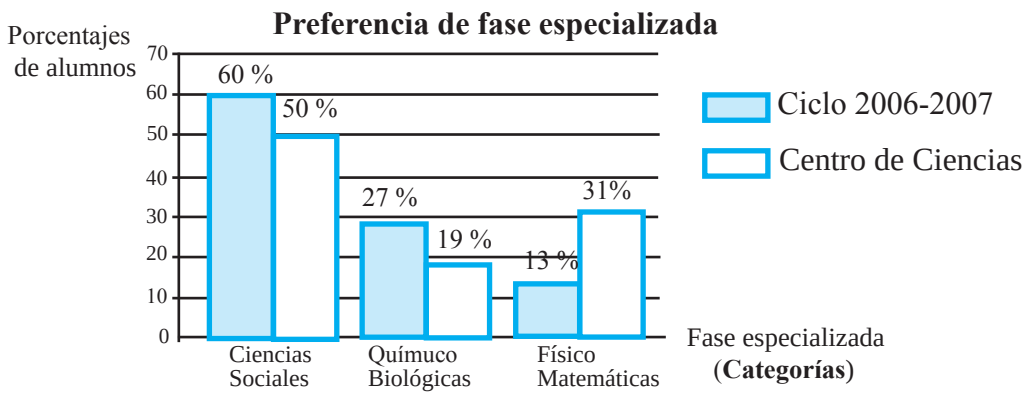
Utilizaremos el gráfico de barras múltiples para contestar la siguiente pregunta:

¿Hay diferencia entre los datos de la variable «preferencia de fase especializada» obtenidos el 2006 (y analizada en la lección 2.2), y los resultados correspondientes a esa misma variable, del Centro de Ciencias?

Construiremos una nueva tabla que nos permita hacer la comparación:

	Encuesta ciclo 2006-2007	Encuesta Centro de Ciencias
Fases especializadas	**Porcentaje**	**Porcentaje**
Ciencias sociales	60 %	50 %
Químico-biológicas	27 %	19 %
Físico-matemáticas	13 %	31 %
Total	100%	100%

A continuación construiremos el gráfico de barras múltiples.



Conclusión

Se aprecian diferencias entre ambas encuestas. Sin embargo hay coincidencia en que la mayor demanda es para ciencias sociales. Llama la atención que en la encuesta del Centro de Ciencias, se presentó gran demanda por Físico Matemáticas por encima de Químico Biológicas. Una explicación podría estar en la naturaleza de la muestra. Posiblemente ésta fue obtenida a partir de los visitantes a ese centro educativo.

Para tener más elementos sobre esta problemática, debes contestar la siguiente pregunta:

¿Hay diferencia entre los datos de la variable «preferencia de fase especializada» obtenidos el 2006, los resultados correspondientes a esa misma variable del Centro de Ciencias, y, resultados correspondientes a esa misma variable, pero de alumnos de este ciclo escolar?

Para contestar esta pregunta, realiza la siguiente actividad.

Actividad 2.3 a

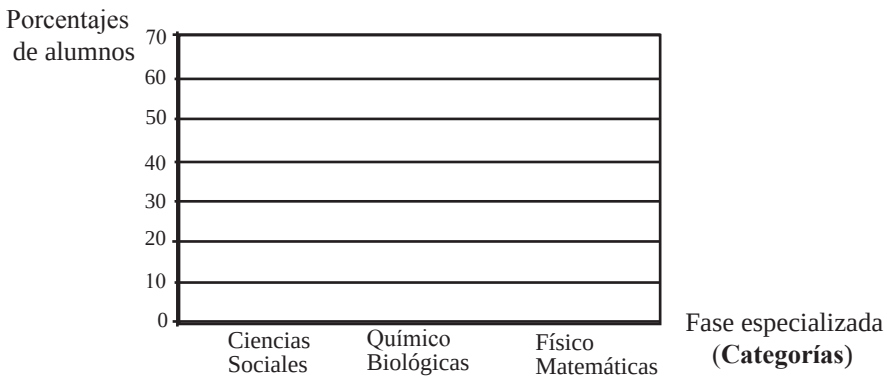
1. Realiza un censo en tu grupo para recolectar datos de la variable «*fase especializada elegida*». Llama a este grupo «*ciclo actual*».

Datos de la variable «*elección de fase especializada*» para el «ciclo actual».

2. Organiza los datos:

Categorías de fase especializada	Frec. abs.	Frec. rel.	Porcentaje
Ciencias sociales			
Químico-biológicas			
Físico-matemáticas			
Total			

3. Representa los datos en un gráfico de barras.



4. Integra en una misma tabla los resultados obtenidos en el ciclo 2006-2007, los del Centro de Ciencias, y, tus resultados del ciclo actual. Puesto que se van a comparar dos grupos, debemos utilizar los porcentajes.

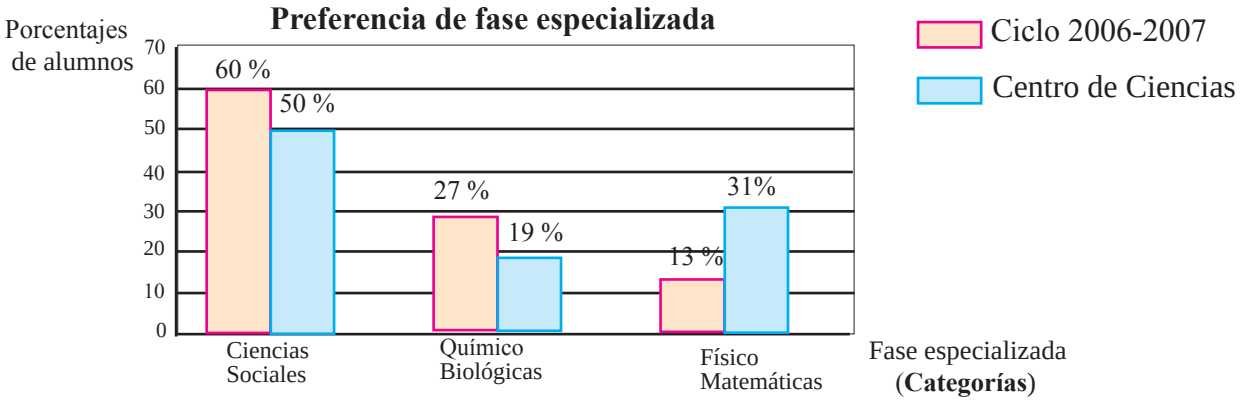
¿Por qué no es conveniente utilizar frecuencias absolutas cuando se hacen comparaciones?

Actividad 2.3 a (Cont.)

Fases especializadas	Encuesta ciclo 2006-2007 Porcentaje	Encuesta Centro de Ciencias Porcentaje	Encuesta Ciclo actual Porcentaje
Ciencias sociales	60 %	50 %	
Químico-biológicas	27 %	19 %	
Físico-matemáticas	13 %	31 %	
Total	100 %	100 %	

Anota tus datos

5. Ahora, con base en esta tabla, modifica el gráfico de barras múltiples que ya construimos. ¿Qué parte debe cambiarse?



6. Finalmente, debes hacer una comparación entre los tres grupos. Apóyate en las siguientes preguntas: ¿Hay variabilidad en los resultados de un grupo a otro? ¿Cuáles son las modas de cada grupo? ¿Qué te indica ésto? Argumenta.

Ejercicio 2.3

1. Completa la siguiente tabla y represéntala en un gráfico de barras múltiples. Realiza una comparación de las variables implicadas.

Población indígena en cuatro países de América

País	Porcentaje de indígenas	Porcentaje de no indígenas
México	10%	
Guatemala	60%	
Ecuador	60%	
Perú	25%	

2. Organiza la siguiente información en una tabla, represéntala en un gráfico de barras múltiples y haz una comparación de la evolución de la población a través de los años.

En Sinaloa, en 1910 la población total era de 323,642 habitantes, de los cuales 159,709 eran hombres y 163,933 mujeres; en 1930 de un total de 395 618 habitantes, 195 023 eran hombres y 200 595 mujeres; en 1950 había 315,877 hombres y 319,804 mujeres, mientras que en 1970 de un total de 1'266,528 habitantes, 646,561 eran hombres y 619,967 mujeres. En 1990 había 1'101, 621 hombres y 1'102,433 mujeres. En el año 2000 el total es 2'534, 835 con 1'257,681 hombres y 1'277,154 mujeres. En el año 2005 había 1'294,146 hombres y 1' 313,854 mujeres. Fuente: INEGI.

3. En una escuela, se realizó en dos grupos diferentes, una encuesta sobre el tipo de música favorita, con los siguientes resultados:

Grupo 1	
Música favorita	Frecuencia
Norteña	8
Banda	12
Rock	4
Total	24

Grupo 2	
Música favorita	Frecuencia
Norteña	5
Banda	11
Rock	14
Total	30

Compara los dos grupos. Puedes agregar los conceptos necesarios a las tablas y hacer gráficos adecuados, para efecto que dichas comparaciones sean más objetivas.

4. La siguiente tabla, muestra datos sobre natalidad y mortalidad en el estado de Sinaloa en 1980 y 2007, así como una proyección para 2012 y 2030. Construye un gráfico de barras múltiples y presenta una conclusión sobre el contraste entre natalidad y mortalidad en Sinaloa.

Año	Natalidad(*)	Mortalidad(*)
1980	34	6.8
2007	17.6	5.2
2012	16.3	5.5
2030	12.5	7.8

(*) Por cada mil habitantes

1. Debes tener una idea clara sobre el significado de los siguientes términos estadísticos:

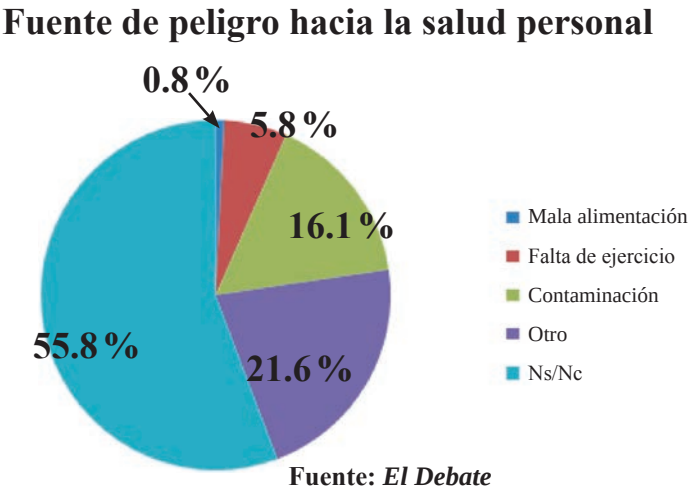
Modalidades de una variable	Frecuencia relativa
Variable cualitativa nominal	Porcentaje
Variable cualitativa ordinal	Gráfico de barras
Variable cuantitativa discreta	Gráfico circular
Variable cuantitativa continua	Moda o categoría modal
Fases de la exploración de datos	Gráfico de barras múltiples
Frecuencia absoluta.	

Agrégalas a tu diccionario. Ejemplifica.

2. De las variables siguientes, indicar cuáles son cuantitativas discretas y cuáles continuas.

a) Número de preguntas de una prueba.
b) Edad de un grupo de universitarios.
c) El peso de una persona.

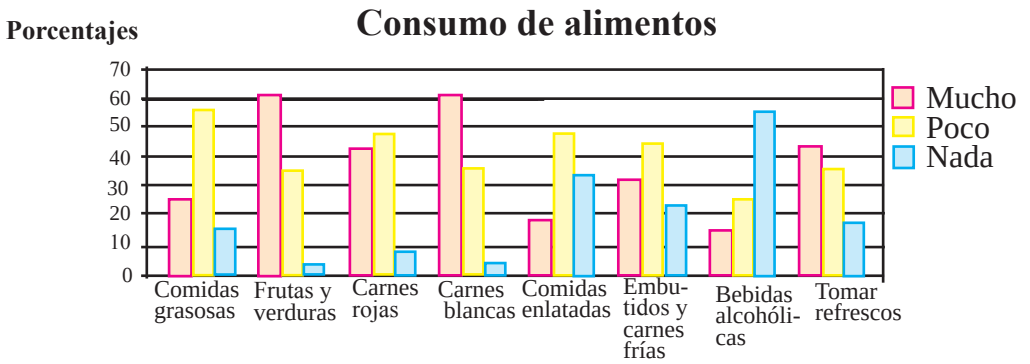
3. A partir del siguiente gráfico de sectores, construye una tabla de frecuencias absolutas y de frecuencias relativas.



Fuente: ***El Debate***

4. En Sinaloa, durante el ciclo escolar 2008-2009, de un total de 626,210 alumnos del nivel básico, 112,517 cursan preescolar, 353,690 cursan primaria y 160,003 secundaria. Representa en un gráfico circular esta información. ¿Por qué crees que hay variabilidad de un nivel a otro?

5. Analiza la información de la encuesta realizada y contesta:



Respuesta de 504 personas encuestadas el 2 y 3 de abril de 2009.

Fuente: El Debate

a. ¿Las personas encuestadas se consideran población o muestra?
b. ¿Cuál es la variable que se está estudiando, y cuáles sus modalidades?
c. ¿Por qué fue necesario utilizar un gráfico de barras múltiples?
d. Saca conclusiones de la gráfica.

6. Las siguientes tablas muestran los resultados de la prueba Enlace del 2006 al 2007 en español y matemáticas. ¿Cómo es la evolución en cada materia? ¿En qué materia salieron mejor? Argumenta.

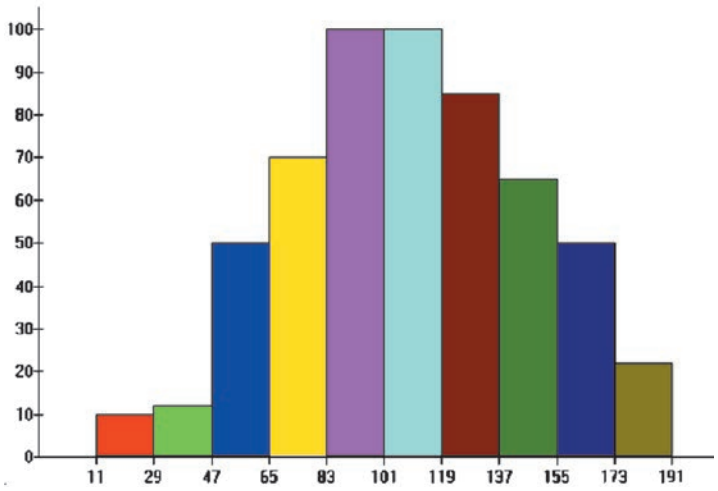
Español

Categoría	2006	2007	2008
Insuficiente y elemental	79 %	74.4%	69.1%
Bueno y excelente	21 %	25.7%	30.9%

Matemáticas

Categoría	2006	2007	2008
Insuficiente y elemental	85.2%	78.5%	79.9%
Bueno y excelente	14.8%	21.5%	26.3%

Exploración de datos cuantitativos



3

UNIDAD

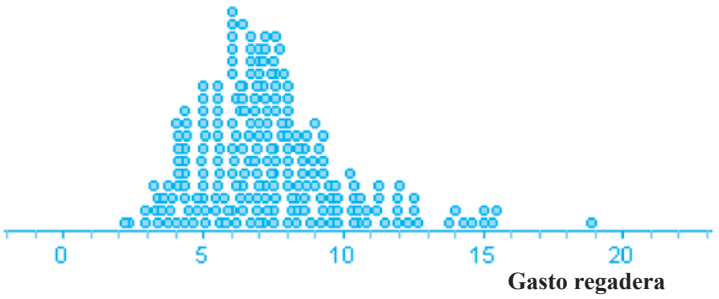
Lección 3.1 Antecedente 1 para la exploración de datos cuantitativos: concepto de Distribución

Objetivos: Aprender a construir gráficos de puntos.
Aprender a reconocer los aspectos que permiten una descripción informal de las distribuciones

Actividad 8

Qué hacer

1. Consulta las **páginas 43 a 45** y al finalizar tu estudio, haz una descripción de la distribución mostrada, la cual corresponde al gasto en litros por minutos de regaderas en algunos hogares.



Antes de estudiar la exploración de datos cuantitativos, necesitas conocer algunos conceptos que resultan clave para la comprensión de dicha exploración.

1. Distribución: forma de la distribución, agrupamientos y valores atípicos.

Para comprender estos conceptos, realiza la siguiente actividad:

Actividad 3.1 a

1. Realiza un censo en tu grupo para recolectar datos de la variable «*número de integrantes en las familias de los alumnos*.

Datos de la variable «*número de integrantes en las familias*»

2. Traza una escala horizontal que contenga los números naturales ordenados desde el valor menor de los datos hasta el mayor.

Número de integrantes

Actividad 3.1 a (Cont.)

3. Coloca un punto para cada estudiante en el número apropiado según los integrantes que sean en la familia. Para valores repetidos, apila los puntos hacia arriba unos a otros.

Número de integrantes

La representación gráfica que has construido para los integrantes en la familia, se llama **gráfico de puntos.**

4. Encierra con un círculo el punto que representa el dato que corresponde a tu familia. Compara tu dato con el de los demás.

5. Ahora, si en vez de colocar los números naturales ordenados (modalidades) en el eje horizontal, los colocas en la primer columna de una tabla, y en una segunda columna escribes la frecuencia con que aparece cada valor, entonces se forma una **distribución de frecuencias.** Forma una distribución de frecuencias de estos datos. Observa que el número de puntos en el gráfico para algún valor, corresponde a su frecuencia en la distribución.

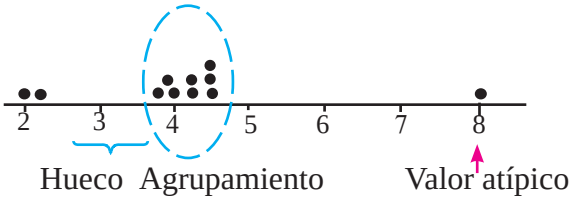
Valores distintos (modalidades) x	**Frecuencia** f

Podemos ya establecer una definición de distribución.

Distribución, es el *patrón de variabilidad* que presentan los datos de una variable. La distribución exhibe la frecuencia de cada valor de la variable.

Los gráficos de puntos, son muy útiles para mostrar la **distribución de datos.** La distribución de frecuencia también nos muestran esta distribución, pero como se verá posteriormente, su mayor utilidad está en los cálculos estadísticos.

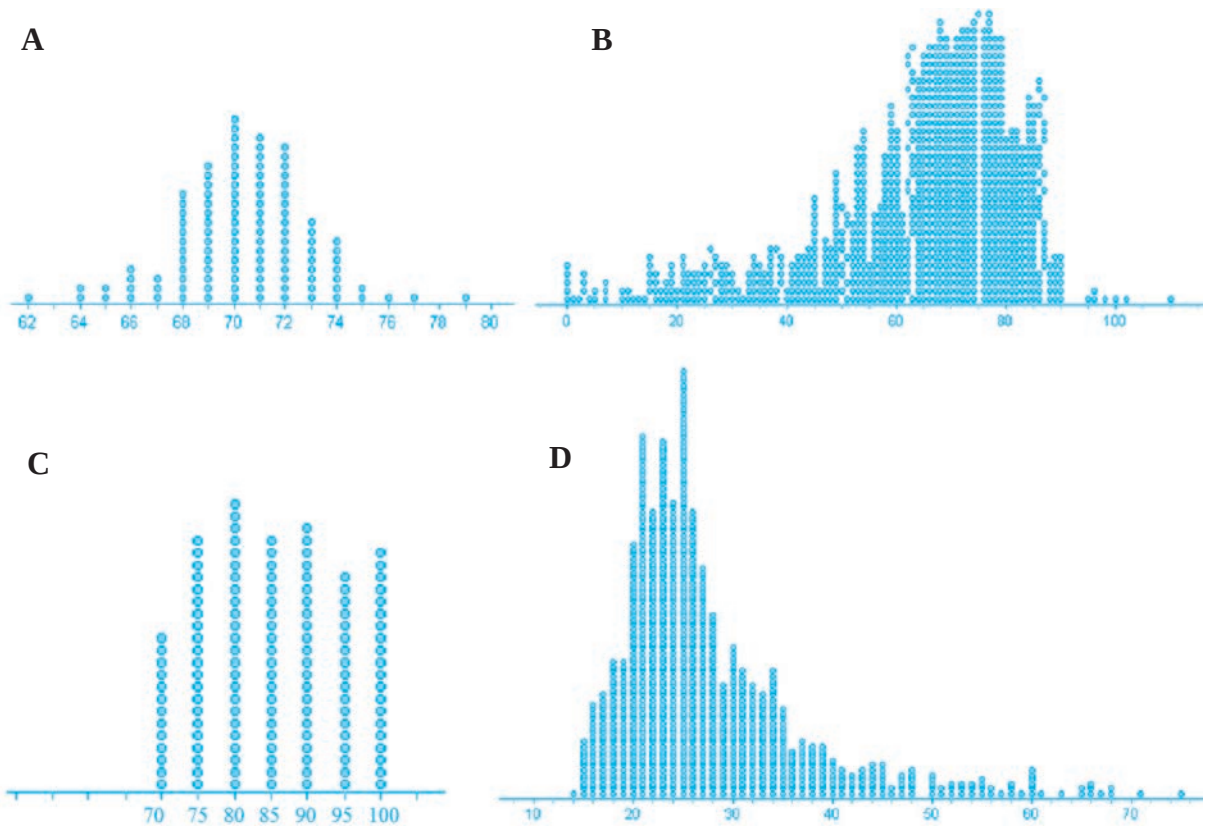
El gráfico de puntos muestra mejor que la tabla de la distribución de frecuencias, características tales como *agrupamientos, huecos y valores atípicos.*



Actividad 3.1 b

Para comprender algunos aspectos de las distribuciones, realiza la siguiente actividad:

1. A continuación se presentan gráficas de puntos de algunas variables cuyas distribuciones son bien definidas. Obsérvalas con atención y contesta lo indicado.



a) Traza sobre cada uno de los gráficos una curva suave que envuelva a los puntos.

b) Si se sabe que los gráficos de puntos corresponden a las variables, e*dad de una población de mujeres al casarse*, *edad de mortalidad de una población de hombres, calificaciones de un examen fácil* de un grupo de estudiantes, y *calificaciones de un examen difícil de un grupo de estudiantes*, ¿qué gráfico corresponde a cada una de estas variables?_____
Explica tus elecciones._____
_____

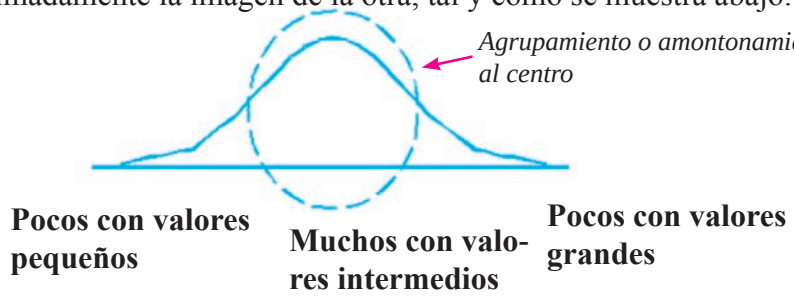
c) ¿Qué diferencias más destacables encuentras entre esas distribuciones?_____
_____
_____

d) Escribe un párrafo de al menos cuatro oraciones describiendo varias características de cada distribución. Imagina que estás intentando explicar lo que estas distribuciones muestran, a alguien que no puede ver los gráficos y no tiene ninguna idea en lo absoluto de las variables implicadas.
_____
_____
_____

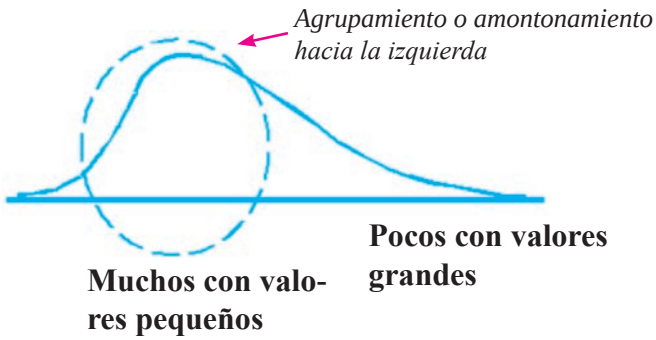
Basados en la actividad anterior, podemos establecer algunas de las características que permiten la descripción de las distribuciones:

1. **La forma** de una distribución puede revelar mucha información. A pesar de que existe una variedad ilimitada de formas, las mostradas a continuación aparecen con bastante frecuencia por lo que tienen sus propios nombres.

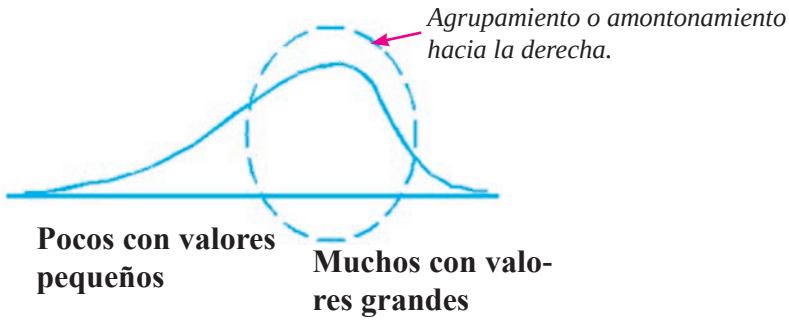
Distribución simétrica. Una distribución es simétrica si una mitad es aproximadamente la imagen de la otra, tal y como se muestra abajo.



Distribución sesgada a la derecha. Una distribución es sesgada a la derecha si presenta una cola hacia los valores grandes.

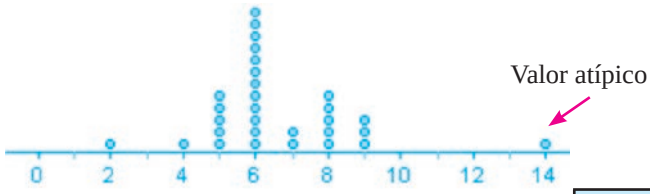


Distribución sesgada a la izquierda. Una distribución es sesgada a la izquierda si presenta una cola hacia los valores pequeños.



2. **Valores atípicos.** Son observaciones que difieren marcadamente del patrón establecido por la gran mayoría de los datos. Con frecuencia, será necesario examinar un posible error de inclusión de estos valores.

Ejemplo



Número de integrantes en las familias de una muestra de estudiantes

A partir de este gráfico, forma una tabla de distribución de frecuencias.

Valores distintos (modalidades) ***x***	**Frecuencia** ***f***

Actividad 3.1 c

a) Utilizando tus nuevos conocimientos sobre distribuciones, vuelve a describir las distribuciones de la actividad (3.1 a). Identifica forma, valores atípicos, agrupamientos.

b) Construye una distribución de frecuencias para los gráficos ***A*** y ***C*** presentados en dicha actividad.

Ejercicio 3.1

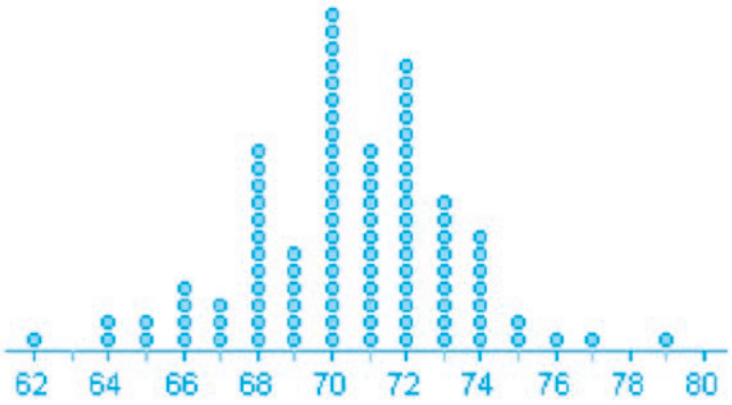
1. El siguiente conjunto de datos corresponden a la edad en que contrajeron matrimonio una muestra de 37 mujeres.

   30, 27, 56, 40, 30, 26, 31, 24, 23, 35, 29, 33, 29, 22, 33, 29, 46, 25, 34, 19, 23, 23, 44, 29, 30, 25, 23, 60, 25, 27, 37, 24, 22, 27, 31, 24, 26.

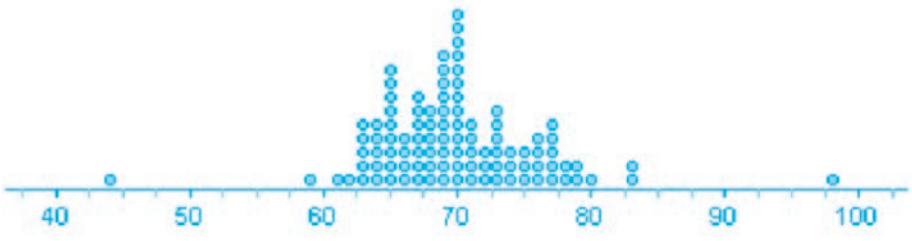
   a) Construye un gráfico de puntos
   b) Construye una tabla de distribución de frecuencias
   c) Describe la distribución

2. Los datos siguientes corresponden a la variable *número de automóviles* que poseen 25 familias. Construye un gráfico de puntos y forma una tabla de distribución de frecuencias.
   0, 1, 2, 3, 1, 0, 1, 1, 1, 4, 3, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1.

3. Los siguientes gráficos de puntos corresponden a las calificaciones obtenidas por dos grupos de alumnos. Haz una descripción completa de cada una de las distribuciones.



Calificaciones grupo A



Calificaciones grupo B

Lección 3.2 Antecedente 2 para la exploración de datos cuantitativos: Medidas de Tendencia central

Objetivos:

- Aprender a calcular la media, mediana y moda para resumir el centro de una distribución.
- Investigar y descubrir propiedades de esos resúmenes estadísticos.
- Desarrollar una conciencia de situaciones en las cuales ciertas medidas son o no, apropiadas.

Actividad 9

Qué hacer

1) Consulta las **páginas 69 a 75** y al finalizar tu estudio contesta:

a) El siguiente conjunto de datos corresponden a la edad en que contrajeron matrimonio una muestra de 37 mujeres.

30, 27, 56, 40, 30, 26, 31, 24, 23, 35, 29, 33, 29, 22, 33, 29, 46, 25, 34, 19, 23, 23, 44, 29, 30, 25, 23, 60, 25, 27, 37, 24, 22, 27, 31, 24, 26.

Calcula la **media, mediana y moda** de dichas edades.
Describe la distribución.

La recolección de datos nos proporciona una gran masa de cifras cuya interpretación generalmente no resulta fácil. Para obtener de estos datos una imagen más clara, se organizan y se presentan mediante gráficos estadísticos. Pero, por lo general, para poder obtener de tal masa algún conocimiento que valga la pena, es necesario extraer unos cuantos valores que sirvan de indicadores y que los resuman satisfactoriamente. Un valor típico descriptivo como ese, se llama **promedio**. Aparejado con estos valores de resumen, deben calcularse otros más, llamados medidas de dispersión que indican cuánto se separan los datos respecto del promedio.

Cuando intentamos resumir numéricamente los datos, la característica básica que focalizamos es el «centro» de la distribución de esos datos. Es por esta razón que los promedios son también llamados medidas de tendencia central.

Desde nuestra enseñanza primaria al término de cada ciclo escolar, hemos calculado promedios. Para obtener este promedio, sumamos las calificaciones alcanzadas en las diferentes asignaturas y dividimos el resultado entre el número de ellas. Sin embargo, este es sólo un tipo de promedio; existen otros más. Los tres promedios más usados son: **media aritmética**, **mediana** y **moda**.

Los promedios o medidas de tendencia central, son valores medios en torno al cual parecen agruparse los datos.

Media aritmética.

La media aritmética o simplemente media, es el promedio o medida de tendencia central más usado, a tal grado, que se utiliza como sinónimo de promedio.

Se calcula sumando todas las observaciones obtenidas, dividiendo después esa suma entre el número total de elementos involucrados.

Ejemplo

Paula obtuvo las siguientes calificaciones bimestrales en 11 materias:

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

¿Cuál es su calificación promedio (o media)?

$$\textit{Calificaciones promedio (media)} = \frac{\textit{suma de calificaciones}}{\textit{número de calificaciones}}$$

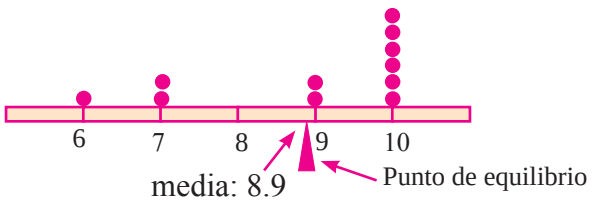
$$= \frac{6+10+10+10+9+7+10+9+10+7+10}{5} = 8.9$$

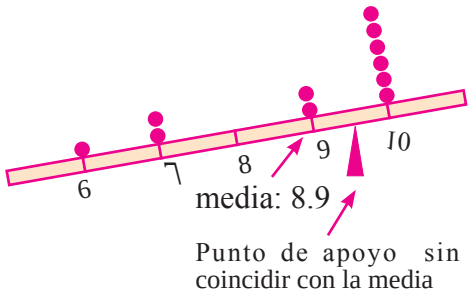
La calificación de Paula que normalmente se llama promedio, es 8.9. Este promedio realmente se llama media.

La media representada por $\bar{X}$ (que se lee como «x barra») de un conjunto de datos es la suma de los datos dividida entre el número de datos:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

No es fácil visualizar la media. Una manera útil para lograr ésto, es imaginar que el gráfico de puntos forma un sistema de «sube y baja» constituido por un tablón y un punto de apoyo. Entonces la media es simplemente el centro de gravedad del sistema, esto es, el punto donde el gráfico de puntos está balanceado.





Mediana

La mediana (denotada ***Med***) es simplemente el valor en la posición central cuando todas las observaciones están escritas de manera ordenada en forma ascendente (de menor a mayor) o descendente (de mayor a menor).

Por ejemplo, consideremos el siguiente conjunto de datos:

6 6 7 8 10 10 10

Med

Observamos que la mediana de estos datos es 8. También puede observarse que la mediana es la 4ª observación medida a partir de cualesquiera de los extremos. En general se cumple que:

> Si ***n*** es el número de datos (tamaño de la muestra), entonces, la mediana es la observación ubicada en la posición:
>
> $$\frac{n+1}{2}$$

Para un número de datos impar, la mediana es el valor de enmedio, como en el ejemplo anterior:

6 6 7 8 10 10 10

Med

Para un número par, la mediana es la semisuma de los dos elementos centrales.

Por ejemplo, consideremos los ocho datos: 10, 10, 12, 14, 15, 16, 20, 27

$n = 8$, $\frac{n+1}{2} = \frac{8+1}{2} = 4.5$ Significa que la mediana está entre el 4º y 5º dato

10 10 12 14 15 16 20 27

4º 5º

$$Mediana = \frac{14+15}{2} = 14.5$$

Procedimiento para calcular la mediana:

1º Ordene todos los datos de menor a mayor o viceversa.

2º Calcule la posición de la mediana, mediante la siguiente expresión:

$$Posición\ de\ la\ mediana = \frac{Número\ de\ elementos + 1}{2} = \frac{n+1}{2}$$

3º Identifique el elemento de la posición determinada en el 2º paso. (Si el número de elementos es par, deberá obtenerse la semisuma de los dos elementos centrales). En este caso la mediana no es un elemento del conjunto de datos.

Ejemplos a) Determinar la mediana de las calificaciones,

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

Solución

1º Ordenar los datos: 6, 7, 7, 9, 9, 10, 10, 10, 10, 10, 10.

2º Posición de la mediana:

$n = 11$

$$\textit{Posición de la mediana} = \frac{n+1}{2} = \frac{11+1}{2} = 6$$

3º La mediana es el 6º valor de la serie ordenada:

6, 7, 7, 9, 9, **10**, 10, 10, 10, 10, 10

5 valores — Mediana — 5 valores

b) Determinar la mediana de los siguientes datos: 6, 7, 5, 8, 10, 9.

Solución

1º Ordenar los datos: 5, 6, 7, 8, 9, 10.

2º Posición de la mediana:

$n = 6$

$$\textit{Posición de la mediana} = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

Significa que la mediana está entre el 3° y 4° dato.

3º La mediana es la semisuma del 3º y 4º valor:

5, 6, 7, 8, 9, 10

3° 4°

$$\textit{Mediana} = \frac{7+8}{2} = 7.5$$

La moda

La **moda** es el valor más frecuente (el que más se repite) de un conjunto de datos.

En ocasiones se presentarán dos o más valores que se repiten con mayor frecuencia. En este caso, a los datos se les conoce como **bimodales** o **multimodales** respectivamente.

Ejemplos

a) Determina la moda para las calificaciones:
6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

La moda es 10 pues es el dato que tiene mayor frecuencia: **aparece 6 veces.**

b) Determina la moda de: 6, 7, 5, 8, 10, 9.

En este conjunto de datos, no hay moda, porque ningún dato aparece más que otro.

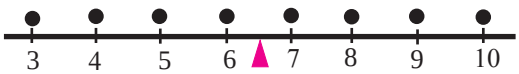
c) Determina la moda de: 5, 5, 4, 4, 6, 6, 6, 3, 3, 3

En este conjunto de datos, hay dos modas: 6 y 3, pues ambos se repiten más veces.

Realiza la siguiente actividad que te permitirá comprender algunas propiedades de la media y mediana.

Actividad 3.2a

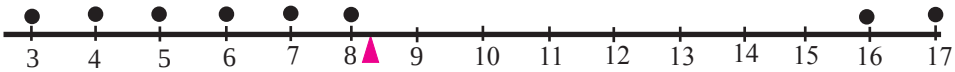
1. El siguiente gráfico de puntos, es completamente simétrico. Calcula la media y la mediana.



Debes comprobar que *mediana* = *media* = 6.5

Completa: Si el gráfico de puntos es perfectamente **simétrico,** entonces la media y la mediana son_____

2. Ahora, el siguiente gráfico es construido a partir del anterior. La diferencia está en que los dos datos del extremo derecho, han aumentado de valor. Vuelve a calcular la media y mediana.



3. Debes comprobar que la mediana sigue siendo 6.5, pero la media es afectada por los valores extremos que han aumentado de valor. La media al igual que los dos valores extremos, también se mueve hacia la derecha.

4. Para reforzar esta propiedad de la media, realiza lo siguiente:

a) Calcula la media de los datos: **12.4, 13.5, 13.6, 11.2, 15.1, 10.6, 12.4, 14.3, 113.5**

b) Con excepción del último valor (113.5) ¿en qué intervalo están estos datos?

c) Con relación a este intervalo, ¿qué puede decirse de la media?

Una vez más haz verificado que la media es muy sensible a valores muy grandes o muy pequeños. En los últimos datos, el valor 113.5, influyó en la media (la arrastró hacia un valor más grande). En este caso, tenemos razones para pensar que pudo haber un error al digitar el valor 113.5, digitando un 1 de más. Si en vez de 113.5, el valor correcto fuera 13.5, ¿cuál será el valor de la media?

Una medida cuyo valor es relativamente indiferente ante la presencia de valores atípicos de una distribución, se dice que es **resistente**.

¿Qué puedes decir acerca de la **resistencia** de la media y de la mediana? Explica por qué esto tiene sentido, basando tu argumento en la definición de cada una de ellas._____

_____

De la actividad anterior, concluimos las siguientes propiedades de la media y mediana:

- En distribuciones que tienden a ser **simétricas**, la media es muy cercana a la mediana.
- En distribuciones **sesgadas a la derecha**, la media es más grande que la mediana.
- En distribuciones **sesgadas a la izquierda**, la media es menor que la mediana.
- Mientras que la mediana es completamente insensible a valores atípicos, la media puede resultar muy afectada por estos valores.

¿Qué promedio usar: media o mediana?

La propiedad de no resistencia de la media ante valores atípicos, sugiere que en presencia de aquellos, la media no es un buen representante de los datos en cuestión. En este caso puede ser preferible usar la mediana. En otras palabras, para conjuntos de datos muy sesgados, la mediana puede ser una mejor medida de resumen. Sin embargo, siempre que debamos decidir, debe prevalecer el sentido común.

Un ejemplo clasico sobre la necesidad de tomar una decisión sobre cuál medida usar, lo constituye el ingreso salarial de algún grupo de personas. Por lo general habrá una o dos personas que ganan mucho más que el grueso del grupo. En este caso, para citar el sueldo promedio, se usa la mediana ya que al estar situada justo en la mitad de los datos, suele ser mejor representante que la media. Realiza la siguiente actividad para verificar este comentario.

Actividad 3.2b

Los datos siguientes, muestra un ordenamiento de los salarios mensuales pagados por una hipotética pequeña empresa que da empleo a 15 personas, presidente y vicepresidente incluidos.

30000, 24000, 14000, 10000, 9000, 9000, 9000, 7000, 6000, 6000, 6000, 5000, 5000, 5000, 5000

a) Construye el gráfico de puntos.

b) Calcula la media salarial.

Actividad 3.1 b (Cont.)

c) Calcula la mediana salarial.

d) Si tuvieras que reportar un salario promedio para este grupo, ¿cuál reportarías? Argumenta tu decisión._____
_____
_____

Ejercicio 3.2

1. ¿Qué lugar ocupa la media en un gráfico de puntos?

2. Si al realizar el cálculo de $\frac{n+1}{2}$ resulta un número no entero, ¿cómo se calcula la mediana?

3. ¿Por qué podría ser recomendable el uso de la mediana en vez de la media en datos sesgados?

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

   (a) 3, 2, 8, 4, 5, 7, 11, 6, 10.
   (b) 4, 3, 5, 4, 3, 8, 10, 7.
   (c) 4, 9, 8, 4, 8, 7, 6, 6.
   (d) 9, 17, 6, 5, 7, 7, 11, 8, 10, 11, 14, 15.

5. De 10 puntos posibles, un grupo de 20 estudiantes obtuvieron las siguientes calificaciones:

   0, 0, 1, 2, 4, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10.

   Obten la media, la mediana y la moda. ¿Cuál de estas tres medidas es la menos representativa del conjunto de calificaciones?

6. ¿Cuánto dura una canción? Proporciona una respuesta_____. Ahora, analiza las duraciones en minutos y segundos de 20 canciones mostradas a continuación y vuelve a contestar la pregunta.

   3:43, 5:16, 3:24, 3:29, 3:46, 4:53, 3:24, 2:38, 2:35, 3:03, 4:56, 3:27, 4:07, 3:56, 3:07, 3:48, 3:28, 3:49, 3:35, 3:32.

Lección 3.3 Antecedente 3 para la exploración de datos cuantitativos: medidas de posición

Objetivo: ♦ Aprender a calcular los cinco-números de resumen denominados de posición.

Actividad 10

Qué hacer

1) Consulta las **páginas 76 a 82** y al finalizar tu estudio contesta:

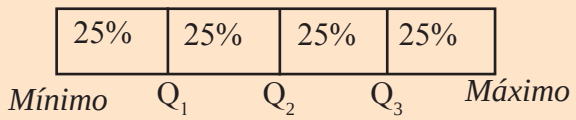
Vuelve a considerarar el conjunto de datos correspondientes a la edad en que contrajeron matrimonio un grupo de 37 mujeres.

30, 27, 56, 40, 30, 26, 31, 24, 23, 35, 29, 33, 29, 22, 33, 29, 46, 25, 34, 19, 23, 23, 44, 29, 30, 25, 23, 60, 25, 27, 37, 24, 22, 27, 31, 24, 26.

a) Calcula los ***cinco-números de resumen*** de dichas edades.
b) Dibuja el gráfico de caja.
c) Describe el conjunto de datos.

La utilización de un único número para resumir un conjunto de datos, muy raras veces es suficiente. Por tal razón, es necesario determinar otros indicadores que nos orienten en donde están los datos. La mediana junto con otras cuatro medidas, constituyen los llamados ***cinco-números de resumen.*** Estos cinco-números son: *el mínimo, el primer cuartil, la mediana (o segundo cuartil), el tercer cuartil y el máximo.* Estas medidas se llaman de posición puesto que se utilizan para describir la posición que un dato específico posee en relación con el resto de los datos cuando están ordenados de menor a mayor.

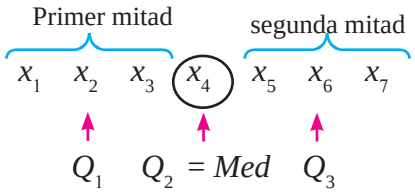
Cuartiles: Son los valores de la variable que dividen los datos ordenados en cuartos.



El primer cuartil (denotado Q_1), es un número tal que a lo sumo 25% de los datos son menores que Q_1 y a los sumo 75% son mayores. El segundo cuartil es la mediana. El tercer cuartil (denotado Q_3), es un número tal que a lo sumo 75% de los datos son menores que Q_3 y a lo sumo 25% son mayores.

La determinación de los cuartiles la basaremos en las siguienes propiedades:

- La mediana divide al conjunto de datos ordenados en dos mitades. El primer cuartil divide la *primer mitad* de los datos ordenados en dos, y el tercer cuartil divide la *segunda mitad* de los datos en dos.



En otras palabras, el primer cuartil es la mediana de las observaciones que están antes de la posición de la mediana de todos los datos, y el tercer cuartil es la mediana de las observaciones que están después de la mediana.

Ejemplos a) Determinar los cuartiles de los datos:

5, 9, 8, 10, 6, 8, 10, 8, 9

Solución

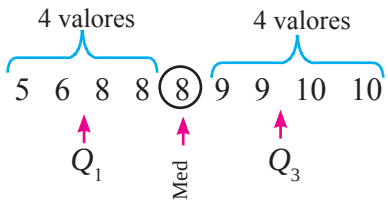
1º Ordenar los datos: 5, 6, 8, 8, 8, 9, 9, 10, 10

2º Posición de la mediana:

$n = 9$

$$\textit{Posición de la mediana} = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

3º La mediana es la 5ª observación de la serie ordenada:



Mediana = 8

4º Ignorando la posición (y el valor de la mediana), el primer cuartil es la mediana de los números 5, 6, 8 y 8. Entonces,

$$Q_1 = \frac{6+8}{2} = 7$$

De manera similar,

$$Q_3 = \frac{9+10}{2} = 9.5$$

Entonces, $Q_1 = 7$

$Q_2 = Med = 8$

$Q_3 = 9.5$

Ejemplos b) Determinar los cuartiles de los datos:

6, 8, 9, 7, 8, 10

Solución

1º Ordenar los datos: 6, 7, 8, 8, 9, 10

2º Posición de la mediana:

$n = 6$

$$\textit{Posición de la mediana} = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

3º La mediana está en la posición 4.5:

6 7 8 8 9 10

Med

Entonces

$$Med = \frac{8+8}{2} = 8$$

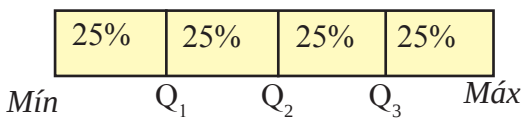
4º El primer cuartil es la mediana de los números 6, 7 y 8, y el tercer cuartil es la mediana de 8, 9 y 10. Entonces, $Q_1 = 7$ y $Q_3 = 9$.

6 7 8 8 9 10

Q_1 Med Q_3

Los *cinco-números* de resumen

El valor mínimo y el valor máximo de los datos, junto con la mediana, el primer y tercer cuartil constituyen los cinco-números de resumen.

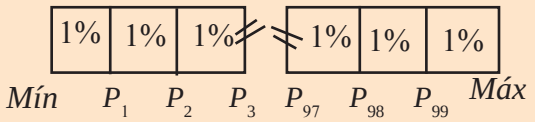


Aquí *Min* y *Max* son las observaciones más pequeña y más grande respectivamente.

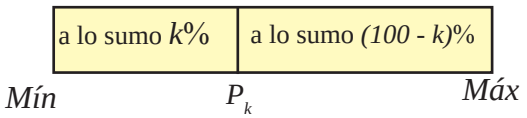
Percentiles

Para obtener los cuartiles, dividimos al conjunto de datos ordenados en cuatro partes aproximadamente iguales (es aproximado porque algunos cuartiles caen en las fronteras). Ahora, si tenemos un conjunto muy grande de datos, también lo podemos dividir en 100 partes aproximadamente iguales, obteniendo de esta manera ***99 percentiles.***

Percentiles, son los valores de la variable que dividen un conjunto de datos ordenados en *100* partes aproximadamente iguales. Cada conjunto de datos tiene *99 percentiles.*



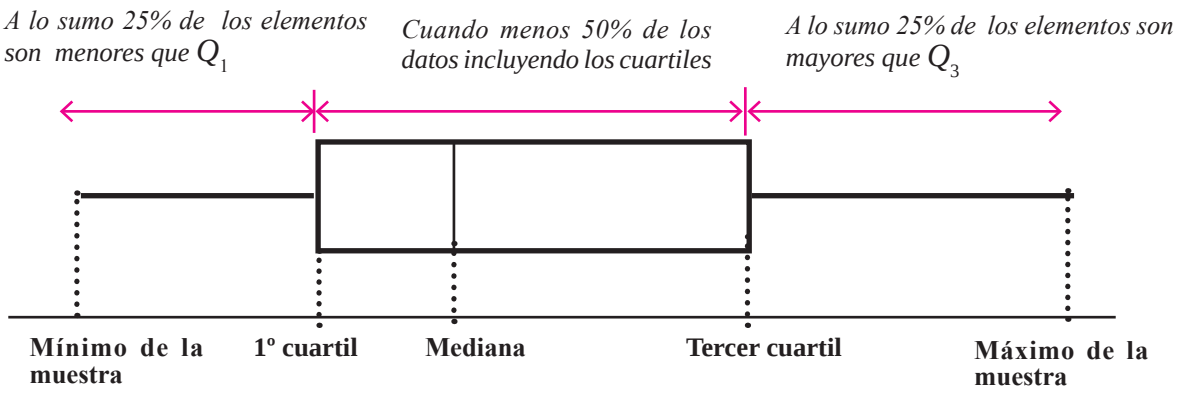
El *k-ésimo percentil,* P_k, es un valor tal que a lo sumo *k*% de los datos son menores que P_k y a lo sumo $(100-k)\%$ de los datos son mayores.



Atendiendo las definiciones de los cuartiles y de los percentiles, observamos que: el *primer cuartil* es el *25avo percentil*, puesto que a lo sumo el 25% de las observaciones caen abajo de Q_1. Similarmente, la *mediana* es el *50avo percentil,* y el *tercer cuartil* es el *75avo percentil.*

Gráfico de caja

Es un tipo de gráfico en el que se remarcan los *cinco-números* de resumen de los conjuntos de datos. El conjunto de valores de los datos comprendidos entre el primer y tercer cuartil, Q_1 y Q_3 se representan por un rectángulo (caja) con la mediana indicada por un segmento. Utiliza las definiciones de Q_1 y Q_3 para comprender la validez de los porcentajes indicados en el gráfico.



Ejemplo Para ilustrar la construcción de un diagrama de caja, utilizaremos los siguientes datos que corresponden a la variable *temperaturas mínimas diarias* registradas en la ciudad de Culiacán durante el mes de febrero de 2008.

10, 10.5, 13.5, 14, 11.5, 9, 9, 11, 11, 11.5, 14, 14, 11.5, 11.5, 11, 12, 10.5, 10.5, 11, 12.5, 12, 12, 11.5, 13.5, 13, 14, 14, 14.5, 15.5

Procedimiento

1. Determinar los cinco-números de resumen. Para ello, recuerda que debemos ordenar los datos.

9,
9,
10,
10.5,
10.5,
10.5,
11, ← Q_1
11,
11,
11,
11.5
11.5,
11.5,
11.5,
11.5, ← ***Med***
12,
12,
12,
12.5,
13,
13.5,
13.5, ← Q_3
14,
14,
14,
14,
14,
14.5,
15.5

$$\textit{Posición de la mediana} = \frac{n+1}{2} = \frac{29+1}{2} = 15$$

La mediana es el dato 15º

Med = 11.5

A partir de los datos ordenados obtenemos que:

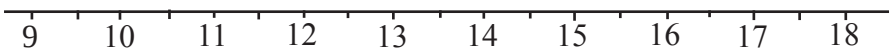
$$Q_1 = \frac{11+11}{2} = 11$$

$$Q_3 = \frac{13.5+14}{2} = 13.75$$

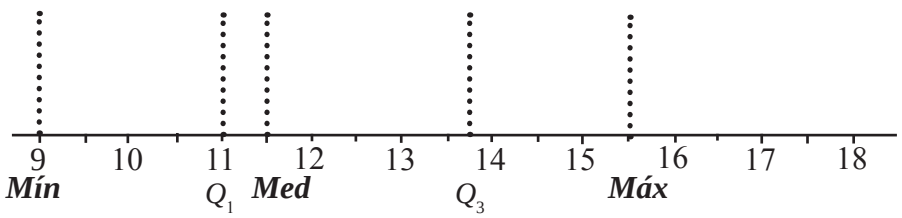
Por lo tanto, los *cinco-números* de resumen son:

Mín = 9
Primer cuartil = 11
Med = segundo cuartil = 11.5
Tercer cuartil = 13.75
Máx = 15.5

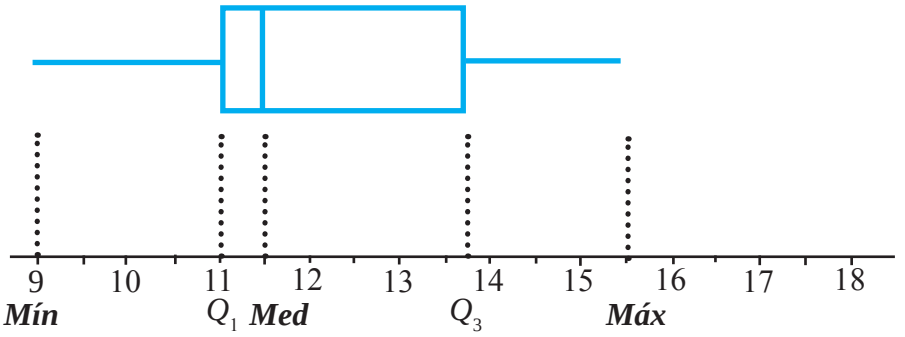
2. Localizar en un eje horizontal todos los posibles valores que puede tomar la variable (modalidades)



3. Trazar perpendiculares al eje anterior en los valores correspondientes a: valor mínimo, primer cuartil, segundo cuartil, tercer cuartil y valor máximo.

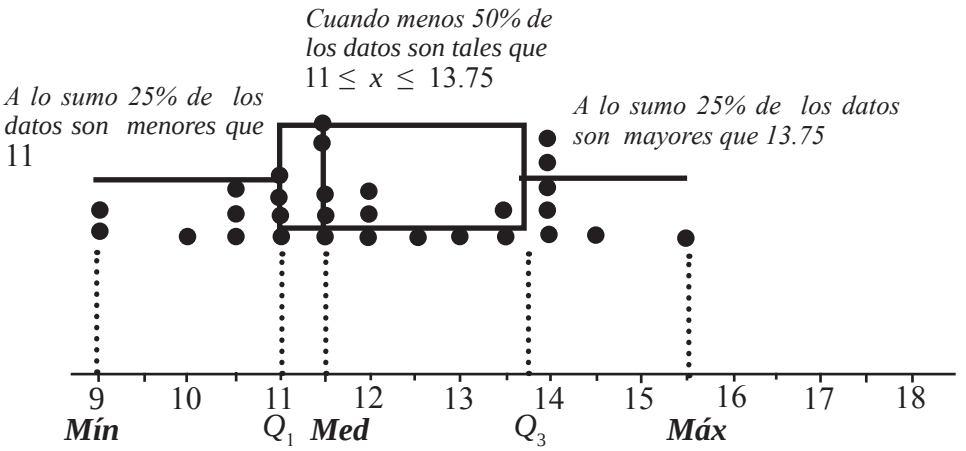


4. El conjunto de valores de la muestra comprendidos entre el primer y tercer cuartil, Q_1 y Q_3 se representa por un *rectángulo (caja)* con la mediana indicada por un segmento.



Si decimos que una cantidad a lo sumo es 25, significa que puede valer 25 o menos. Entonces, en la primer región del gráfico de caja, pueden estar 25%, 24%, 23%, 22%.... de los datos.

Para poder utilizar el gráfico de caja, es necesario recordar las definiciones de cada una de las medidas que involucra. A continuación sobreponemos al gráfico de caja, un gráfico de puntos, así como las condiciones que deben cumplir los datos para ubicarse en cada una de las regiones.



Si decimos que una cantidad cuando menos es 50, significa que puede valer 50 o más. Entonces, en la segunda región, pueden estar 50%, 51%, 52%, 53%.....de los datos.

Verifiquemos que se cumplen estas condiciones el caso que nos ocupa:

-En la primer región, hay 6 datos, los cuales son menores que el primer cuartil. 6 con respecto al total es: $\left(\frac{6}{29}\right)\times 100\% = 20.7\%$.

-En la segunda región (que incluye sus orillas), hay 16 datos, los cuales cumplen con $11 \leq x \leq 13.75$. 16 con respecto al total es: $\left(\frac{16}{29}\right)\times 100\% = 55.2\%$.

-En la tercer región, hay 7 datos, los cuales son mayores que el tercer cuartil. 7 con respecto al total es: $\left(\frac{7}{29}\right)\times 100\% = 24.1\%$.

De esta manera, quedan repartidos todos los datos en las regiones del gráfico de caja.

Actividad 3.3a

Los datos siguientes corresponden a la variable *temperaturas mínimas diarias* registradas en la ciudad de Culiacán durante el mes de febrero de 2009.

11, 11, 13, 13.5, 14, 14, 14, 13.5, 12.5, 12, 10, 9.5, 11, 12.5, 12.5, 13.5, 12.5, 12.5, 13, 12.5, 13.5, 14.5, 17, 17, 16.5, 16.5, 17, 16.5.

a) Calcula los cinco-números de resumen.
b) Construye el gráfico de caja.
c) Verifica los porcentajes de datos que deben ubicarse en cada una de las regiones del gráfico.

Ejercicio 3.3

1. ¿Cuáles son los cinco-números de resumen?
2. ¿Qué entiendes por cuartiles?
3. ¿Qué entiendes por percentiles?
4. ¿Qué semejanzas encuentras entre los cuartiles y los percentiles?
5. Hace algunos años, el presidente de la República en turno, declaró que los nuevos impuestos al salario sólo afectaban al último ***decil*** de la población. ¿Qué crees signifique esto?
6. Calcula los *cinco-números* de resumen y dibuja el gráfico de caja de los conjuntos de datos siguientes:
   a) 3, 2, 8, 4, 5, 7, 11, 6, 10.
   b) 4, 3, 5, 4, 3, 8, 10, 7.
   c) 4, 9, 8, 4, 8, 7, 6, 6.
   d) 9, 17, 6, 5, 7, 7, 11, 8, 10, 11, 14, 15.
7. Los datos siguientes, muestran los puntos totales ganados por los 18 equipos de futbol de primera división en el torneo de clausura 2009.

   36, 36, 28, 26, 26, 25, 23, 23, 22, 22, 21, 21, 21, 17, 17, 14, 14, 13.

   Calcula los *cinco - números* de resumen y el gráfico de caja. Verifica los porcentajes de datos que deben caer en cada una de las regiones del gráfico de caja.
8. De 10 puntos posibles, un grupo de 20 estudiantes obtuvieron las siguientes calificaciones:
   0, 0, 1, 2, 4, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10.
9. ¿Cuántas páginas tiene un libro típico editado por la DGEP para ayudarte en tu aprendizaje?
   Escribe aquí tu respuesta_____

   Ahora, analiza los siguientes datos que corresponden al número de páginas de todos los libros del plan 2006 editados haste el momento.
   Calcula los *cinco - números* de resumen y el gráfico de caja. Verifica los porcentajes de datos que deben caer en cada una de las regiones del gráfico de caja.

   187, 181, 140, 112, 205, 223, 271, 169, 67, 208, 229, 160, 255, 224, 156, 244, 224, 93, 299, 208, 181, 127, 182, 247, 206, 171, 88, 289, 323, 192, 211, 135, 262, 192, 223, 111, 256, 304, 188, 247, 144,

   Puedes utilizar gráficos, promedios y medidas de posición para argumentar tu respuesta.

Lección 3.4 Antecedente 4 para la exploración de datos cuantitativos: medidas de dispersión

Objetivo: Aprender a calcular las medidas que resumen la variabilidad de una distribución de datos: rango, rango intercuartílico, varianza y desviación estándar.

Actividad 11

Qué hacer

1) Consulta las **páginas 83 a 93** y al finalizar tu estudio contesta:

a) Vuelve a considerar el conjunto de datos correspondientes a la edad en que contrajeron matrimonio un grupo de 37 mujeres.

30, 27, 56, 40, 30, 26, 31, 24, 23, 35, 29, 33, 29, 22, 33, 29, 46, 25, 34, 19, 23, 23, 44, 29, 30, 25, 23, 60, 25, 27, 37, 24, 22, 27, 31, 24, 26.

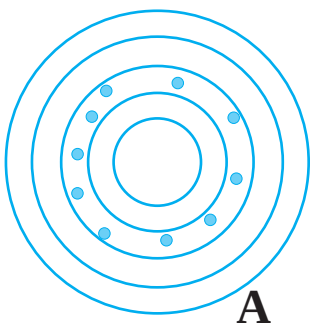
Calcula el **rango, rango intercuartílico, varianza y desviación estándar.**
Vuelve a describir la distribucion.

Las medidas del centro y de posición son indispensables para describir una distribución. Sin embargo, no son suficientes, por lo que junto con estos valores de resumen, deben calcularse otros más, llamados medidas de dispersión o de variabilidad que indican cuánto se separan los datos respecto del promedio. Para comprender esta afirmación, realiza la siguiente actividad.

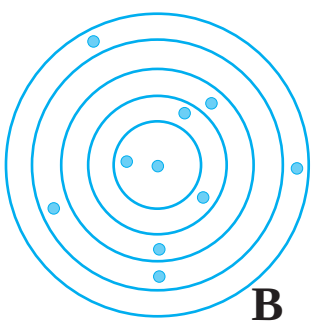
Actividad 3.4 a

Estudia la situación plateada y contesta lo indicado.

La siguiente prueba de destreza en tiro, es aplicada a dos competidores, A y B:



A



B

Valores

Círculo central	=	10
1ª. Franja circular	=	9
2ª. Franja circular	=	8
3ª. Franja circular	=	7
4ª. Franja circular	=	6

a) ¿Qué competidor tiene la mejor media de puntuación? Calcula las medias.

¿Quién es el mejor tirador? Argumenta_____

_____

Tus respuestas a las cuestiones anteriores deben ser parecidas a las siguientes:

Las puntuaciones de *A* fueron: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8.

$\Sigma x = 80$ $\qquad \bar{X} = \frac{\Sigma x}{n} = \frac{80}{10} = 8$

Las puntuaciones de *B* fueron: 10, 10, 9, 9, 8, 8, 7, 7, 6, 6.

$\Sigma x = 80$ $\qquad \bar{X} = \frac{\Sigma x}{n} = \frac{80}{10} = 8$

¿Qué blanco tiene la mejor media? *Ninguno: las medias son idénticas*.

Si solo conociéramos la puntuación media, podríamos decir que el competidor *A* es tan bueno como el competidor *B* (ambos competidores tienen la misma puntuación media).

Sin embargo, observando la figura, el blanco *A* muestra *menor variación* que el blanco *B*.

Si juzgamos a partir de la diseminación de los "impactos" en los blancos de la figura, ¿cuál competidor es el tirador más consistente? _____

Por lo tanto, *conocer el promedio no es suficiente para describir a los datos.*

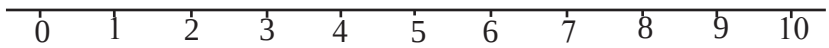
Como un segundo ejemplo, realiza la siguiente actividad que consiste en analizar las calificaciones obtenidas por cuatro grupos en un examen de geometría. Tu análisis deberá consistir en:

- Construir un gráfico de puntos.
- Calcular la media.
- Describir la distribución.

Actividad 3.4 b

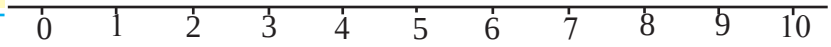
Grupo A

4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 8,



Grupo B

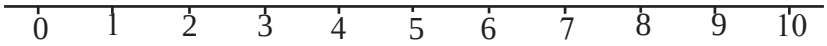
1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 9, 9, 10



Actividad 3.4 b (Cont.)

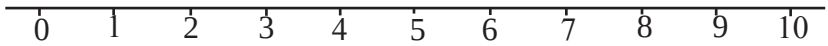
Grupo C

0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 10, 10



Grupo D

0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 4, 5, 5, 6, 6, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10,



Al calcular la calificación media en cada grupo, obtenemos unos valores muy parecidos, próximos a 5. Sin embargo, observando las gráficas, **las distribuciones de las calificaciones son muy distintas**.

¿Qué significa esto? Que conociendo sólo la calificación media de un grupo, no podemos hacernos una idea de cómo se distribuyen las calificaciones, es decir, no sabemos lo dispersas que están las notas con relación a la media.

Así pues, una medida de tendencia central por sí sola, no describe ni resume adecuadamente una distribución de datos; es necesario acompañarla de un indicador que dé cuenta del grado de heterogeneidad o dispersión con que se distribuyen los datos de la variable. Una medida de dispersión dice cuánto se desvían los datos respecto a las tendencias centrales.

Demostrada la insuficiencia de una medida de tendencia central para describir adecuadamente, por sí sola, un conjunto de datos, procederemos al estudio de algunos indicadores de dispersión.

El rango

El rango es la medida más simple para indicar qué tan dispersos están los datos.

El rango denotado por *R*, es la diferencia entre el valor mayor y el valor menor de los datos.

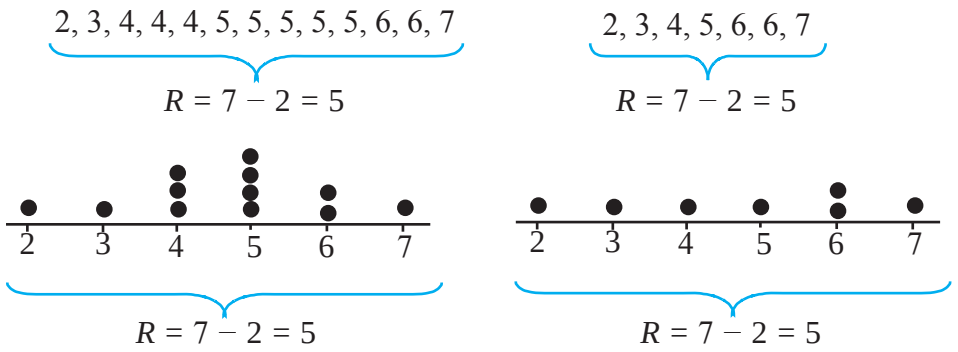
Rango = Dato mayor - dato menor

$$R = x_{máx} - x_{mín}$$

El rango, si bien brinda una primera idea de la dispersión de un conjunto de datos, tiene el inconveniente de que **sólo toma en cuenta los dos valores extremos** y descuida los intermedios; es decir, no dice cuánto se desvía un dato intermedio de la tendencia central. Por esta razón no sirve, por sí solo, para dar cuenta objetivamente de la desviación en su conjunto; más que nada, se le debe usar como complemento de otras medidas de dispersión.

Ejemplo

Para ilustrar la insuficiencia de este indicador, consideremos los siguientes conjuntos de datos:



En ambos conjuntos, el rango es el mismo, pero observando la distribución de los datos en cada conjunto, descubrimos que la primera muestra una mayor concentración que la segunda en torno a sus medidas de tendencia central.

El rango intercuartílico

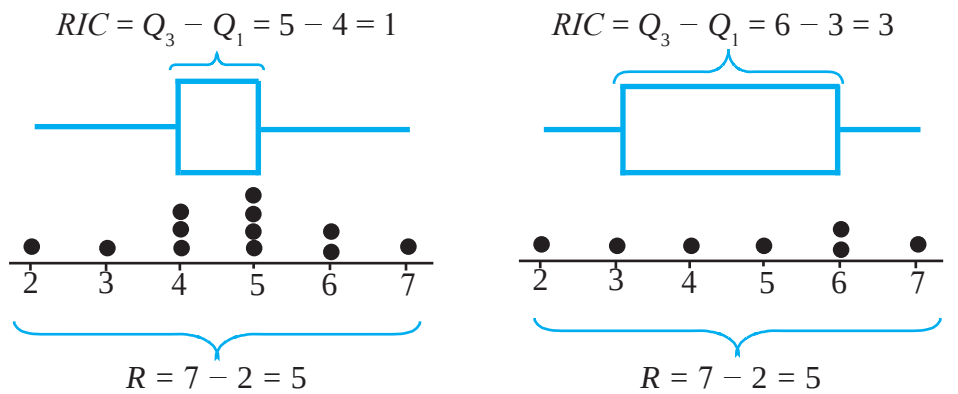
A diferencia del rango, el **rango intercuartílico** es mucho menos afectado por la presencia de valores extremos tales como los valores atípicos.

El rango intercuartílico denotado por *RIC*, es la diferencia entre los cuartiles primero y tercero. Es el rango que tienen el 50% de los datos centrales.

Rango intercuartílico = Tercer cuartil - primer cuartil

$$RIC = Q_3 - Q_1$$

Utilizaremos los datos del último ejemplo para ilustrar este concepto. El *RIC* nos da una idea sobre el largo de la caja. En el conjunto de la izquierda, hay un agrupamiento que se refleja con una caja reducida; en cambio en el de la derecha, los datos están más dispersos sin ningún agrupamiento, lo cual se refleja con una caja amplia.



Desvío o desviación

El desvío es el concepto fundamental que nos permitirá comprender posteriormente otras medidas de dispersión.

El desvío de cada observación es **la diferencia entre la observación y la media**, se denota por ***d***:

$$d = x_1 - X$$

Propiedad que relaciona la media y sus desvíos

La media tiene una propiedad interesante que consiste en los siguiente:

Si calculamos los desvíos de todas las observaciones con respecto a la media y sumamos esos desvíos el resultado es igual a cero.

$$(x_1 - \overline{X})\ (x_2 - \overline{X}) + \ldots + (x_n - \overline{X}) = 0$$

Ejemplo

Supongamos que en una fiesta se repartieron chocolates en barra a los presentes. Los chocolates se lanzaban al aire y quien más corría los atrapaba. La siguiente distribución contiene el número de chocolates que cada invitado consiguió atrapar:

Carlos	8
Ariana	7
Ana	3
María	5
Teresa	4
José	6

Calcula la media de chocolates repartidos_____

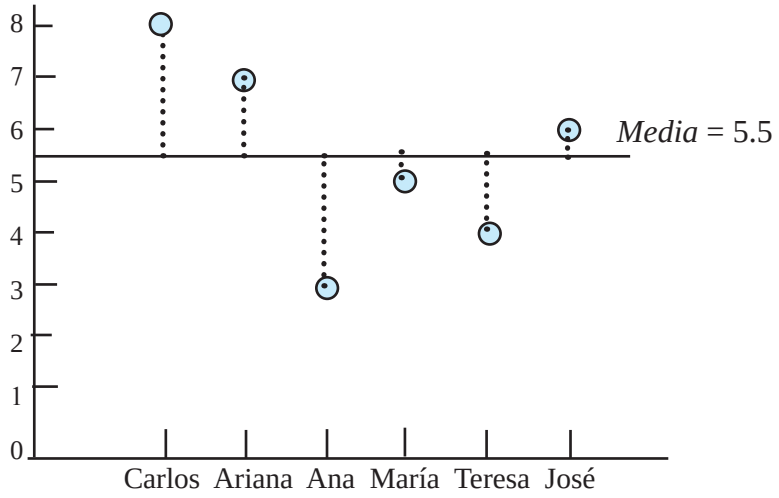
El acuerdo fue que a cada invitado le tocara una cantidad igual a la media de chocolates. Sin embargo, algunos invitados atraparon menos chocolates de lo que indica la media, mientras que otros atraparon más. Los invitados que atraparon más chocolates resolvieron repartirlos entre los que atraparon menos de forma que todos recibieran la misma cantidad. ¿Es posible ésto?

Calculemos los desvíos con respecto a la media:

		Desvío
Carlos	8	8 - 5.5 = 2.5
Ariana	7	7 - 5.5 = 1.5
Ana	3	3 - 5.5 = -2.5
María	5	5 - 5.5 = -0.5
Teresa	4	4 - 5.5 = -1.5
José	6	6 - 5.5 = 0.5

$$2.5 + 1.5 - 2.5 - 0.5 - 1.5 + 0.5 = 0$$

Gráficamente tenemos:



Los resultados anteriores verifican el por qué la suma de los desvíos positivos son iguales a la suma de los desvíos negativos.

Actividad 3.4 c

Verifica esta propiedad de la media con los siguientes datos: 1, 2, 3, 4, 4, 4, 5, 6, 7.

Primero. Calcula la media:

$$\bar{X} = \frac{\Sigma x}{n} =$$

Segundo. Calcula los desvíos. Realiza tus cálculos en la tabla.

x_i	$d = x_i - \bar{X}$
1	
2	
3	
4	
4	
4	
5	
6	
7	

$\Sigma (x_i - \bar{X}) =$ ____

Desviación media

Como la suma de todos los desvíos es cero, tendremos que pensar en calcular el valor absoluto del desvío.

La **desviación media,** se define como la media de los valores absolutos las desviaciones de los datos de una variable con respecto a la media.

$$\textit{Desviación media: D.M.} = \frac{\sum_{i=1}^{N} |x_i - \bar{X}|}{n}$$

La desviación media siempre será positiva porque se obtiene a partir de una suma de números positivos.

Si la desviación media es muy pequeña, indica que hay una gran concentración de valores en torno a la media.

Ejemplo

Calcular la desviación media de los datos:

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

Solución

Para hallar la desviación media seguiremos cuatro pasos:

1. Calcular la media.
2. Se resta la media de cada dato de la variable, lo que produce la separación de cada dato respecto a la media (desviación).
3. Se toman los valores absolutos de cada desviación.
4. Se suman los valores absolutos de las desviaciones y el resultado se divide entre el total de datos.

Todos estos pasos se sistematizan en una tabla.

	x_i	Desvío $x_i - \bar{X}$	Valor absoluto $\lvert x_i - \bar{X}\rvert$
	6	-2.9	2.9
	10	1.1	1.1
	10	1.1	1.1
	10	1.1	1.1
	9	0.1	0.1
	7	-1.9	1.9
	10	1.1	1.1
	9	0.1	0.1
	10	1.1	1.1
	7	-1.9	1.9
	10	1.1	1.1
Sumas	98	$0.1 \approx 0$	13.5

Simplificaremos la notación omitiendo los subíndices de los datos x_i.

$\Sigma x = 98$

$n = 11$

$\bar{X} = \frac{\Sigma x}{N} = \frac{98}{11} = 8.9$

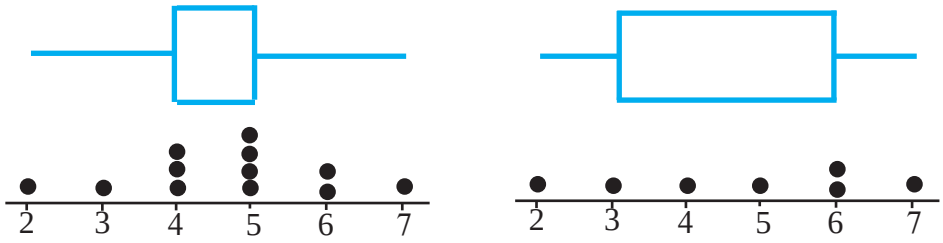
$\sum |x - \bar{X}| = 13.5$

$D.M. = \frac{\sum |x - \bar{X}|}{n} = \frac{13.5}{11} = 13.5$

Por tanto, los datos se desvían 1.23 unidades en promedio con respecto a su media.

Actividad 3.4 c

Calcula la desviación media de cada conjunto de datos indicados en los gráficos de puntos. En cada caso, observa el valor de la desviación media, la forma en que están distribuidos los datos y la forma de los gráficos de caja. Comenta los resultados.



Varianza y desviación estándar

El concepto de desviación media se origina cuando los desvíos se toman en valor absoluto eliminando así el efecto de $\Sigma(x - X) = 0$.

En el estudio de la llamada estadística matemática, las funciones juegan un papel muy relevante.

En matemáticas, efectuar operaciones con valores absolutos de funciones es, usualmente difícil. Por tal razón, y puesto que el propósito de los valores absolutos es eliminar el signo negativo de los desvíos, otra manera de lograr esto, es elevar al cuadrado cada uno de los desvíos y sumarlos de manera similar a como se hace con la desviación media.

Sin embargo, el cambio a elevar al cuadrado en vez de valores absolutos, no es el único. Por cuestiones más formales, cuando se trata de una muestra, esta suma de los cuadrados se divide entre $n-1$. Únicamente si se trabaja con una población, se permite la división entre ***n.*** Bajo estas consideraciones, se plantean las siguientes definiciones:

Varianza de una población, es el promedio de los cuadrados de los desvíos de los datos respecto a su media.

$$Varianza\ poblacional = \frac{\Sigma(x_i - \bar{X})^2}{N}$$

Varianza de una muestra, denotada por s^2, se calcula con:

$$s^2 = \frac{\Sigma(x_i - \bar{X})^2}{n-1}$$

Puesto que lo más común es que uses muestras en vez de poblaciones, debes concentrarte en la fórmula que divide entre ***n - 1.***

La **desviación estándar** denotada por **s**, es simplemente la raíz cuadrada de la varianza.

Procedimiento de cálculo de la varianza y desviación estándar

1. Se calcula la media y se resta de cada uno de los valores de la variable. Esto produce la desviación de cada valor a partir de la media $(x-\bar{X})$. Enseguida, se elevan al cuadrado estas desviaciones para obtener las desviaciones cuadráticas $(x-\bar{X})^2$.
2. Se efectúa la suma de las desviaciones cuadráticas respecto a la media: $\Sigma(x-X)^2$. Este valor se conoce brevemente como la suma de los cuadrados.

3. Se divide la suma de los cuadrados entre ***n -1.***
Este cociente representa la media de las desviaciones cuadráticas y es precisamente la **varianza s^2**.

4. Finalmente, para hallar la desviación estándar se extrae raíz cuadrada de la varianza:

Ejemplo

Calcular la varianza y desviación estándar de los datos correspondientes a las calificaciones bimestrales planteadas en el ejemplo de la página 70.

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

Solución

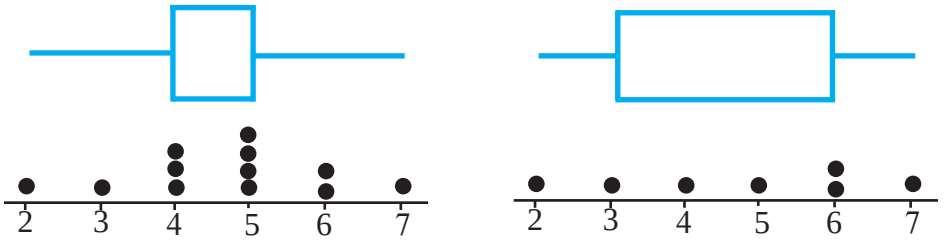
Preparamos una tabla con tres columnas: una para los datos ***x***, otra para las desviaciones de los datos con respecto a la media $(x-\bar{X})$ y otra más para las desviaciones cuadráticas $(x-\bar{X})^2$.

Observación. Para efectos de simplificación omitiremos los subíndices de x_i.

Paso 1 Calcular Σx	Paso 2 Calcular $\bar{x}$	Paso 3 Calcular los desvíos	Paso 4 Calcular $\Sigma\left(x-\bar{x}\right)^2$	Paso 5 Calcular s^2 y s
x		$x-\bar{X}$	$\left(x-\bar{X}\right)^2$	$s^2=\dfrac{\Sigma\left(x-\bar{x}\right)^2}{n-1}$
6	$\bar{x}=\dfrac{\Sigma x}{n}$	6 - 8.9 = -2.9	8.41	$s^2=\dfrac{22.91}{11-1}=2.291$
10		10 - 8.9 = 1.1	1.21	
10		10 - 8.9 = 1.1	1.21	
10	$\bar{x}=\dfrac{98}{11}$	10 - 8.9 = 1.1	1.21	
9		9 - 8.9 = 0.1	0.01	
7	$\bar{x}=8.9$	7 - 8.9 = -1.9	3.61	$s=\sqrt{2.291}=1.51$
10		10 - 8.9 = 1.1	1.21	
9		9 - 8.9 = 0.1	0.01	
10		10 - 8.9 = 1.1	1.21	
7		7 - 8.9 = -1.9	3.61	
10		10 - 8.9 = 1.1	1.21	
$\Sigma x=98$		$\Sigma x=0.1\approx 0$	$\sum\left(x-\bar{X}\right)=22.91$	

Actividad 3.4 d

Calcula la desviación estándar de cada conjunto de datos indicados en los gráficos de puntos. En cada caso, observa el valor de la desviación estándar, la forma en que están distribuidos los datos y la forma de los gráficos de caja. Comenta los resultados.



Método abreviado para calcular la varianza y desviación estándar

El cálculo de la varianza mediante la fórmula: $s^2 = \dfrac{\sum(x-\bar{X})^2}{N}$ es una tarea laboriosa hasta en el caso en que se tengan pocos datos. Para poder deducir una fórmula alternativa, aplicaremos un tratamiento algebraico a la fórmula que ya conocemos.

Primeramente, necesitamos de algunas propiedades de la sumatoria:

El símbolo Σx indica que hay que sumar los valores que toma la variable x. Si por ejemplo, x representa una variable que toma los valores: 1, 3, 5, 2, 4, entonces:

$$\Sigma x = 1+3+5+2+4 = 15$$

En general, la notación: $\sum_{i=1}^{N} x$ nos indica sumar todos los valores de x desde la posición 1 hasta la posición N. Es decir: $\sum_{i=1}^{N} x = x_1 + x_2 + \ldots + x_N$

Por ejemplo, con los mismos datos tenemos: $\sum_{i=1}^{3} x = 1+3+5 = 9$

(Nos indicaron sumar desde el primer valor hasta el tercero).

Para efectos de simplificación, el símbolo: Σx significará sumar todos los N valores de la variable.

Supongamos ahora que x toma N valores constantes; por ejemplo: $x = 3, 3, 3, 3, 3$.

Entonces: $\Sigma x = 3+3+3+3+3 = 5(3) = 15$.

Regla 1. La sumatoria de una constante que aparece N veces en un conjunto, es simplemente N veces la constante.

$$\sum_{i=1}^{N} C = NC \quad \text{ó} \quad \Sigma C = NC$$

Regla 2. La sumatoria de una variable multiplicada por una constante, es igual a la constante multiplicando a la sumatoria de los datos de la variable.

$$\Sigma cx = c\Sigma x$$

Por ejemplo: $x = 1, 3, 5, 5, 2, 4$ y $C = 10$

$$\begin{aligned}\Sigma cx &= 10(1)+10(3)+10(5)+10(2)+10(4)\\ &= 10(1+3+5+2+4)\\ &= 10\Sigma x\end{aligned}$$

Regla 3. Si después de la sumatoria se encuentra, entre paréntesis, una expresión que incluye sólo operaciones de suma o resta, la sumatoria puede ser distribuida entre los términos de la expresión:

$$\Sigma\left(x+y-c\right) = \Sigma x+\Sigma y-\Sigma c$$

Procederemos ahora, a simplificar la expresión: $s^2 = \dfrac{\sum\left(x-\bar{X}\right)^2}{n-1}$

$$s^2 = \frac{\sum\left(x-\bar{X}\right)^2}{n-1} = \frac{\Sigma\left[x^2-2x\bar{X}+\left(\bar{X}\right)^2\right]}{n-1}$$

$$= \frac{\Sigma x^2-2\bar{X}\Sigma x+\Sigma\left(\bar{X}\right)^2}{n-1}$$

$$= \frac{\Sigma x^2-2\bar{X}\dfrac{\Sigma x}{n}\bullet n+\Sigma\left(\bar{X}\right)^2}{n-1}$$

$$= \frac{\Sigma x^2-2n\left(\bar{X}\right)^2+n\left(\bar{X}\right)^2}{n-1}$$

$$= \frac{\Sigma x^2-n\left(\bar{X}\right)^2}{n-1}$$

$$= \frac{\Sigma x^2-n\left(\dfrac{\Sigma x}{n}\right)^2}{n-1}$$

$$= \frac{\Sigma x^2-\dfrac{\left(\Sigma x\right)^2}{n}}{n-1}$$

Hemos demostrado que: $\Sigma\left(x-\bar{x}\right)^2=\Sigma x^2-\dfrac{\left(\Sigma x\right)^2}{n}$.

Es frecuente que estas cantidades que son los numeradores en cada una de las fórmulas de la varianza, se denomine suma de cuadrados de *x* y se simbolice por *SC(x)*.

Por lo tanto, la fórmula simplificada de la varianza es:

$$s^2=\frac{\text{Suma de cuadrados de } x}{n-1}=\frac{SC(x)}{n-1}=\frac{\Sigma x^2-\dfrac{\left(\Sigma x\right)^2}{n}}{n-1}$$

Ejemplo

Calcular con el método simplificado la varianza y desviación estándar de los datos correspondientes a las calificaciones bimestrales planteadas en el ejemplo de la página 70.

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

Paso 1 Calcular Σx	Paso 2 Calcular Σx^2	Paso 3 Calcular *SC (x)*	Paso 4 Calcular s^2 y s
x	x^2	$SC(x)=\Sigma x^2-\dfrac{\left(\Sigma x\right)^2}{n}$	$s^2=\dfrac{SC(x)}{n-1}=\dfrac{22.91}{11-1}=2.291$
6	36	$SC(x)=896-\dfrac{(98)^2}{11}$	$s=\sqrt{2.291}=1.51$
10	100	$SC(x)=896-873.09$	
10	100	$SC(x)=22.91$	
10	100		
9	81		
7	49		
10	100		
9	81		
10	100		
7	49		
10	100		
$\Sigma x=98$	$\Sigma x^2=896$		

Actividad 3.4 d

Calcula con el método simplificado la varianza y desviación estándar de los gráficos presentados en la actividad (3.4 c)

Interpretación de la desviación estándar

Una consecuencia de la definición de desviación estándar es que cuando la desviación estándar es pequeña, la *media es una buena representante* de los valores del conjunto de datos. Por otra parte, una desviación estándar mayor indica que la media es menos representativa de la localización de los datos.

Ejemplo

Los siguientes conjuntos de datos corresponden a las estaturas en centímetros de los estudiantes de dos clases de educación física. La clase I era mixta y la clase II era solamente para hombres. Calcula la media y desviación estándar para cada clase y analice lo que indica.

Clase I

156	158	182	178	159	176
160	176	174	166	160	172
189	187	154	152	180	198
159	162	176	180	166	192

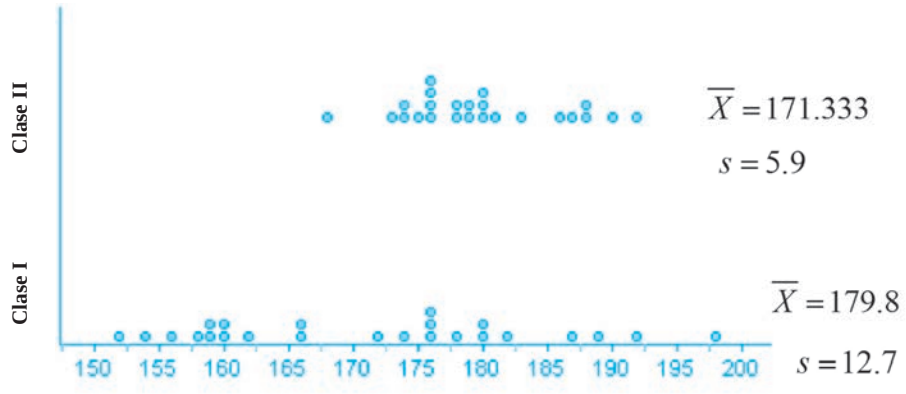
Media 171.333
Desviación estándar: 12.69

Clase II

168	180	183	180	187	190
176	176	178	188	186	174
179	192	188	176	179	181
173	174	180	178	176	175

Media: 179.875
Desviación estándar: 5.88

Con el objeto de ver mejor la relación entre la media y la desviación estándar, representamos los dos conjuntos de datos en gráficos de puntos localizados en el mismo sistema de ejes.



Observa que la clase I comparada con la II, muestra una gran variabilidad o dispersión de los datos. Por consiguiente, la desviación estándar de la clase I es mucho mayor que la de la clase II. La desviación estándar nos indica además que en el caso de la clase II, la media puede representar la localización de los datos no así en la clase I.

Ejercicio 3.4

1. Calcula la varianza y la desviación estándar de los conjuntos de datos siguientes.
   (a) 3, 2, 8, 4, 5, 7, 11, 6, 10.
   (b) 4, 3, 5, 4, 3, 8, 10, 7.
   (c) 4, 9, 8, 4, 8, 7, 6, 6.

2. Los datos siguientes, muestran los puntos totales ganados por los 18 equipos de futbol de primera división en el torneo de clausura 2009.

   36, 36, 28, 26, 26, 25, 23, 23, 22, 22, 21, 21, 21, 17, 17, 14, 14, 13.

   Calcula la varianza y la desviación estándar.

3. De 10 puntos posibles, un grupo de 20 estudiantes obtuvieron las siguientes calificaciones:

   0, 0, 1, 2, 4, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10.

   Calcula la varianza y la desviación estándar..

4. De 10 puntos posibles, un grupo de 20 estudiantes obtuvieron las siguientes calificaciones:

   0, 0, 1, 2, 4, 5, 5, 6, 6, 6, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10.

   Calcula la varianza y la desviación estándar.

5. Calcula la varianza y desviación estándar que corresponde al siguiente conjunto de medidas:

   40 kg, 50 kg, 60 kg, 70 kg

6. La siguiente tabla muestra las calificaciones en matemáticas y biología obtenidas por 12 alumnos. Para cada materia construye un gráfico de puntos y contesta las siguientes cuestiones:

Alumno	Calificación Matemáticas	Calificación Biología
1	2	1
2	3	3
3	4	2
4	4	4
5	5	4
6	6	4
7	6	6
8	7	4
9	7	6
10	8	7
11	6	9
12	7	10

   a) Sin hacer ningún cálculo, únicamente observando los gráficos, ¿en cuál materia hay más dispersión?_____

   b) Calcula la desviación estándar y verifica tu respuesta del inciso a).

Lección 3.5 Antecedente 5 para la exploración de datos cuantitativos: organización y representación de datos agrupados

Objetivo: Aprender a organizar y representar datos agrupados.

Actividad 12

Qué hacer

1) Consulta las **páginas 97 a 117** y al finalizar tu estudio contesta:

Los siguientes datos corresponden a las calificaciones obtenidas por 50 alumnos en un examen de aritmética.

56	72	70	75	49
58	53	77	51	61
72	46	68	69	63
69	61	80	47	58
88	65	70	64	66
60	50	78	64	49
61	59	91	57	56
48	89	56	63	56
73	44	72	61	46
46	59	88	70	74

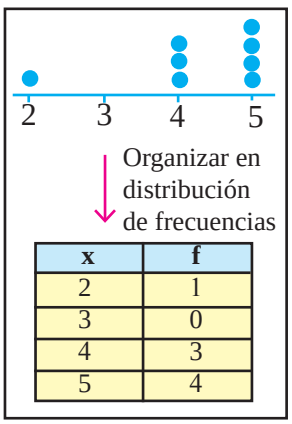
Construye:

a) Un gráfico de tallo y hojas

b) Una distribución de frecuencias agrupadas.

c) Un histograma

d) Una ojiva.



x	f
2	1
3	0
4	3
5	4

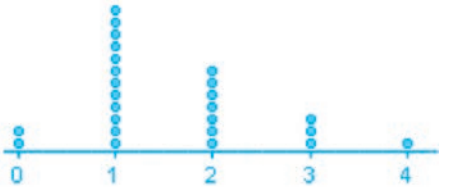
El gráfico de puntos, se ha utilizado como una herramienta fundamental para apoyar gran parte del contenido estadístico, que se ha desarrollado hasta este momento. Es una buena práctica empezar la exploración de datos construyendo un gráfico de puntos. Una ventaja destacada de este gráfico es que presenta los datos ordenados según su valor numérico, así como las frecuencias de cada uno de ellos. Esto permite que organizemos los datos de manera directa en una distribución de frecuencias.

Sin embargo, para que un gráfico de puntos o una distribución de frecuencias sea útil, una simple ojeada a ella, nos debe indicar algo relevante; **un patrón** o **tendencia de los datos**. La búsqueda de estos patrones es uno de los objetivos de la estadística. En algunos casos, el gráfico de puntos resulta insuficiente para este propósito.

Por ejemplo, observa los siguientes conjuntos de datos.

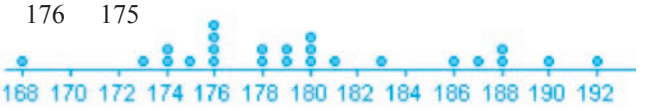
Número de automóviles de 25 familias

0, 1, 2, 3, 1, 0, 1, 1, 1, 4, 3, 2, 2,
1, 1, 2, 2, 1, 1, 1, 2, 1, 3, 2, 1



Estaturas en cm de 24 estudiantes

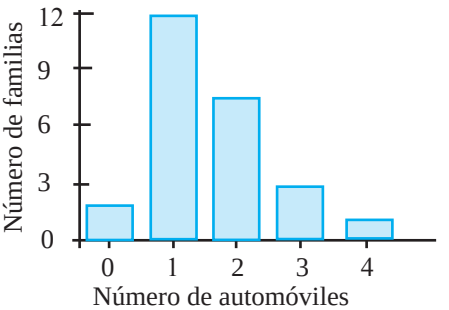
168	180	183	180	187	190
176	176	178	188	186	174
179	192	188	176	179	181
173	174	180	178	176	175



El primer gráfico, representa los datos de manera suficiente para poder darnos una idea visual de la distribución de los datos. Por ejemplo, se aprecia que el valor modal es 1 y que la distribución presenta un sesgo hacia la derecha.

A partir del gráfico de puntos, podemos obtener de manera directa la distribución de frecuencias y un gráfico de barras.

Valores distintos (modalidades) x	**Frecuencia** f
0	2
1	12
2	7
3	3
4	1



El **gráfico de barras** es recomendable para la representación de datos cualitativos. Sin embargo, también puede usarse para datos **cuantitativos** discretos que presentan **pocos valores distintos.**

Con respecto al gráfico de puntos correspondiente al segundo conjunto de datos, puede apreciarse una gran dispersión y se presentan muchos valores sin frecuencia por lo que es difícil identificar un patrón o tendencia. Además, si quisiéramos obtener una distribución de frecuencias esta quedaría de la siguiente manera:

Valores distintos	**Conteo**	f
168	/	1
169		0
170		0
171		0
172		0
173	/	1
174	//	2
175	/	1
176	////	4
177		0
178	//	2
179	//	2
180	///	3
181	/	1

Valores distintos	**Conteo**	f
182		0
183	/	1
184		0
185		0
186	/	1
187	/	1
188	//	2
189		0
190	/	1
191		0
192	/	1

Por estas razones, debemos mejorar la estrategia de exploración.

Gráfico de tallo y hoja

La estrategia fundamental para mejorar el aspecto visual de la distribución de los datos, consiste en **agruparlos en intervalos.** Cabe aclarar, que este agrupamiento se vuelve necesario al trabajar datos cuantitativos discretos con muchos valores distintos o datos contínuos.

Una técnica que agrupa los datos y a la vez los representa gráficamente, es el gráfico de ***tallo y hoja.***

Gráfico de tallo y hoja. Gráfico que presenta los datos en renglones mediante sus dígitos. Cada renglón empieza con uno o más dígitos y representa una posición de tallo, y cada dígito a la derecha de una recta vertical se puede considerar como una hoja.

Para construir un gráfico de tallo y hoja, cada dato numérico se divide en dos partes: El (los) dígito (s) principal (es) se convierte (n) en el tallo, y el dígito posterior se convierte en la hoja. Los tallos se escriben a lo largo del eje principal, y por cada dato se escribe una hoja.

Ejemplo

Los siguientes datos corresponden a la *variable edad* de los diputados del congreso local de Sinaloa.

50, 58, 33, 35, 38, 40, 45, 46, 50, 52, 55, 57, 58, 61, 65, 33, 54, 28, 34, 41, 42, 42, 44, 46, 52, 52, 54, 57, 60, 30, 38, 41, 43, 45, 48, 52, 53, 55, 58, 65.

Verifica que el gráfico de puntos correspondiente a estos datos, es el siguiente:

Edades

Puesto que este gráfico no nos muestra un patrón bien definido, procederemos a agrupar y representar los datos en un gráfico de tallo y hoja.

A simple vista se observa que *hay edades en las decenas*: 20, 30, 40, 50 y 60.

Como tallo se utilizará el primer dígito de cada edad; como hoja, el segundo.

Se traza una recta vertical y se escriben los tallos, en orden, a la izquierda de la recta.

2
3
4
5
6

A continuación, se coloca cada hoja en su tallo. Esto se hace escribiendo el último dígito a la derecha de la recta vertical, enfrente de su dígito principal correspondiente.

Ejemplo

El primer dato es 50; el tallo es 5 y la hoja es 0. Así, enfrente del tallo 5 se escribe 0.

2	
3	
4	
5	0 ← Representación del dato 50
6	

El siguiente dato es 58, de modo que enfrente del tallo 5 y a la derecha del 0 se escribe una hoja 8.

2	
3	
4	
5	0 8 ← El gráfico ya tiene registrados, dos datos: el 50 y el 58.
6	

El siguiente dato es 33, por lo que enfrente del tallo 3 se escribe una hoja 3.

2	
3	3
4	
5	0 8 ← El gráfico ya tiene registrados, los tres primeros datos: el 50, 58 y 33.
6	

Se continúa hasta que cada una de las 37 hojas restantes se escriben en el gráfico. A continuación se muestra todo el gráfico de tallo y hoja.

Edades de diputados locales

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

El gráfico de tallo y hoja, también nos permite ordenar los datos.

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

De esta presentación se obtienen los datos ordenados:

28, 30, 33, 33, 34, 35, 38, 38, 40, 41, 41, 42, 42, 43, 44, 45, 45, 46, 46, 48, 50, 50, 52, 52, 52, 52, 53, 54, 54, 55, 55 , 57, 57, 58, 58, 58, 60, 61, 65, 65.

¿En qué situaciones nos sirven los datos ordenados?_____

_____

Lenguaje de los intervalos

A continuación a partir del gráfico de tallo y hoja construido, precisaremos algunos términos relativos al agrupamiento en intervalos.

Edades de diputados locales

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

Todas las edades con el mismo dígito en las decenas se escribieron en la misma rama. Esto significa que en la primer rama de este ejemplo, *los posibles valores diferentes* que se incluirán son:

20, 21, 22, 23, 24, 25, 26, 27, 28 y 29

Estos posibles valores forman un intervalo el cual puede describirse como: «*todos los valores mayores o iguales a 20 y menores que 30*». En notación de intervalo escribimos: [20,30). Obsérvese que para esta descripción, estamos usando el 30 (número natural que sigue a 29). Se toma esta decisión por una cuestión técnica que después cobrará importancia.

Recuerda que la notación de intervalo [20,30), incluye todos los valores x tales que: $20 \leq x < 30$.

Entonces nuestro gráfico de tallo y hoja, también puede escribirse:

Tallo	Intervalo		
2	[20,30)	2	8
3	[30,40)	3	0 3 3 4 5 8 8
4	[40,50)	4	0 1 1 2 2 3 4 5 5 6 6 8
5	[50,60)	5	0 0 2 2 2 2 3 4 4 5 5 7 7 8 8 8
6	[60,70)	6	0 1 5 5

La **amplitud** del intervalo denotada por i es 10 puesto que incluye 10 valores potenciales.

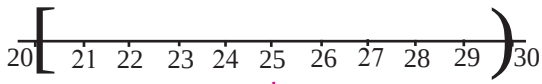
El 20 se llama **límite inferio**r del intervalo, y el 30 límite superior.

La amplitud es igual a la diferencia entre el límite superior y el límite inferior.

$$Amplitud = i = Límite \text{ superior } - Límite \text{ inferior}$$

La **marca de clase** es el punto medio del intervalo, por lo que, para encontrar la marca de clase de un intervalo, se suman los límites inferior y superior y el resultado se divide entre 2.

Marca de clase de [20,30) es: $\frac{20+30}{2} = 25$



Marca de clase = $\frac{20+30}{2} = 25$

Actividad 3.5 a

Calcula la marca de clase de los demás intervalos.

Intervalo	Marca de clase	Tallo	Hoja
[20,30)	2	2	8
[30,40)	3	3	0 3 3 4 5 8 8
[40,50)	4	4	0 1 1 2 2 3 4 5 5 6 6 8
[50,60)	5	5	0 0 2 2 2 2 3 4 4 5 5 7 7 8 8 8
[60,70)	6	6	0 1 5 5

Distribuciones de frecuencias agrupadas

Así como el gráfico de puntos nos permite construir directamente una distribución de **frecuencias no agrupadas**, el de tallo y hoja origina una distribución de **frecuencias agrupadas.**

Del gráfico de tallo y hoja anterior, se obtiene la siguiente distribución de frecuencias agrupadas.

Intervalos	Frecuencia *f*
[20,30)	1
[30,40)	7
[40,50)	12
[50,60)	16
[60,70)	4
Total	40

Recuerda que con esta notación, si un dato coincide con el límite superior del intervalo, debe incluirse en el siguiente intervalo. Por ejemplo el dato 30, no se incluye en el primer intervalo sino en el segundo.

Histograma

El gráfico de barras es uno de los recursos gráficos para una distribución de frecuencias no agrupadas. Asimismo, una representación gráfica parecida al gráfico de barras denominada **histograma,** se usa para representar distribuciones de frecuencias agrupadas.

Histograma, es un gráfico que se utiliza para datos agrupados en intervalos. Un histograma consta de los siguientes componentes:

1. Un **título**, que identifica la población o muestra de interés.
2. Una **escala vertical**, que identifica las frecuencias de los diversos intervalos.
3. Una **escala horizontal,** que identifica la variable estudiada. A lo largo del eje horizontal pueden marcarse los límites de los intervalos o las marcas de clase. Se debe utilizar el método de marcar el eje que mejor presente la variable.

Para construir el histograma se levantan rectángulos que tienen como base la longitud de los distintos intervalos y una altura tal que el área del rectángulo sea proporcional a la frecuencia correspondiente al intervalo.

Cuando los intervalos son de la misma longitud, la altura es igual a la frecuencia; si no es así, hay que modificar la altura para mantener la proporción entre el área y la frecuencia correspondiente. Lo más recomendable es usar intervalos de igual amplitud.

A continuación se presenta todo el proceso que nos permitió llegar hasta el trazo del histograma de los datos de la variable *edad de diputados.*

50, 58, 33, 35, 38, 40, 45, 46, 50, 52, 55, 57, 58, 61, 65, 33, 54, 28, 34, 41,
42, 42, 44, 46, 52, 52, 54, 57, 60, 30, 38, 41, 43, 45, 48, 52, 53, 55, 58, 65.

↓

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

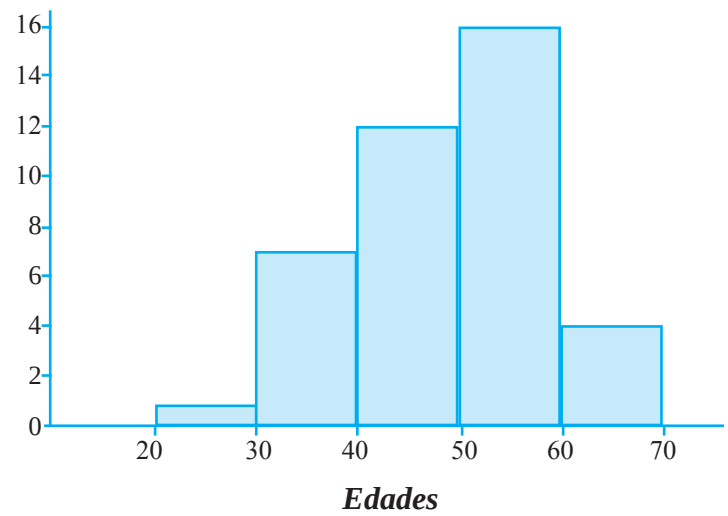
↓

Intervalos	Frecuencia *f*
[20,30)	1
[30,40)	7
[40,50)	12
[50,60)	16
[60,70)	4
Total	40

→

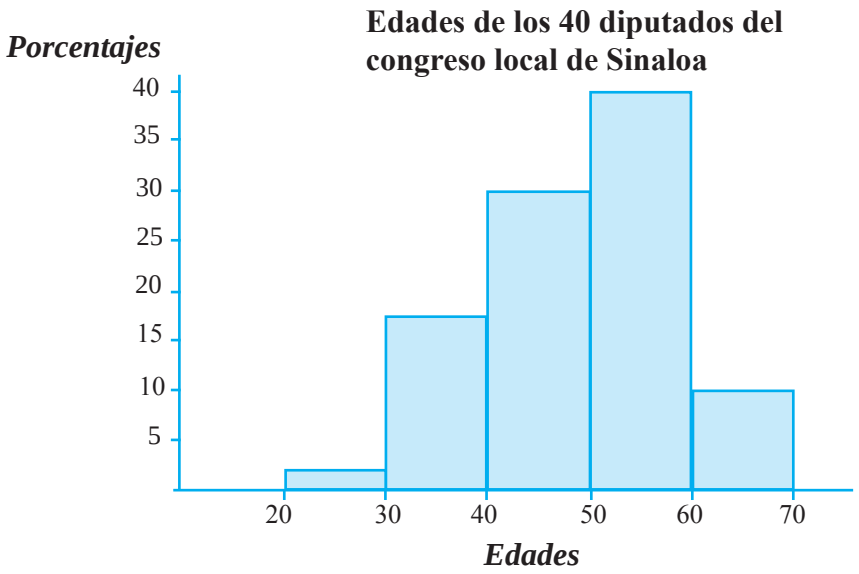
Frecuencia de edades

Edades de los 40 diputados del congreso local de Sinaloa



Recordemos que el nombre de distribución de frecuencias, se refiere a las frecuencias absolutas. Si en vez de presentar las frecuencias absolutas presentamos las relativas, el nombre cambia a distribución de frecuencias relativas. Recuerda que la frecuencia relativa es una medida proporcional de la frecuencia para que ocurra un dato. En este caso se encuentra dividiendo la frecuencia de cada intervalo entre el número total de datos. Esta división tiene como resultado un número decimal. Por lo general, es más útil manejar las frecuencias relativas como porcentajes. Para convertir la frecuencia relativa a porcentaje recuerda que simplemente multiplicamos el número decimal obtenido, por 100. Verifica las frecuencias relativas de la distribución que estamos estudiando.

Intervalos	**Frecuencia** f	**Frecuencia relativa**	**Porcentajes**
[20,30)	1	0.025	2.5 %
[30,40)	7	0.175	17.5 %
[40,50)	12	0.300	30 %
[50,60)	16	0.400	40 %
[60,70)	4	0.100	10 %
Total	40	1.000	100 %



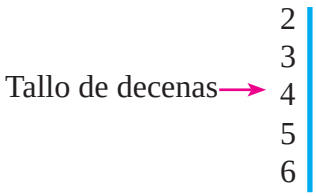
Influencia en el histograma de la amplitud de los intervalos

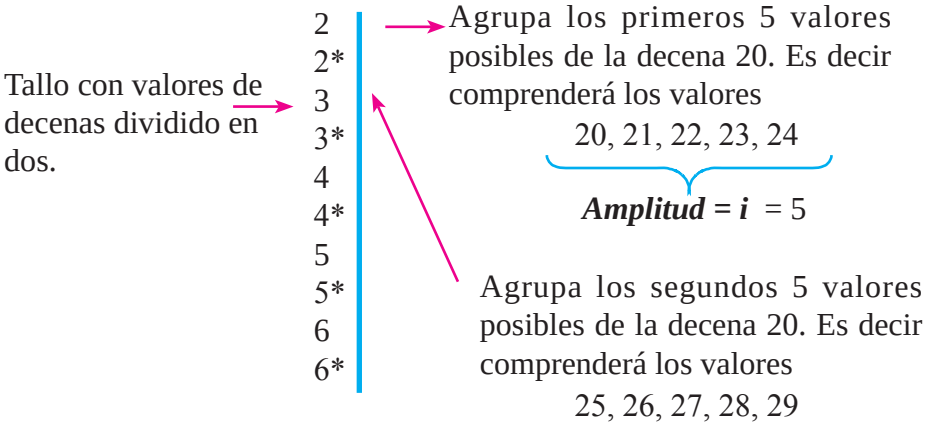
Desafortunadamente, no siempre se logra esto al primer intento. En muchos casos para lograr que el histograma proyecte una imagen adecuada, debemos manipular la amplitud del intervalo o el númereo de ellos. Debemos entender por imagen adecuada, aquella que tenga una forma de distribución muy semejante a la de la población de la cual se extrajo la muestra.

En el caso de la variable que estamos estudiando «*edades de diputados*», el usar una amplitud de intervalo igual a10, fue suficiente. Sin embargo, seguiremos con esta variable para mostrar el efecto que tiene en la forma del histograma el cambio de amplitud.

La amplitud igual a 10, significa que en cada tallo, del gráfico de tallo y hoja, habrá 10 valores posibles. Ahora, reconstruiremos el gráfico agrupando los valores de modo que en cada tallo puedan escribirse sólo 5 valores posibles.

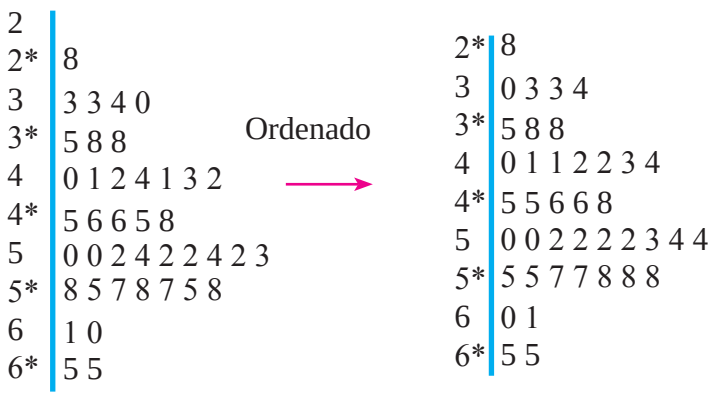
Esto significa que cada tallo representando una decena, se dividirá en dos tallos. El segundo tallo de cada decena se escribirá con un asterisco.





Ahora, agruparemos los datos originales en un tallo dividido:

50, 58, 33, 35, 38, 40, 45, 46, 50, 52, 55, 57, 58, 61, 65, 33, 54, 28, 34, 41, 42, 42, 44, 46, 52, 52, 54, 57, 60, 30, 38, 41, 43, 45, 48, 52, 53, 55, 58, 65.



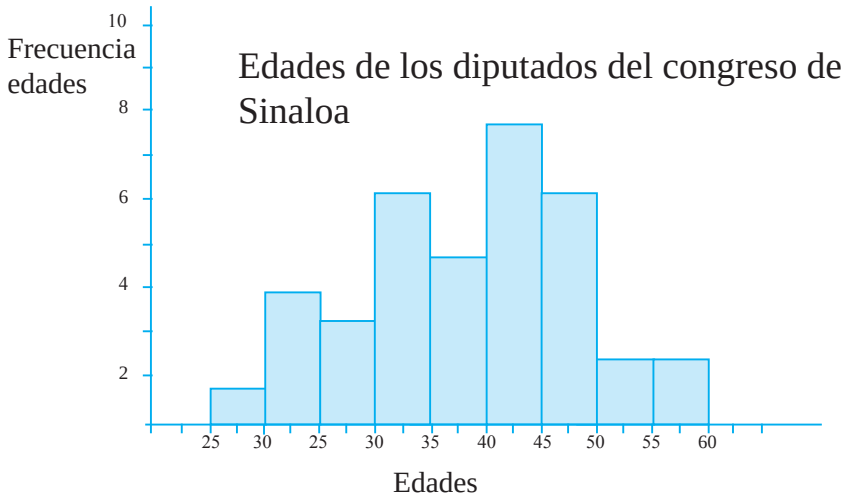
Distribución de frecuencias agrupadas para $i = 5$.

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

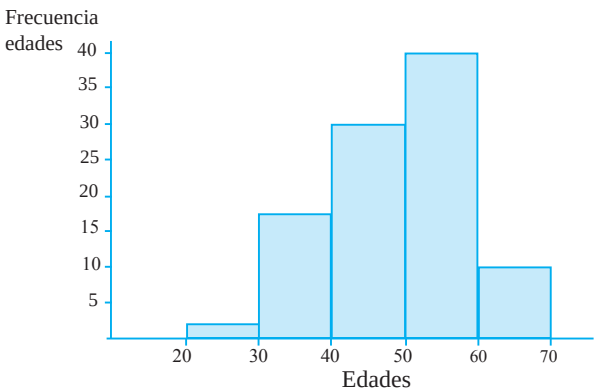
→

Intervalos	Frecuencia f
[25,30)	1
[30,35)	4
[35,40)	3
[40,45)	7
[45,50)	5
[50,55)	9
[55,60)	7
[60,65)	2
[65,70)	2
Total	40

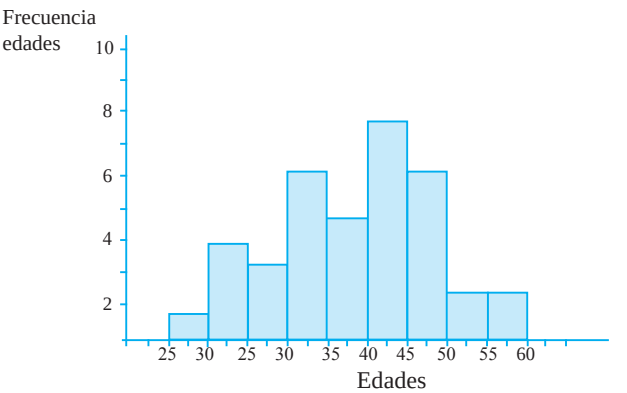
El histograma correspondiente es:



Observa la influencia que tiene el cambio de valor en la amplitud de los intervalos.



$i = 10$



$i = 5$

Revisemos otro ejemplo.

Ejemplo

A continuación se presentan las estaturas en cm de 24 estudiantes.
a) Investiga si es necesario agrupar los datos en intervalos.
b) Si es necesario agrupar, dibuja el gráfico de tallo y hoja.
c) Construye la distribución defrecuencias agrupadas y el histograma.

Estaturas en cm de 24 estudiantes

168	180	183	180	187	190
176	176	178	188	186	174
179	192	188	176	179	181
173	174	180	178	176	175

Solución

a) Para investigar si hace falta agrupar los datos, construimos un gráfico de puntos.

Debido a que los datos muestran mucha dispersión, hay valores sin frecuencia y no hay un patrón definido, concluimos que deben agruparse los datos.

b) Gráfico de tallo y hoja

A simple vista se observa que *las estaturas están en las decenas :* 160, 170, 180 y 190.

Puesto que las hojas siempre deben ser un dígito, se utilizarán como tallo los dos primeros dígitos de las estaturas.

Estaturas de estudiantes

16	8
17	6 9 3 6 4 8 6 8 9 6 4 0
18	0 3 8 0 0 8 7 6 1
19	2 0

Ordenado →

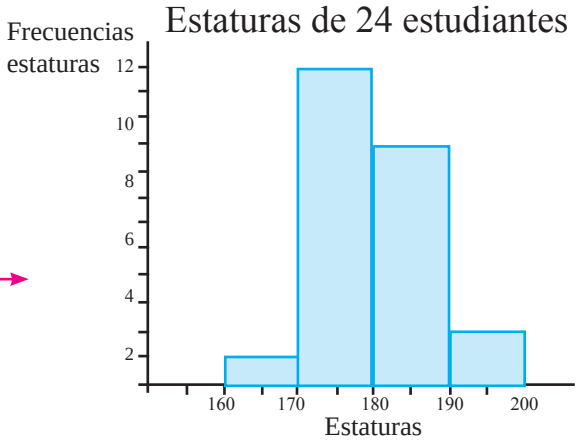
Estaturas de estudiantes

16	8
17	0 3 4 4 6 6 6 6 8 8 9 9
18	0 0 0 1 3 6 7 8 8
19	0 2

↓ Distribución de frecuencias agrupadas

Intervalos	Frecuencia *f*
[160,170)	1
[170,180)	12
[180,190)	9
[190,200)	2
Total	24

→



Actividad 3.5 b

Reconstruye el gráfico de tallo y hoja de la variable *estaturas* del ejemplo anterior, agrupando los valores de modo que en cada tallo puedan escribirse sólo 5 valores posibles.
A partir de este agrupamiento construye:

a) La distribución de frecuencias agrupadas.
b) El histograma
c) La distribución de frecuencias relativas y su histograma correspondiente.

Procedimiento general para formar una distribución de frecuencias agrupadas

El proceso de tallo y hoja fue utilizado para construir una distribución de frecuencias agrupadas y su correspondiente histograma; no obstante, la representación de tallo y hoja no es fácil de construir para amplitudes de intervalos diferentes de 5 y 10. Por ejemplo, si algún conjunto de datos necesita una amplitud de intervalo de 3, 4 ó 6, el procedimiento que acabamos de estudiar que consiste en partir de un gráfico de tallo y hoja, resulta muy difícil de desarrollar. Por lo tanto, es necesario tener un procedimiento general para construir distribuciones de frecuencias agrupadas. Ilustraremos el procedimiento mediante un ejemplo.

Ejemplo

Los siguientes datos, representan las edades en meses de 30 estudiantes de tercero de preparatoria. Construir una distribución de frecuencias agrupadas y su histograma correspondiente.

200, 205, 192, 203, 208, 218, 216, 209, 205, 192,
201, 202, 207, 209, 211, 208, 210, 214, 216, 215,
227, 205, 200, 208, 210, 215, 222, 216, 218, 216

Procedimiento para construir una distribución de frecuencias agrupadas.

Primer paso. Identifica la edad máxima ($x_{máx}$) y la edad mínima ($x_{mín}$) y calcula el rango:

$$R = x_{máx} - x_{mín} = 227 - 192 = 35$$

Segundo paso. Se elige el número de intervalos. No hay reglas absolutas para determinar el número de intervalos. La práctica ha enseñado que el número de intervalos no debe ser menor de 5 ni mayor de 20. La cantidad $\sqrt{n}$ (donde n es el tamaño de la muestra) da ***un valor aproximado para el número apropiado de intervalos***; cinco intervalos cuando n = 25, diez intervalos cuando n = 100, etcétera.

En nuestro ejemplo, si denotamos por k el número de intervalos, tenemos:

$$k = \sqrt{30} = 5.47 \approx 6$$

Tomaremos 6 intervalos.

Tercer paso. Se determina la amplitud o anchura de los intervalos la que hemos denotado por ***i.*** Para ello, se divide el rango entre el número de intervalos (en la práctica, esta división rara vez es exacta; en estos casos se recomienda redondear al entero siguiente).

$$\textit{Amplitud de los intervalos} = i = \frac{R}{\textit{Número de intervalos}} = \frac{}{} = 5.8 \approx 6$$

Se recomienda redondear al entero siguiente para que se cumpla la siguiente condición:

$$\textit{(Número de intervalos) (amplitud de intervalo)} > \textit{Rango}$$

$$k \cdot i > R$$

Se plantea esta condición para asegurar que todos los datos queden incluidos en los intervalos previstos.

(Sin embargo, aún así, en aquellos casos en los que este producto sea ligeramente mayor que *R,* tal vez no sea suficiente el número de intervalos determinado y tengamos que agregar otro más).

En nuestro ejemplo:

$$k \cdot i = 6 \times 6 = 36 > R$$

Cuarto paso. Se selecciona un punto inicial; éste puede ser el dato mínimo o *uno un poco menor que él*. El punto inicial se toma como el límite inferior del primer intervalo y se le suma (***i***) para obtener el límite superior de ese intervalo.

En nuestro ejemplo, el dato menor es 192. Entonces, vamos a tomar como punto inicial al 190.

[*Punto inicial, punto inicial* + *i*)

Primer intervalo:

[190,190 + 6) → [190,196) → *Incluye todos los dato que sean 190 o más a menos de 196*

Segundo intervalo: El límite inferior del segundo intervalo será el límite superior del primer intervalo, 196 en nuestro ejemplo, al cual se le suma (***i*** = 6) para obtener el límite superior:

[196,196 + 6) → [196,202) → *Incluye todos los dato que sean 196 o más a menos de 202.*

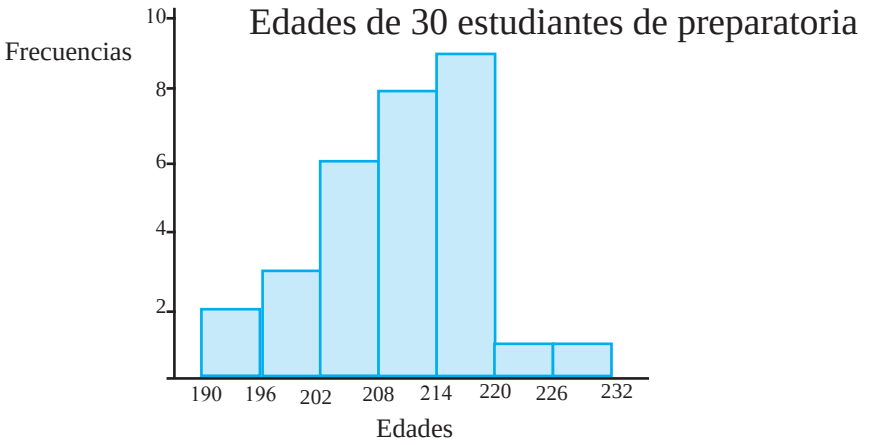
El límite inferior del tercer intervalo es 202 y el superior 208, y así sucesivamente seguimos formando intervalos hasta que el último dato quede incluido.

Quinto paso. Una vez establecidos los intervalos, se realiza el *conteo de frecuencias;* para ello, se recorre la lista original en el orden en que está dada y conforme se avanza en este recorrido, se hacen marcas en la segunda columna enfrente del intervalo correspondiente. *Este paso termina cuando el dato mayor es incluido en el último intervalo.*

	Intervalos	Conteo	Frecuencia f
190 o más a menos de 196 →	[190,196)	//	2
196 o más a menos de 202 →	[196,202)	///	3
202 o más a menos de 208 →	[202,208)	𝍸 /	6
208 o más a menos de 214 →	[208,214)	𝍸 ///	8
214 o más a menos de 220 →	[214,220)	𝍸 ////	9
220 o más a menos de 226 →	[220,226)	/	1
226 o más a menos de 232 →	[226,232)	/	1
	Total		30

El procedimiento de agrupamiento en intervalos está terminado. Nos hemos privado de las cifras reales y en su lugar aceptamos una pequeña lista que solamente indica el número de valores en cada uno de los intervalos. Por ejemplo, sabemos que existen 6 valores entre 202 y 208 (o igual a 202) pero no sabemos dónde están. Perdemos un poco de exactitud, pero se gana en el conocimiento del patrón. Es decir, con la distribución de frecuencias agrupadas, se logra la presentación de la masa de datos en una sencilla estructura que facilita el hallazgo de las relaciones que puedan haber entre ellos, pero, se da una pérdida inevitable de una buena parte del detalle original de los datos; *esta pérdida de detalle será mayor cuanto menor sea el número de intervalos, o lo que es lo mismo, cuanto mayor sea la amplitud de los intervalos.* Con todo, la sola ventaja del hallazgo de un patrón, justifica el agrupamiento de datos. Probablemente la distribución de frecuencias agrupadas es la forma fundamental de presentar datos estadísticos con fines de análisis.

Construyamos ahora el histograma correspondiente a esta distribución de frecuencias.



Polígono de frecuencias

Son gráficos lineales que se utilizan en el caso de una variable cuantitativa. Al igual que el histograma, el polígono de frecuencias debe incluir los siguientes componentes:

1. Un **título**, que identifica la población o muestra de interés.
2. Una **escala vertical**, que identifica las frecuencias de los diversos intervalos.
3. Una **escala horizontal,** que identifica la variable estudiada. A lo largo del eje horizontal se escriben las marcas de clase.

Ejemplo

Vamos a construir el polígono de frecuencias asociado al último conjunto de datos relacionado con la variable *edades de estudiantes.*

Intervalos	Frecuencia f
[190,196)	2
[196,202)	3
[202,208)	6
[208,214)	8
[214,220)	9
[220,226)	1
[226,232)	1
Total	30

Cálculo de marcas de clase:

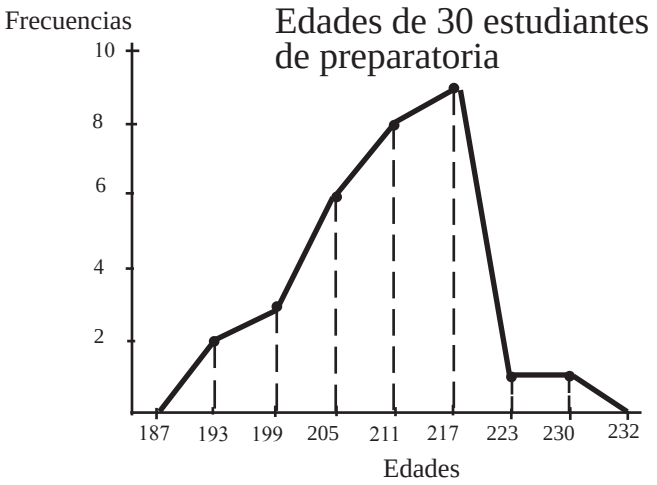
$$\frac{190+196}{2} = 193$$

$$\frac{196+202}{2} = 199$$

$$\frac{202+208}{2} = 205$$

Intervalos	Marca de clase	Frecuencia f
[190,196)	193	2
[196,202)	199	3
[202,208)	205	6
[208,214)	211	8
[214,220)	217	9
[220,226)	223	1
[226,232)	229	1
Total		30

Ahora, trazamos dos ejes perpendiculares: en el eje horizontal se localizan las marcas de clase de cada intervalo, y en el vertical las frecuencias. Para cada marca de clase, levantamos sobre el eje horizontal una línea de altura igual a su frecuencia, y colocamos un punto a esta altura. A continuación unimos puntos consecutivos con segmentos rectos. Para que este polígono quede cerrado, se agregan al inicio y al final marcas de clase con frecuencia cero. Finalmente escribimos el título del polígono.



Actividad 3.5 c

A continuación se repite la distribución de frecuencias de la variable *estaturas* de 24 estudiantes realizada en la página 104.

Intervalos	Frecuencia f
[160,170)	1
[170,180)	12
[180,190)	9
[190,200)	2
Total	24

a) Construye el polígono de frecuencias para esta distribución.

b) Calcula la distribución de frecuencias relativas, y construye su polígono de frecuencias.

Ejemplo

Un investigador agrícola, investigando el crecimiento de árboles frutales, registró los datos siguientes sobre 40 árboles jóvenes. Los datos son las alturas de los árboles medidos en cm.

102.6	106.7	101.3	100.2	107.4
109.3	104.2	102.6	105.2	111.2
105.7	110.5	109.4	101.6	120.4
104.2	104.3	106.6	116.6	105.1
109.8	99.8	112.9	101.3	119.2
104.2	111.1	112.7	112.6	107.8
104.3	117.9	101.3	111.6	106.2
107.2	111.1	117.2	102.1	109.6

a) Realiza una distribución de frecuencias agrupadas.
b) Dibuja el histograma y el polígono de frecuencias.

Solución

Primer paso.

$$R = x_{máx} - x_{mín} = 120.4 - 99.8 = 20.6$$

Segundo paso.

$$Número\ de\ \text{intervalos}\ k = \sqrt{número\ de\ datos} = \sqrt{40} = 6.3$$

Usaremos 6 intervalos.

Tercer paso. $$Amplitud\ de\ \text{intervalos}\ i = \frac{R}{k} = \frac{20.6}{6} = 3.43$$

¿ Es $k \times i > R$?

$$(6)(3.43) = 20.58 < 20.6$$

Entonces, tomaremos ***i = 3.5*** ***(manejar un decimal porque los datos así están presentados)***

Ejemplo (Cont.)

Cuarto paso.

Primer intervalo: [*Punto inicial, punto inicial + i*)

Tomaremo como punto inicial el dato menor: 99.8

[99.8, 99.8 + 3.5) → [99.8, 103.3)

Segundo intervalo:

[103.3, 103.3 + 3.5)→ [103.3, 106.8)

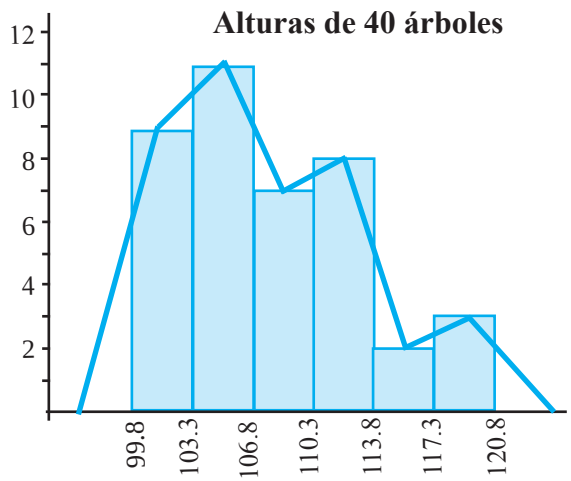
Quinto paso.

Completar los intervalos y efectuar el conteo

Intervalos	Conteo	Frecuencia *f*
[99.8, 103.3)	𝍸 ////	9
[103.3, 106.8)	𝍸 𝍸 /	11
[106.8, 110.3)	𝍸 //	7
[110.3, 113.8)	𝍸 //	7
[113.8, 117.3)	//	2
[117.3,120.8)	///	3
Total		40

Histograma y polígono de frecuencias

A continuación construimos el histograma. Uniendo con segmentos rectilíneos los puntos medios superiores de cada rectángulo, se construye el polígono de frecuencias. Esta es otra forma de trazar dicho polígono.



Distribución de frecuencias acumuladas para distribuciones no agrupadas

Las frecuencias acumuladas y su representación gráfica denominada ojiva, son una herramienta muy útil para determinar medidas de posición. Por tal razón, procederemos a estudiarlas.

En una distribución de frecuencias no agrupadas, la frecuencia acumulada para cualquier valor determinado, es la suma de la frecuencia para ese valor y las frecuencias de todos los valores menores. La frecuencia acumulada es más útil presentada como porcentaje.

Ejemplo Construiremos la distribución de frecuencias acumuladas para los datos de la variable *integrantes en la familia*

7, 2, 6, 8, 5, 6, 5, 6, 6, 5, 6, 4, 5, 7, 9, 5, 6, 6, 8, 9, 6, 8, 6, 12, 6, 8, 6, 9, 6, 8.

Valores	Frecuencia absoluta	Frecuencia acumulada	Frec. acumulada relativa en porcentajes	
2	1	1	3.3 %	← *3.3 % de las familias tienen 3 hijos o menos*
3	0	1	3.3 %	← *6.7 % de las familias tienen 4 hijos o menos*
4	1	2	6.7 %	← *23.3 % de las familias tienen 5 hijos o menos*
5	5	7	23.3 %	← *63.3 % de las familias tienen 6 hijos o menos*
6	12	19	63.3 %	← *70 % de las familias tienen 7 hijos o menos*
7	2	21	70.0 %	← *86.7 % de las familias tienen 8 hijos o menos*
8	5	26	86.7 %	← *96.7 % de las familias tienen 9 hijos o menos*
9	3	29	96.7 %	← *96.7 % de las familias tienen 10 hijos o menos*
10	0	29	96.7 %	← *96.7 % de las familias tienen 11 hijos o menos*
11	0	29	96.7 %	
12	1	30	100 %	
Total	30			

Ojiva o polígono de frecuencias acumuladas para distribuciones no agrupadas

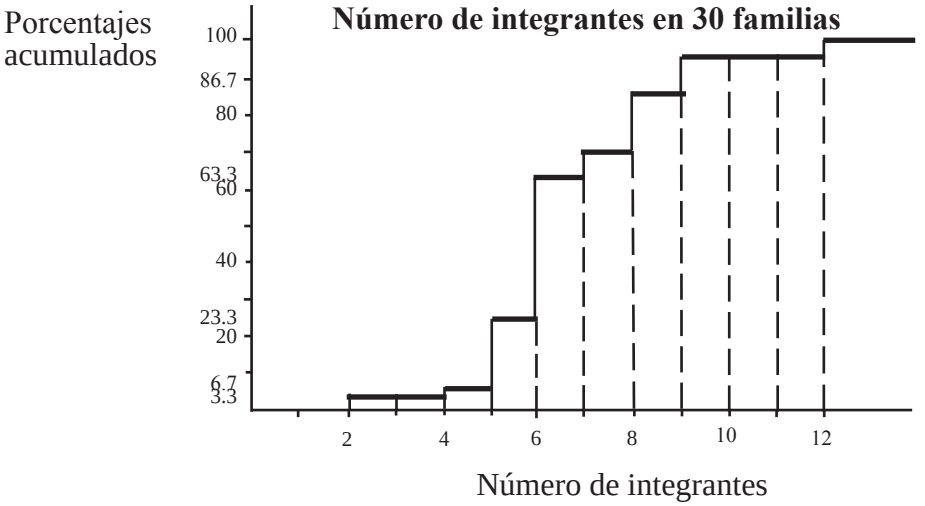
La representación gráfica de las distribuciones de frecuencias acumuladas se llama *ojiva o polígono de frecuencias acumulada.*

Al igual que el histograma y el polígono de frecuencias, una ojiva debe incluir los siguientes componentes:

1. Un **título**, que identifica la población o muestra de interés.
2. Una **escala vertical**, que identifica las frecuencias acumuladas, preferentemente las relativas expresadas en porcentajes.
3. Una **escala horizontal,** que identifica la variable estudiada. A lo largo del eje horizontal se escriben cada uno de los valores de la variable.

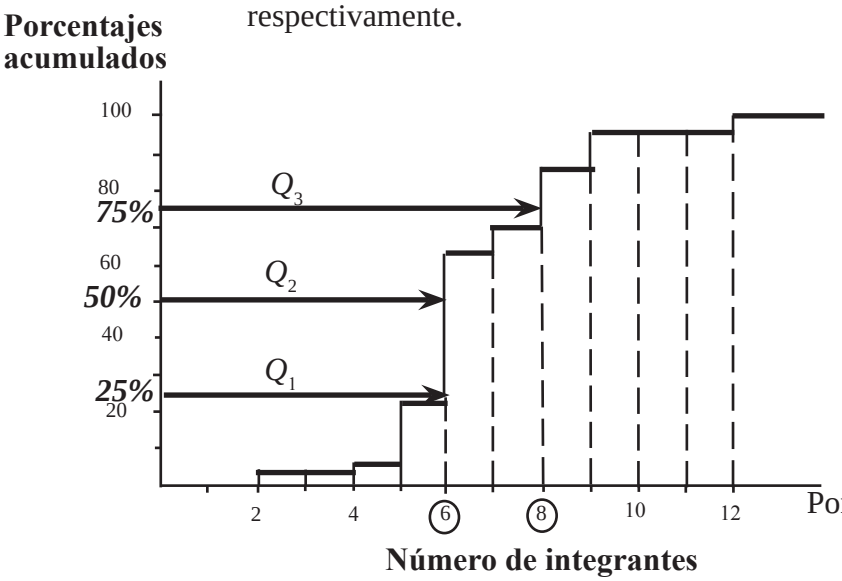
Procedimiento para construir una ojiva para distribuciones no agrupadas

Para cada uno de estos valores, levantamos sobre el eje de abscisas una línea de altura igual a la frecuencia acumulada; a continuación trazamos desde el extremo de cada línea una paralela al eje X, que corte a la línea siguiente. Después, se completa el diagrama, como se muestra en la siguiente figura. En esta gráfica podemos ver cómo las frecuencias acumuladas experimentan un aumento en cada valor de la variable.



La ojiva se puede usar para determinar las medidas de posición. Como ejemplo, usaremos la ojiva anterior para determinar los cuartiles del número de integrantes en las familias.

Para el primer cuartil, trazamos a la altura que indica 25% del eje Y, una horizontal hasta que toque la gráfica; a partir de ese punto , trazar una vertical hasta que cruce el eje X. Este punto de cruce será el primer cuartil. Para el segundo y tercer cuartil, se procede de manera idéntica, pero la recta horizontal se traza a partir de los porcentajes correspondientes a 50 y 75 respectivamente.



Por lo tanto, los cuartiles son

$Q_1 = 6$

$Q_2 = Med = 6$

$Q_3 = 8$

Distribución de frecuencias acumuladas para distribuciones agrupadas

En una distribución de frecuencias agrupadas, la frecuencia acumulada para cualquier intervalo determinado, es la suma de la frecuencia para ese intervalo y las frecuencias de todos los valores menores.

Ejemplo Construiremos la distribución de frecuencias acumuladas para los datos de la variable *edades de estudiantes*

200, 205, 192, 203, 208, 218, 216, 209, 205, 192,
201, 202, 207, 209, 211, 208, 210, 214, 216, 215,
227, 205, 200, 208, 210, 215, 222, 216, 218, 216

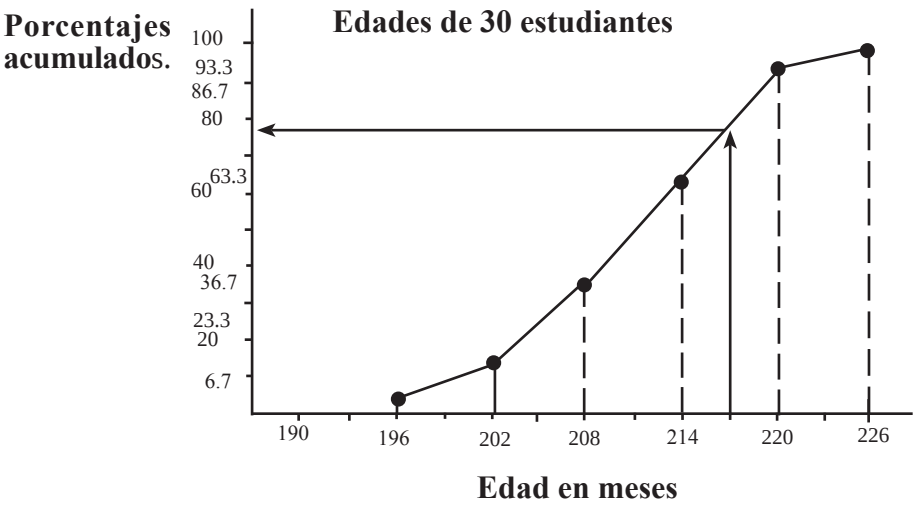
Intervalos	Frecuencia absoluta	Frecuencia acumulada	Frec. acumulada relativa en porcentajes
[190,196)	2	2	6.7 %
[196,202)	3	5	16.7 %
[202,208)	6	11	36.7 %
[208,214)	8	19	63.3 %
[214,220)	9	28	93.3 %
[220,226)	1	29	96.7 %
[226,232)	1	30	100 %
Total	30		

← *6.7% de los estudiantes tienen menos de 196 meses de edad*
← *16.7% de los estudiantes tienen menos de 202 meses de edad.*
← *36.7% de los estudiantes tienen menos de 208 meses de edad.*
← *63.3% de los estudiantes tienen menos de 214 meses de edad.*
← *93.3% de los estudiantes tienen menos de 220 meses de edad.*
← *96.7% de los estudiantes tienen menos de 226 meses de edad.*
← *100% de los estudiantes tienen menos de 232 meses de edad.*

Ojiva para distribuciones de frecuecnias agrupadas

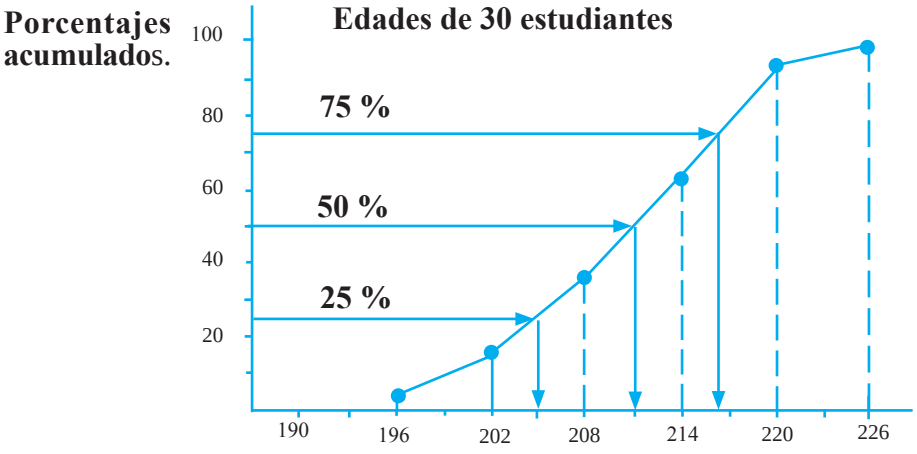
La ojiva para datos agrupados tiene los mismos componentes que para datos no agrupados, pero en este caso, en el eje horizontal, se escriben los límites superiores de cada intervalo. También cambia el trazado de la curva.

Tracemos la ojiva para la variable edades.



Los polígonos de frecuencias acumuladas son útiles porque permiten responder distintas preguntas sin necesidad de cálculo. Para ello, debemos interpretar a la frecuencia acumulada de una clase, como la cantidad de elementos que tienen menor valor que el límite superior de esa clase. Por ejemplo: ¿qué porcentaje de alumnos son menores de 18 años (216 meses)? (Trazamos una paralela al eje vertical a partir del punto que representa 216, hasta que corte a la ojiva, y proyectamos este punto horizontalmente hasta el eje vertical, estimando que el 78 % son menores de 216 meses).

Con un procedimiento semejante calcular los cuartiles. Tener en cuenta que este es un procedimeinto gráfico, y por tanto aproximado.



$Q_1 = 205$

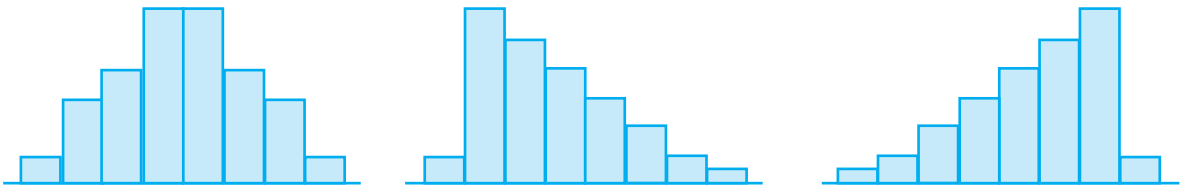
$Q_2 = Med = 211$

$Q_3 = 216$

Formas de los histogramas

Los histogramas son herramientas valiosas, ya que a través de ellos, se podrán interpretar varios hechos importantes de la variable implicada. Pero, para ello, el histograma de la muestra debe tener una forma de distribución bastante semejante a la de la población de donde se extrajo la muestra.

Al igual que los gráficos de puntos, los histogramas pueden presentar distintas formas; los más comunes son los siguientes:



Simétrico, normal o triangular | Sesgado a la derecha | Sesgado a la izquierda

Curvas de frecuencias

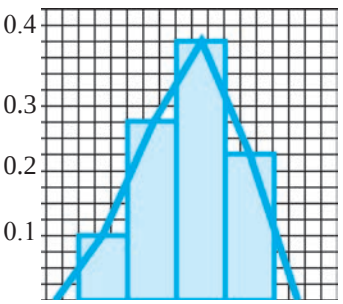
Recordemos que el trazo del histograma y su polígono de frecuencias se realiza sobre ejes cartesianos: en el eje de las abscisas se localizan los límites de los intervalos o las marcas de clase y en el eje de las ordenadas se localizan las frecuencias absolutas o las relativas. Para el concepto de curva de frecuencia, necesitamos que en el eje de las ordenadas se localicen las frecuencias relativas.

El siguiente proceso nos ayudará a comprender el concepto de curva de frecuencia.
Se ha anotado el peso en kilogramos de 40 estudiantes de una escuela.

41	46	46	46	51	51	52	54	54
57	58	58	58	59	60	61	61	61
64	64	65	65	66	67	67	67	68
68	68	69	69	72	72	73	74	75
75	78	78	80					

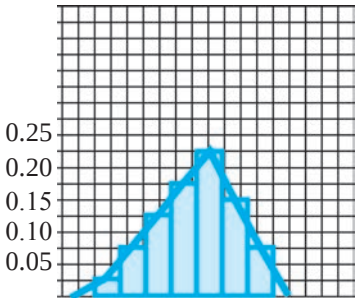
- Si agrupamos los datos de 10 en 10 kg obtenemos la siguiente distribución de frecuencias con su histograma y polígono de frecuencias asociado.

Intervalos	f	fr
[41,51)	4	0.100
[51,61)	11	0.275
[61,71)	16	0.400
[71,81)	9	0.225
Total	40	1.000

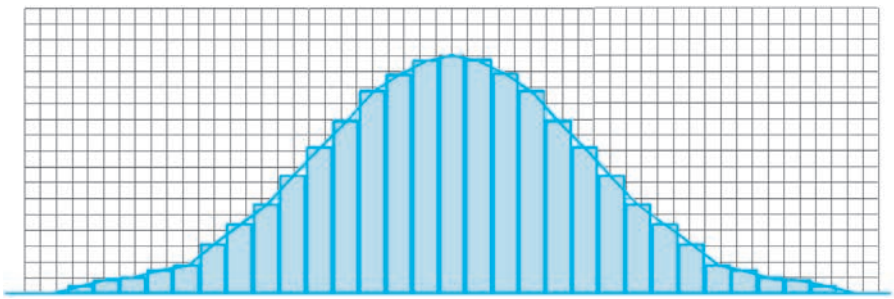


■ Si agrupamos los datos de 5 en 5 kg obtenemos la siguiente distribución de frecuencias con su histograma y polígono de frecuencias asociado.

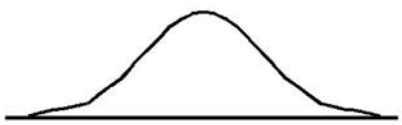
Intervalos	f	fr
[41,46)	1	0.025
[46,51)	3	0.075
[51,56)	5	0.125
[56,61)	6	0.150
[61,66)	7	0.175
[66,71)	9	0.225
[71,76)	6	0.150
[76,81)	3	0.075
Total	40	1.000



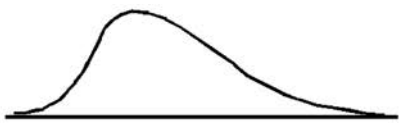
■ Si se hace el mismo estudio para todos los individuos de un país y agrupamos los datos aproximados por los decigramos y centigramos obtenemos un polígono de frecuencias que se aproxima a una curva que se llama **curva de frecuencias**.



Como ya lo hemos mencionado, las curvas de frecuencias más comunes son la simétrica (normal) y las sesgadas (a la derecha o a la izquierda)



Simétrica o normal



Sesgada a la derecha

Sesgada a la izquierda

Ejercicio 3.5

1. Los datos que se muestran a continuación, corresponden a la edad de mortalidad de una población.

   1, 8, 16, 21, 25, 27, 27, 28, 29, 31, 31, 33, 34, 35, 38, 38, 40, 42, 42, 45, 45, 48, 48, 48, 49, 50, 51, 52, 55, 55, 56, 58, 58, 58, 59, 59, 59, 60, 61, 61, 62, 62, 63, 64, 67, 67, 68, 68, 68, 69, 69, 70, 70, 71, 72, 72, 73, 74, 74, 74, 74, 75, 75, 77, 78, 78, 78, 79, 79, 79, 79, 80, 80, 80, 81, 81, 84, 84, 85, 86, 86, 86, 87, 83, 83, 82, 85, 85, 82, 84, 88, 88, 89, 89, 90, 93, 103

   a) Construye un gráfico de tallo y hoja apropiado.
   b) Construye una distribución de fecuencias apropiada.
   c) Construye el histograma
   d) Construye el polígono de frecuencias.
   e) Construye una distribuciónd efrecuencias acumuladas y la ojiva
   f) A partir de la ojiva determina los cuartiles.
   g) Describe la distribución.

2. Los datos que se muestran a continuación, corresponden a la edad en que contrajeron matrimonio un grupo de mujeres.

   30, 27, 56, 40, 30, 26, 31, 24, 23, 35, 29, 33, 29, 22, 33, 29, 46, 25, 34, 19, 23, 23, 44, 29, 30, 25, 23, 60, 25, 27, 37, 24, 22, 27, 31, 24, 26

   a) Construye un gráfico de tallo y hoja apropiado.
   b) Construye una distribución de fecuencias apropiada.
   c) Construye el histograma
   d) Construye el polígono de frecuencias.
   e) Construye una distribuciónd efrecuencias acumuladas y la ojiva
   f) A partir de la ojiva determina los cuartiles.
   g) Describe la distribución.

3. A continuación se presentan los pesos en libras de 32 dorados pescados en un torneo.

   52.60, 51.60, 50.50, 50.50, 50.20, 49.55, 49.50, 49.05, 48.30, 48.25, 48.15, 47.65, 46.85, 46.25, 46.20, 45.95, 43.95, 43.85, 43.75, 43.55, 41.15, 41.10, 39.75, 39.40, 39.30, 38.10, 37.25, 36.55

   a) Construye una distribución de fecuencias apropiada.
   c) Construye el histograma
   d) Construye el polígono de frecuencias.
   e) Describe la distribución.

4. A continuación se indican el número de resfriados experimentados durante un invierno por un grupo de 30 niños.

   7, 1, 1, 0, 3, 4, 5, 5, 3, 2, 3, 3, 6, 6, 2, 4, 2, 1, 0, 0, 3, 4, 5, 6, 3, 1, 4, 1, 3, 4

   Organiza los datos en una distribución de frecuencias y represéntalos gráficamente. Debes decidir si es suficiente una distribución de frecuencia simples o si debes agruparla. Describe la distribución.

Lección 3.6

Antecedente 6 para la exploración de datos cuantitativos: cálculo de medidas de resumen para distribuciones de frecuencias simples y agrupadas.

Objetivo: Aprender a calcular medidas de resumen de distribuciones de frecuencias simples agrupadas.

Actividad 13

Qué hacer

Consulta las **páginas 121 a 134** y al finalizar tu estudio contesta:

1) Se ha preguntado a un grupo de 70 alumnos universitarios sobre el número de zapatos que calzan, obteniendo los resultados de la siguiente tabla:

Nº de calzado	Nº de alumnos
25	4
26	15
27	17
28	20
29	10
30	4

a) ¿Cuál es el número de calzado más frecuente?
b) ¿Cuál es el número medio de calzado?
c) ¿Cuál es el número mediano de calzado?
d) ¿Cuál es la varianza y desviación estándar?

2) El consumo de gasolina, en litros, de una flota de camiones a lo largo de un día está tabulado en la siguiente tabla:

Consumo	Nº de camiones
[0,11)	8
[11,21)	12
[21,31)	10
[31,41)	14
[41,51)	21
[51,61)	16
[61,71)	9

a) ¿Entre qué valores está la cantidad más frecuente de combustible?
b) ¿Cuál es el consumo medio de combustible?
c) ¿Cuál es el consumo mediano de combustible?
d) ¿Cuál es la varianza y desviación estándar?

En las páginas anteriores, aprendiste el significado y cálculo de los principales resúmenes estadísticos. El buen dominio de estos conceptos y el uso de tecnología, te garantizan un buen desempeño en tareas de exploración de datos. La lección actual, tenía mucha importancia en épocas anteriores, pero en esta era tecnológica pierde relevancia, debido a que, para calcular las medidas de resumen, es mucho más recomendable trabajar con los datos originales que sobre datos organizados en tablas. Sin embargo, estos temas aún no pasan de moda y por lo tanto debes dominarlos.

a) Media

Cálculo de la media en distribuciones de frecuencias simples (datos no agrupados)

Una tabla como la siguiente, presenta una distribución de frecuencias no agrupadas, conocida también como distribución de frecuencias simples.

x	*f*
1	3
2	2
3	2
4	3
5	8
6	6
7	4
8	3
9	3
10	6

Estas son las calificaciones de un examen. Como puede verse, tres alumnos han sacado 1, dos alumnos han obtenido 2, ...

Con la fórmula $\bar{X} = \frac{\Sigma x}{n}$, calcularíamos la media así:

$$\bar{X} = \frac{(1+1+1)+(2+2)+(3+3)+(4+4+4)+...}{3+2+2+3+...}$$

El resultado de cada paréntesis del numerador es igual al producto de cada calificación por su frecuencia. El denominador es la suma de las frecuencias.

Sería más cómodo calcular previamente cuánto vale cada paréntesis. Por eso, en la práctica, para calcular la media, ampliamos la tabla con **una columna en la que escribimos el producto de cada *x* (calificación) por su frecuencia *f* (frecuencia correspondiente).**

x	*f*	*fx*
1	3	3
2	2	4
3	2	6
4	3	12
5	8	40
6	6	36
7	4	28
8	3	24
9	3	27
10	6	60
	40	**240**

De la tabla:

$\Sigma f = 40$ = *Número total de alumnnos*

$\Sigma fx = 240$

Entonces, la media será:

$$\bar{X} = \frac{240}{40} = 6$$

En general, para **una distribución de frecuencias simples (no agrupadas)** la media es:

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{\Sigma fx}{n}$$

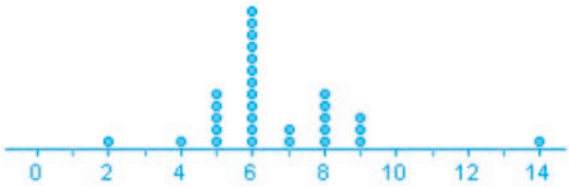
En donde:

$$\Sigma fx = fx_1 + f_2x_2 + \ldots + f_nx_n$$

$$\Sigma f = f_1 + f_2 + \ldots + f_n = n$$

Ejemplo

A partir del siguiente gráfico de puntos, construiremos una distribución de frecuencias simples.



Número de integrantes en las familias de una muestra de estudiantes

En la primer columna escribimos todos los valores distintos que puede tomar la variable y en la segunda la frecuencia con que aparece cada valor.

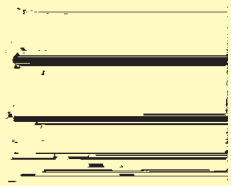
Valores distintos (modalidades) ***x***	**Frecuencia** ***f***
2	1
3	0
4	1
5	5
6	12
7	2
8	5
9	3
10	0
11	0
12	0
13	0
14	1

Distribución de frecuencias simples para la variable número de integrantes en las familias.

A la tabla agregamos una tercer columna en donde escribimos **los productos fx**.

Ejemplo

	Valores distintos (modalidades) x	Frecuencia f	$f \bullet x$	Cálculo de la media
	2	1	2	
	3	0	0	
	4	1	4	
	5	5	25	
	6	12	72	
	7	2	14	
	8	5	40	
	9	3	27	
	10	0	0	
	11	0	0	
	12	0	0	
	13	0	0	
	14	1	14	
Sumas		$\Sigma f = 30$	$\Sigma f \bullet x = 198$	



Cálculo de la media en distribuciones de frecuencias agrupadas en intervalos

En este caso, sigue siendo válida la fórmula anterior, $\overline{X} = \frac{\Sigma f x}{\Sigma f}$;

sin embargo, se requiere especificar que valores debe tomar *x*, puesto que un intervalo está en lugar de un grupo de valores posibles. La solución consiste en utilizar como valores de *x*, la marca de clase del intervalo.

Ejemplo

En la página 110 se presenta la tabla construida para la variable edad de 30 estudiantes. A continuación se repite dicha tabla.

Intervalos	Conteo	Frecuencia f
[190,196)	//	2
[196,202)	///	3
[202,208)	𝍸 /	6
[208,214)	𝍸 ///	8
[214,220)	𝍸 ////	9
[220,226)	/	1
[226,232)	/	1
Total		30

Recuerda que esta tabla indica, por ejemplo, que hay dos alumnos con edades mayores o iguales a 190 y menores que 196, pero sin especificar sus edades exactas. En estos casos, es cuando se utiliza la **marca de clase** de cada intervalo como el valor de ***x*** representando al intervalo.

Ejemplo (Cont.)

Por lo tanto, para calcular la media en una distribución de frecuencias agrupadas, además de la columnna para ***fx,*** ocuparemos una columna adicional que contenga las marcas de clase.

Intervalos	Marca de clase	Frecuencia f	$f\,x$	Cálculo de la media
[190,196)	193	2	386	$\overline{X} = \frac{\Sigma f\,x}{\Sigma f}$
[196,202)	199	3	597	
[202,208)	205	6	1230	$\overline{X} = \frac{6306}{30} = 210.2$
[208,214)	211	8	1688	
[214,220)	217	9	1953	
[220,226)	223	1	223	
[226,232)	229	1	229	
Sumas		$\Sigma f = 30$	$\Sigma fx = 6306$	

b) Mediana

Cálculo de la mediana para una distribución de frecuencias simples (no agrupada)

En la lección (3.4), estudiamos el concepto de frecuencia acumulada y su representación gráfica llamada ojiva, y, en las páginas 114 y 115, utilizamos la ojiva para calcular las medidas de posición, y por ende, el valor de la mediana. Debido a que para una distribución de frecuencias simples no existe una fórmula única que nos proporcione el valor de la mediana, cuando se presenten estos casos, seguiremos utilizando la ojiva para determinar la mediana. Vuelve a repasar las páginas 112 a 114 y resuelve la siguiente actividad.

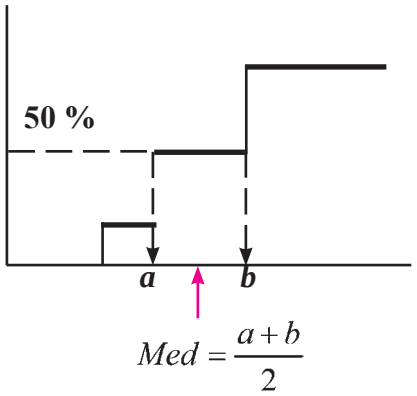
Actividad 3.6 a

Calcula la mediana de la siguiente distribución.

Salario por hora	Número de obreros
10.0	4
15.0	9
20.0	25
25.0	13
30.0	3
35.0	1

Caso especial

Si la línea a 50 % cae justamente en una zona plana de la ojiva, se calcula la semisuma de los dos valores extremos de dicha zona.



Ejemplo Determina la mediana de la siguiente distribución

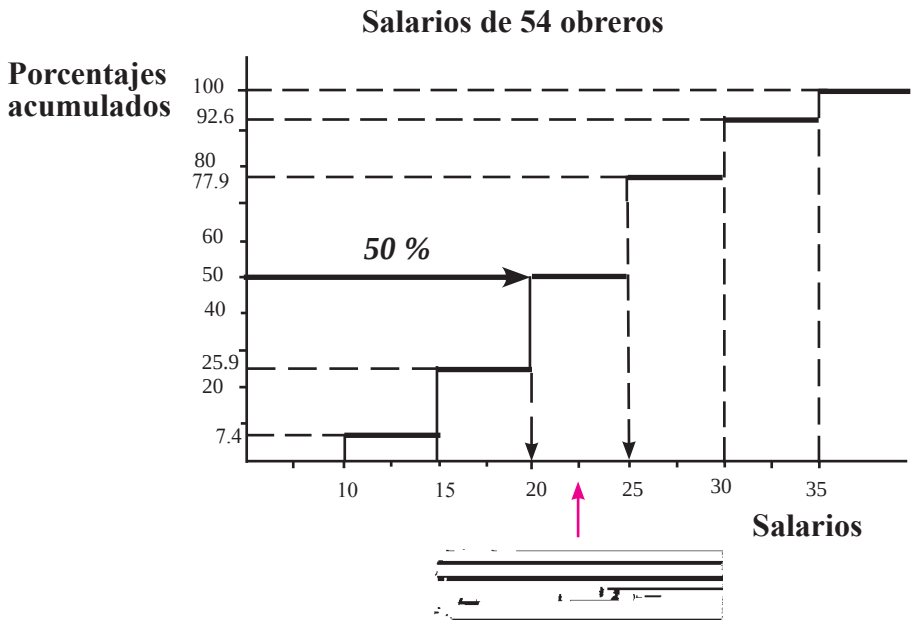
Salario por hora	Número de obreros
10.0	4
15.0	10
20.0	13
25.0	15
30.0	8
35.0	4

Solución

1º Se agregan a la tabla dos columnas, una con las frecuencias acumuladas absolutas y la otra con las frecuencias acumuludas relativas en porcentajes.

x	f	*f acum.*	*f acum. (porcentaje)*
10.0	4	4	7.4 %
15.0	10	14	25.9 %
20.0	13	27	50.0 %
25.0	15	42	77.8 %
30.0	8	50	92.6 %
35.0	4	54	100.0 %

2º Dibujar la ojiva (recuerda que en este ejemplo, se trata de una distribución de frecuencias simples, y la ojiva tiene forma de « escalera »).



Cálculo de la mediana para una distribución de frecuencias agrupadas.

En este caso, la mediana puede obtenerse directamente del histograma o bien mediante la aplicación de una fórmula.

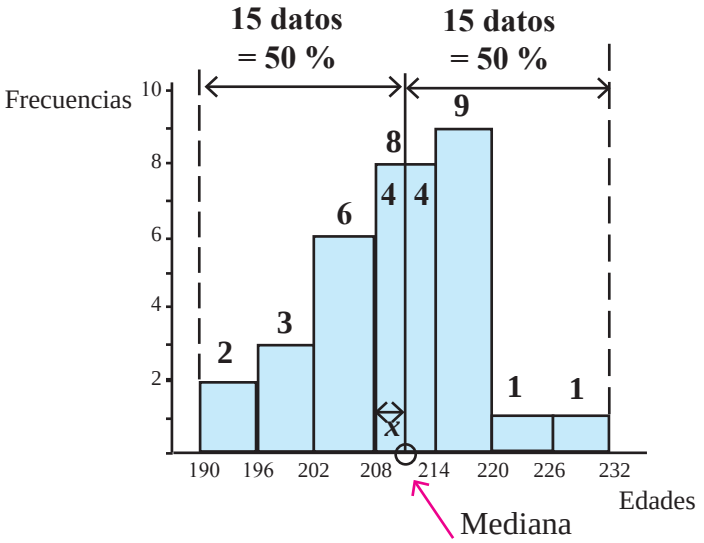
Ejemplo Calcularemos la mediana de la distribución de frecuencias agrupadas de la variable edades de la página 110.

Intervalos	Frecuencia f
[190,196)	2
[196,202)	3
[202,208)	6
[208,214)	8
[214,220)	9
[220,226)	1
[226,232)	1
Total	30

Procedimiento 1

Cálculo gráfico de la mediana utilizando el histograma

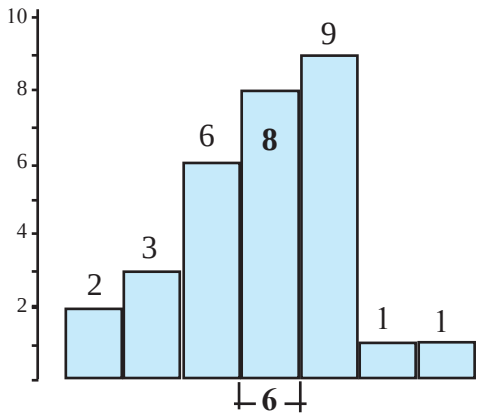
El histograma correspondiente para esta distribución es:

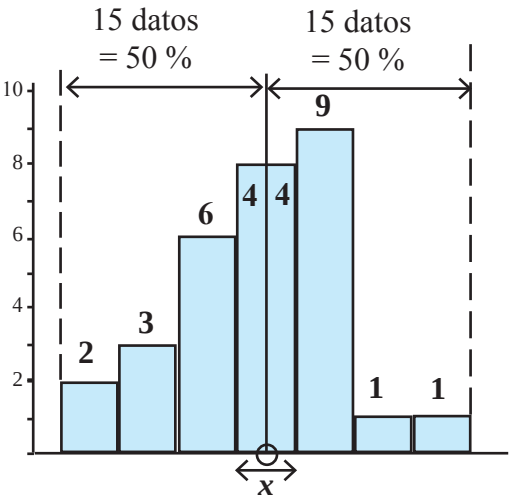


Mediana = 208 + *x*

Recordemos que la mediana es el punto que está a la mitad con el 50% de valores a su derecha y el 50% de valores a su izquierda (en este histograma 15 valores a cada lado).

La mediana "cae" en el intervalo **208 - 214**. Este intervalo se llama "**clase mediana**". El valor exacto de la mediana se obtiene por interpolación. Para ello, debemos observar, que **4** valores de la clase mediana están a la izquierda de la mediana y **4** a su derecha. Considerando que la longitud del intervalo es 6, planteamos la siguiente proporción.





$$\frac{8}{6} = \frac{4}{x} \longrightarrow 8x = (6)(4)$$

$$x = \frac{(6)(4)}{8} = 3$$

Por lo tanto, el valor de la mediana es: *Mediana* = 208 + *x*

= 208 + = 211

Procedimiento 2

En términos generales, podemos deducir la siguiente fórmula de la mediana para una distribución de frecuencias agrupadas:

$$Me = L_I + \left[\frac{\frac{n}{2} - f_{aA}}{f_c}\right] i$$

en donde:

L_I = Límite inferior de la clase mediana (es decir, la clase que contiene la mediana).

n = Número total de datos (suma de frecuencias).

f_{aA} = La frecuencia acumulada de la clase que precede ("antes") a la clase mediana.

f_c = Frecuencia de la clase mediana.

i = Tamaño del intervalo de la clase mediana.

Ejemplo

Aplicando la fórmula, calcularemos la mediana de la distribución de frecuencias agrupadas de la variable edades.

Intervalos	Frecuencia f
[190,196)	2
[196,202)	3
[202,208)	6
[208,214)	8
[214,220)	9
[220,226)	1
[226,232)	1
Total	30

Solución

1º Agregamos una columna de frecuencias acumuladas y determinamos la clase que contiene la mediana (clase mediana). **La clase mediana es la primera cuya frecuencia acumulada iguala o excede la mitad del total de datos.**

Intervalos	Frecuencia absoluta	Frecuencia acumulada
[190,196)	2	2
[196,202)	3	5
[202,208)	6	11
[208,214)	8	19
[214,220)	9	28
[220,226)	1	29
[226,232)	1	30
Total	30	

$\frac{n}{2} = 15$

Clase mediana: aquella cuya frecuencia acumulada iguala o excede a $\frac{n}{2} = 15$

2º Identificamos los elementos de la clase mediana:

Clase mediana: 208 - 214

Límites de la clase mediana: 208 y 214

Tamaño del intervalo que contiene la clase mediana:

$$\begin{array}{r} 214 \\ -\ 208 \\ \hline 6 \end{array}$$

$L_1 = 208$

***i* = 6**

De la tabla (columna de *f)* obtenemos $f_c = 8$.

De la tabla (columna de *frec. acum.)* obtenemos $f_{aA} = 11$.

Sustituyendo:

$$Me = L_I + \left[\frac{\frac{n}{2} - f_{aA}}{f_c}\right] i = 208 + \left[\frac{\frac{30}{2} - 11}{8}\right](6)$$

$$= 208 + \left(\frac{4}{8}\right)(6)$$

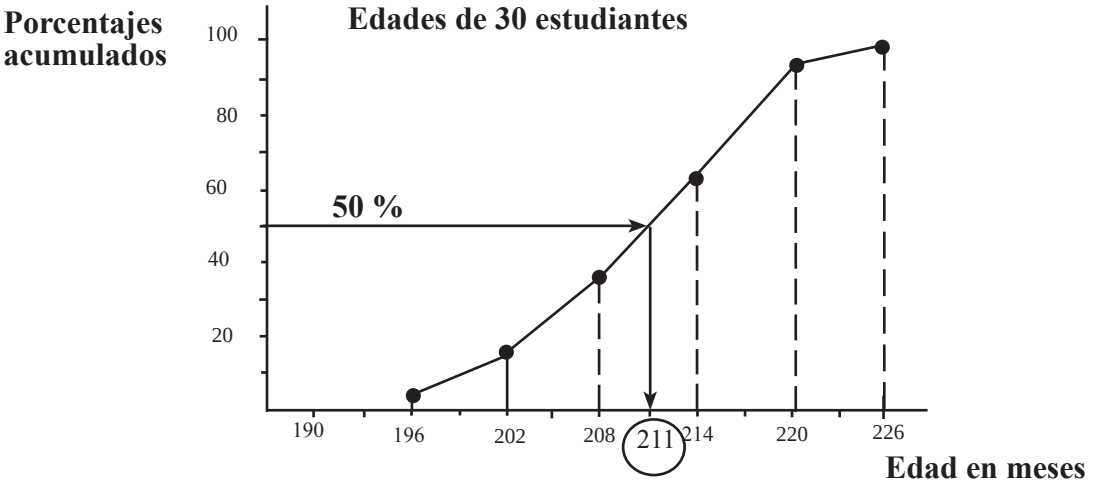
$$= 208 + 3$$

$$= 211$$

La edad mediana de los estudiantes es 211 meses.

Procedimiento 3

Un tercer método para calcular la mediana, es el ya visto de la ojiva. En la página 116, determinamos la mediana de esta distribución de edades.



$Q_2 = Med = 211$

La moda

Cálculo de la moda para una distribución de frecuencias simples (no agrupada).

En este caso, no hay nada que calcular: la moda es simplemente el valor de la variable cuya frecuencia es la más alta.

La moda de la siguiente distribución es 6.

x	**f**
2	**1**
4	1
5	5
6	12
7	2
8	5
9	3
12	1

← **Moda**: valor con la mayor frecuencia: **6**

Para una distribución de frecuencias agrupadas, la moda es el punto medio del intervalo que contiene mayor frecuencia. Este intervalo se llama **clase modal**.

Ejemplo Determinar la moda de la siguiente distribución.

Intervalos	Frecuencia f
[190,196)	2
[196,202)	3
[202,208)	6
[208,214)	8
[214,220)	9
[220,226)	1
[226,232)	1
Total	30

Clase modal: [214,220)

$$\textit{Moda} = \textit{marca de clase de la clase modal} = \frac{214+220}{2} = 217$$

Varianza y desviación estándar para datos organizados en distribuciones de frecuencias

Para distribuciones de **frecuencias simples**, la fórmula de la varianza muestral es:

$$s^2 = \frac{\sum f\left(x-\bar{X}\right)^2}{n-1}$$

Y la desviación estándar es:

$$s = \sqrt{\frac{\sum f\left(x-\bar{X}\right)^2}{n-1}}$$

Para distribuciones de **frecuencias agrupadas**, la fórmula es la misma, pero ***x*** representa la **marca de clase** del intervalo.

Ahora bien, así como transformamos la fórmula $s^2 = \frac{\sum f\left(x-\bar{X}\right)^2}{n-1}$ en la fórmula simplificada: $s^2 = \frac{\sum x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$ para datos individuales, de manera semejante obtenemos **la fórmula de la varianza simplificada para datos organizados en distribuciones de frecuencias**:

$$s^2 = \frac{\Sigma f \bullet x^2 - \frac{(\Sigma f \bullet x)^2}{n}}{n-1}$$

Y para la desviación estándar:

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

Para el cálculo de cada uno de los términos de estas fórmulas, formaremos una tabla con las siguientes columnas:

Para distribuciones de frecuencias simples (datos no agrupados):

x	f	x^2	$f \cdot x$	$f \cdot x^2$
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
Sumas	$n = \Sigma f$		Σfx	Σfx^2

$$\overline{X} = \frac{\Sigma fx}{n}$$

Para distribuciones de **frecuencias agrupadas**:

Intervalos	**Marca de clase (x)**	f	x^2	$f \cdot x$	$f \cdot x^2$
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
Sumas		$n = \Sigma f$		Σfx	Σfx^2

$$\overline{X} = \frac{\Sigma fx}{n}$$

Ejemplo

En la página 94 calculamos la varianza y desviación estándar de los siguientes datos:

6, 10, 10, 10, 9, 7, 10, 9, 10, 7, 10.

Estos datos pueden organizarse en la siguiente distribución de frecuencias simples:

x	f
6	1
7	2
9	2
10	6

Para calcular la varianza y desviación estándar a partir de esta tabla, necesitamos las nuevas fórmulas. Para ello, a continuación ampliamos la tabla.

x	f	x^2	fx	fx^2
6	1	36	6	36
7	2	49	14	98
9	2	81	18	162
10	6	100	60	600
Sumas	$\Sigma f = 11$		$\Sigma fx^2 = 11$	$\Sigma fx^2 = 896$

Sustituyendo:

$$s^2 = \frac{\Sigma f \bullet x^2 - \frac{(\Sigma f \bullet x)^2}{n}}{n-1} = \frac{896 - \frac{(98)^2}{11}}{11-1}$$

$$= \frac{896 - 873.09}{10} = 2.291$$

$$s = \sqrt{2.291} = 1.51$$

Resultado idéntico al obtenido en la página 94.

Ejemplo

Calculemos la varianza y desviación estándar para la variable edades, agrupada en la distribución de frecuencias ya conocida.

Intervalos	**Marca de clase** (x)	**Frecuencia** f	x^2	fx	fx^2
[190,196)	193	2	37249	386	74498
[196,202)	199	3	39601	597	118803
[202,208)	205	6	42025	1230	252150
[208,214)	211	8	44521	1688	356168
[214,220)	217	9	47089	1953	423801
[220,226)	223	1	49729	223	49729
[226,232)	229	1	52441	229	52441
Sumas		$\Sigma f = 30$		$\Sigma fx = 6306$	$\Sigma fx^2 = 1327590$

$$\overline{X} = \frac{\Sigma fx}{\Sigma f}$$

$$\overline{X} = \frac{6306}{30} = 210.2$$

$$s^2 = \frac{\Sigma f \bullet x^2 - \frac{(\Sigma f \bullet x)^2}{n}}{n-1} = \frac{1327590 - \frac{(6306)^2}{30}}{30-1} = 71.34$$

$$s = 8.4$$

Ejercicio 3.6

1. Se ha realizado una prueba compuesta de 10 preguntas a 40 alumnos de un grupo, obteniéndose los siguientes resultados.

N° DE RESPUESTAS CORRECTAS	N° DE ALUMNOS
[0,2)	1
[2,4)	4
[4,6)	9
[6,8)	14
[8,10)	7
[10,12)	5

a) Representar gráficamente la distribución.
b) Calcular todas las medidas estadísticas que haz estudiado hasta el momento.

2. Trabaja con los datos de tu proyecto sobre el alumno(a) típico(a). Presenta un reporte.

3. Los siguientes datos muestran las calificaciones obtenidas por un grupo de 25 estudiantes.

3, 3, 4, 5, 5, 5, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10.

Con estos datos, realiza lo que se indica:

a) Construye una distribución de frecuencias simples.
b) Calcula cada una de las medidas de tendencia central.
c) Calcula cada una de las medidas de dispersión.
d) Construye una ojiva y a partir de ella, calcula cada una de las medidas de posición.

Lección 3.7 Exploración datos cuantitativos

Objetivo: Compara distribuciones y hace inferencias informales, usando las formas de las distribuciones y medidas de tendencias central y de dispersión.

Actividad 14

Qué hacer

Consulta las **páginas 136 a 140** y al finalizar tu estudio contesta:

1) El gerente de un equipo de beisbol de ligas mayores está evaluando a un par de lanzadores de ligas menores. El número de ponches que cada lanzador ha logrado en sus últimos 6 juegos están registrados abajo:

Juego	Luis	José
1	4	8
2	3	9
3	9	12
4	16	6
5	10	9
6	12	10

Si la consistencia es de interés primordial, ¿qué lanzador deberá ser más interesante para el gerente?

A través de las páginas anteriores, has aprendido los elementos básicos para explorar datos cuantitativo. A continuación, podrás integrar todas estas herramientas y lograr una visión global de gran parte del proceso de exploraación de datos.

a) Comparación de grupos

Fase 1. Planteamiento del problema

La experiencia anecdótica nos señala que en el futbol los equipos tienen más posibidades de ganar cuando juegan de local que de visitante. ¿En qué medida es cierto ésto? ¿Realmente jugar de local beneficia a los equipos de futbol?

Fase 2. Recolección de datos

Para contestar esta pregunta utilizaremos los puntajes obtenidos por los equipos mexicanos de primera división en el torneo de clausura 2008.

Estadísticas de la primera división.
Torneo de clausura 2008.

Equipo	Puntos como local	Puntos como visitante
Pachuca	25	11
Toluca	23	13
UNAM	15	11
Monterrey	17	09
Puebla	09	17
UAG	16	10
Cd. Juárez	13	11
América	07	16
Santos	16	06
Morelia	16	06
Guadalajara	15	06
Chiapas	11	09
Atlas	15	06
Atlante	11	06
San Luis	09	08
UANL	09	05
Necaxa	12	02
Cruz Azul	06	07

Fuente: *El Debate*

Fase 3. Exploración de datos

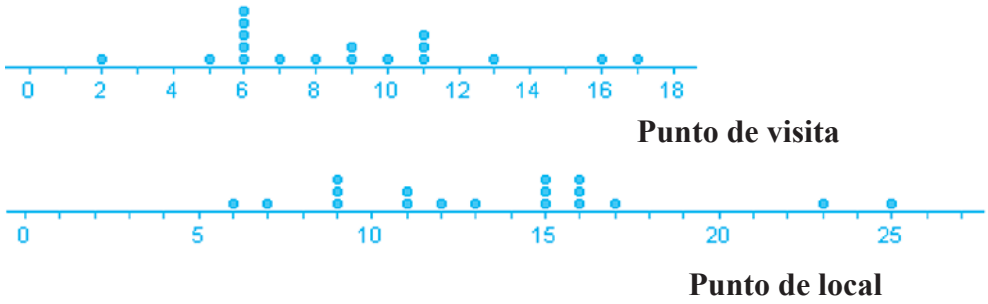
3.1 Análisis de datos

-Organización y representación gráfica

-Cálculo de medidas de resumen.

Para realizar esta fase de análisis podemos seguir varios caminos. Pero es recomendable empezar trazando un gráfico de puntos.

Ejemplo



A continuación nos auxiliaremos con la hoja de calculo excel para determinar las medidas estadísticas.

Primer paso. Abrir el programa excel y capturar los dos grupos de datos.

	A	B	C	D	E
1					
2					
3					
4					
5					
6		Puntos local		Puntos visita	
7		25		11	
8		23		13	
9		17		9	
10		16		6	
11		16		10	
12		16		6	
13		15		6	
14		15		6	
15		15		11	
16		13		11	
17		12		2	
18		11		9	
19		11		6	
20		9		8	
21		9		17	
22		9		5	
23		6		7	
24		7		16	

Ejemplo (Cont.)

Seguno paso. Excel trae funciones para determinar cada una de las medidas estadísticas. Estas funciones y su uso se explican con el ejemplo que estamos estudiando.

Para juegos como local

Media. Su función se llama PROMEDIO. Para calcularlo procede como se indica:

Para la media de puntos, digita:

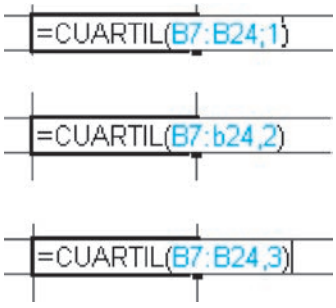
Clic y aparece el valor de la media: $\overline{X} = 13.6$

Desviación estándar. Para la desviación estándar muestral, la función se llama DESVEST.

Clic y aparece el valor de la desviación estándar: $s = 5.01$

Cuartiles. Para los cuartiles la función se llama CUARTIL La instrucción:

permite obtener el cuartil ***n***. El número ***n*** toma los valores 0, 1, 2 ,3, 4: el cero para obtener el valor mínimo, el uno para obtener el primer cuartil, el dos para obtener la mediana, el tres para obtener el tercer cuartil y el cuatro para obtener el valor máximo.



$Q_1 = 9$

$Med = Q_2 = 14$

$Q_3 = 16$

Ejemplo

Siguiendo las instrucciones anteriores, verifica las siguiente medidas para el grupo de puntos de visita:

Medidas estadísticas para el grupo de puntos de visita:

$\overline{X} = 8.8$ $\quad Q_2 = 8.5$

$s = 3.88$ $\quad Q_3 = 11$

$Q_1 = 6$

Observando los valores medios de puntos, una vez más se corrobora que de local es más beneficioso.

$$\overline{X}_{local} = 13.6$$

$$\overline{X}_{visita} = 8.8$$

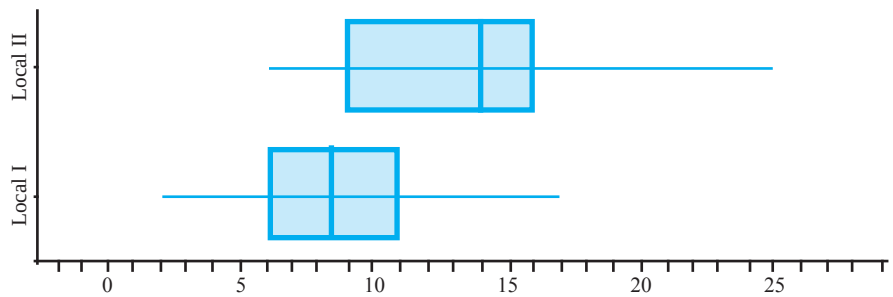
Sin embargo, la desviación estándar de local es mayor que la de visita.

$$s_{local} = 5.01$$

$$s_{visita} = 3.88$$

Por lo tanto, hay más variabilidad jugando de local.

Un gráfico de caja nos muestra mejor la variabilidad:



Se aprecia mayor variabilidad de local que de visita. El rango de visita es $\boldsymbol{R} = 17 - 2 = 15$, y de local $\boldsymbol{R} = 25 - 6 = 19$. El tamaño de la caja de local es mayor que de visita. Esto significa que el rango intercuartílico es mayor de local que de visita:

$$RIC_{local} = -Q_3 - Q_1 - 16 - 7$$

$$RIC_{local} = -Q_3 - Q_1 = 11 \quad 6 = 5$$

Fase 4. Interpretación de resultados

Los resultados efectivamente corroboran la creencia de que jugar futbol de local es mejor que de visita. Sin embargo, existe variabilidad importante. Por ejemplo, un equipo ganó 7 puntos de local y 16 de visita. Con respecto a la muestra, parece ser representativa ya que corresponde a un torneo completo.

b) Exploración de una distribución

Fase 1. Planteamiento del problema

Muchas variables exhiben una distribución normal. Por ejemplo estaturas y pesos de personas, calificaciones de estudiantes. En este problema, investigaremos la distribución exhibida por un grupo de calificaciones obtenidas por estudiantes para contestar la pregunta: ¿La distribución de la variable calificaciones es de forma normal?

Fase 2. Recolección de datos

Para contestar esta pregunta utilizaremos los puntajes obtenidos por 80 aspirantes aceptados para ingresar en la carrera de ingeniería mecánica del Tecnológico de Culiacán para el ciclo 2009-2010.

Puntajes de alumnos seleccionados

1143	1108	1108	1107	1098	1097	1089	1081	1077	1077
1076	1074	1073	1069	1068	1067	1059	1056	1055	1053
1053	1052	1049	1047	1047	1046	1045	1044	1044	1044
1043	1043	1043	1041	1040	1034	1032	1028	1027	1019
1014	1014	1012	1009	1008	1007	1007	1006	1004	1003
1003	1002	1001	1000	1000	999	998	995	995	993
992	987	986	985	983	981	980	979	978	974
971	971	971	971	970	968	968	967	966	965

Fase 3. Exploración de datos

3.1 Análisis de datos

-Organización y representación gráfica

- Cálculo de medidas de resumen.

En este caso, dada la cantidad de datos, procede organizarlos en una distribución de frecuencias. El excel nos ayuda, pero necesita que le proporcionemos los límites superiores de los intervalos. Entonces procederemos a formar los intervalos.

$R = 1143 - 178 = 178.$

Número de intervalos: $k = \sqrt{80} = 8.8$

Utilizaremos 9 intervalos.

Amplitud de intervalos: $i = \frac{R}{k} = \frac{168}{9} = 19.7$

Utilizaremos $i = 20$

Empezaremos el primer intervalo con el dato menor:

[965,985)

[985,1005)

[1005,1025)

[1025,1045)

[1045,1065)

[1065,1085)

[1085,1105)

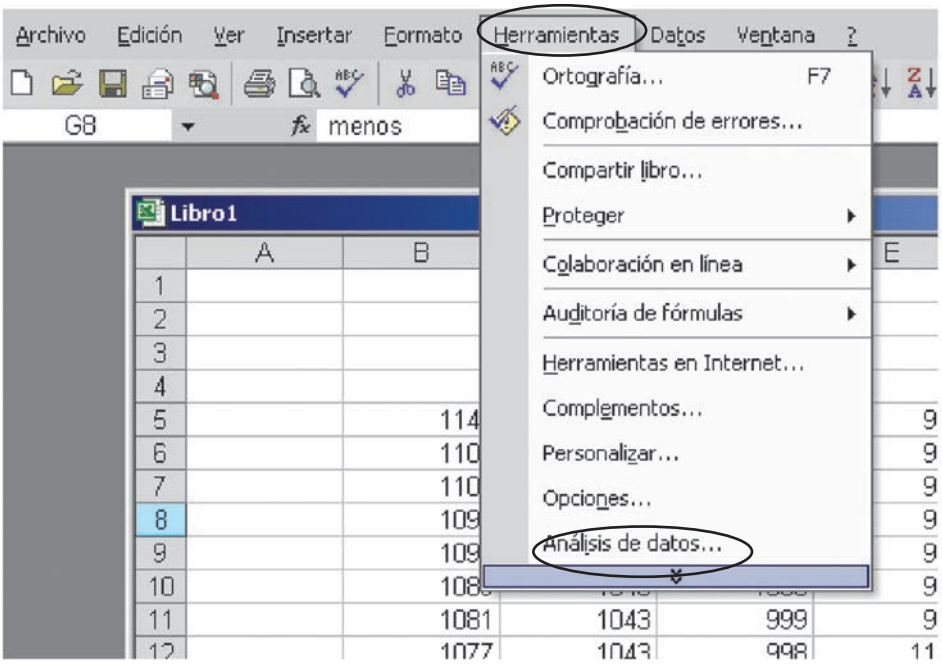
[1105,1125)

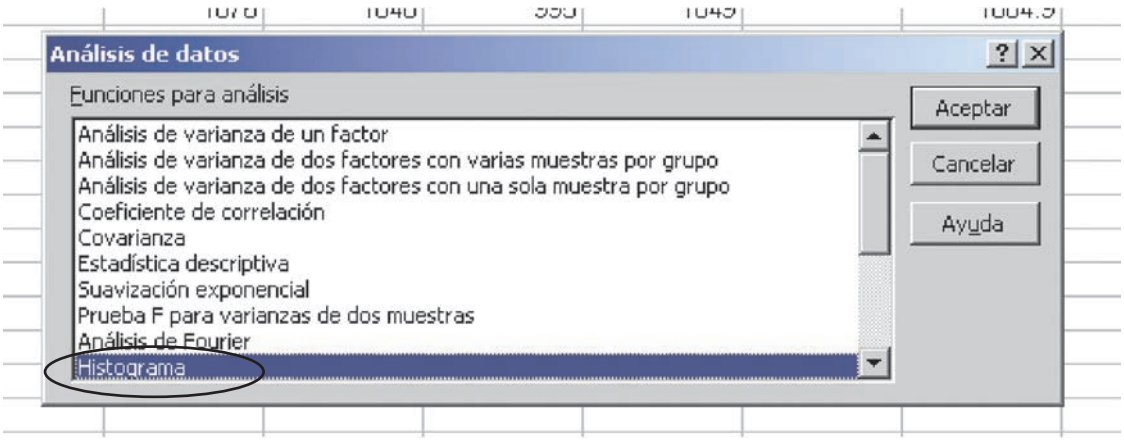
[1125,1145)

Una vez formados los intervalos abrimos el programa ***Excel*** y capturamos todos los datos, y, en un lugar aparte también capturamos los límites superiores, pero, ***¡atención!***: el programa requiere que ***cada límite superior sea capturado como se indica: Si los datos están redondeados a enteros, quitarle a cada límite superior un décimo; si están redondeados a décimos quitarles un centésimo, y así sucesivamente dependiendo del número de cifras decimales en los datos.*** Es decir, debemos capturar los siguientes límites superiores: 984.9, 1004.9, 1024.9, 1044.9, 1064.9, 1084.9, 1104.9, 1124.9, 1144.9.

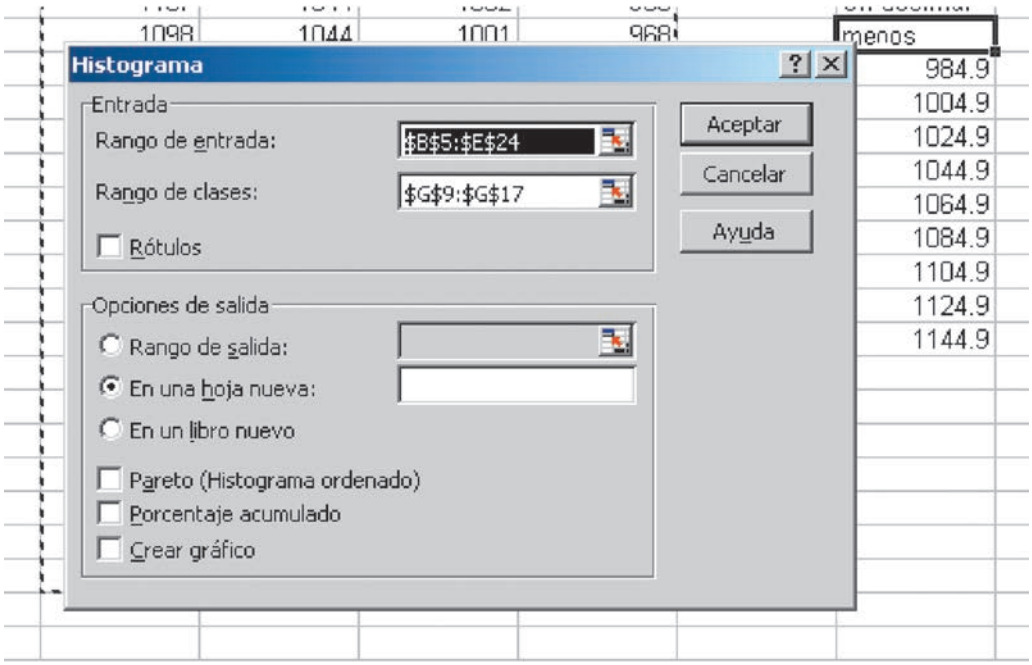
	A	B	C	D	E	F	G
1							
2							
3							
4							
5		1143	1046	1003	971		
6		1108	1045	1003	970		Límites Sup.
7		1107	1044	1002	968		Un decimal
8		1098	1044	1001	968		menos
9		1097	1044	1000	967		984.9
10		1089	1043	1000	966		1004.9
11		1081	1043	999	965		1024.9
12		1077	1043	998	1108		1044.9
13		1077	1041	995	1052		1064.9
14		1076	1040	995	1049		1084.9
15		1074	1034	993	1047		1104.9
16		1073	1032	992	1047		1124.9
17		1069	1028	987	1007		1144.9
18		1068	1027	986	1007		
19		1067	1019	985	1006		
20		1059	1014	983	1004		
21		1056	1014	981	974		
22		1055	1012	980	971		
23		1053	1009	979	971		
24		1053	1008	978	971		

A continuación, seleccionamos ***herramientas***, después ***análisis de datos*** y finalmente ***histogramas.***





Solución



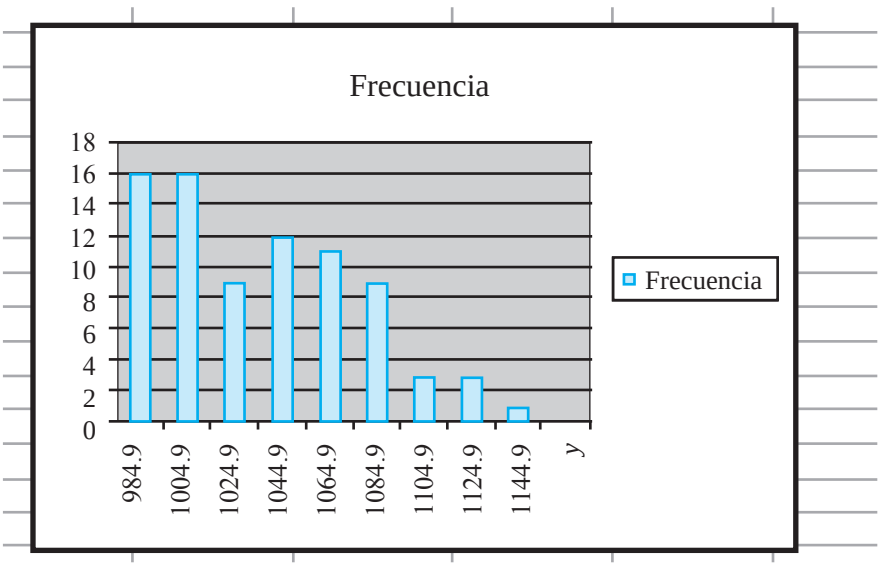
***Clic* en aceptar** y aparece la distribución de frecuencias:

	A	B
1	*Clase*	*Frecuencia*
2	984.9	16
3	1004.9	16
4	1024.9	9
5	1044.9	12
6	1064.9	11
7	1084.9	9
8	1104.9	3
9	1124.9	3
10	1144.9	1
11	y mayor...	0

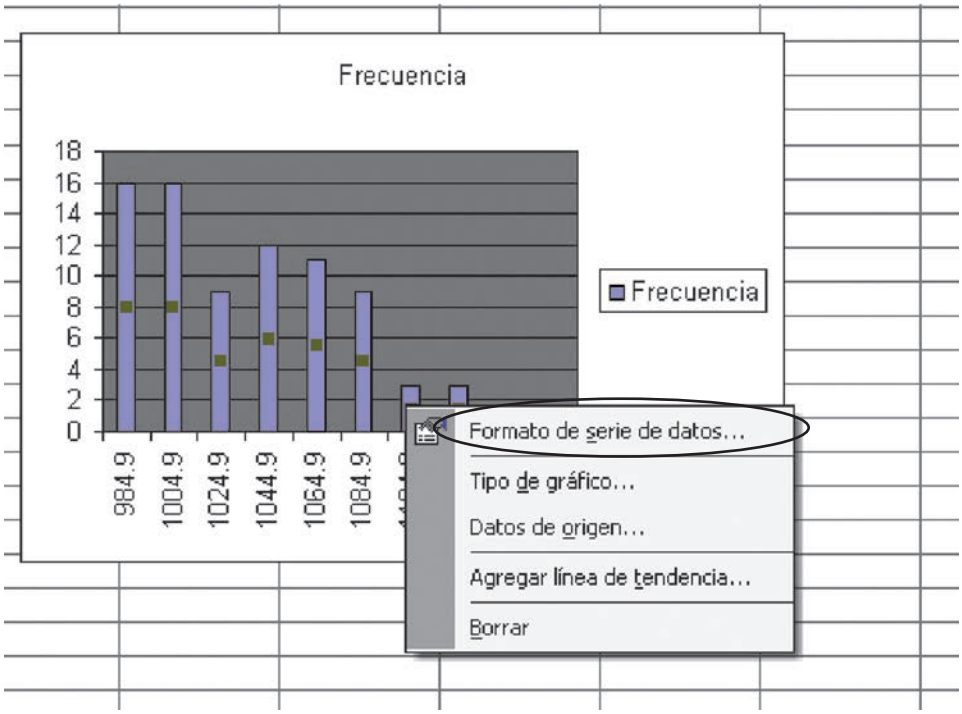
Ignorar este renglón.

A continuación **seleccionamos gráficas** y aparece lo siguiente:

Clic en la primera opción y a continuación repetidamente clic en ***siguiente.***

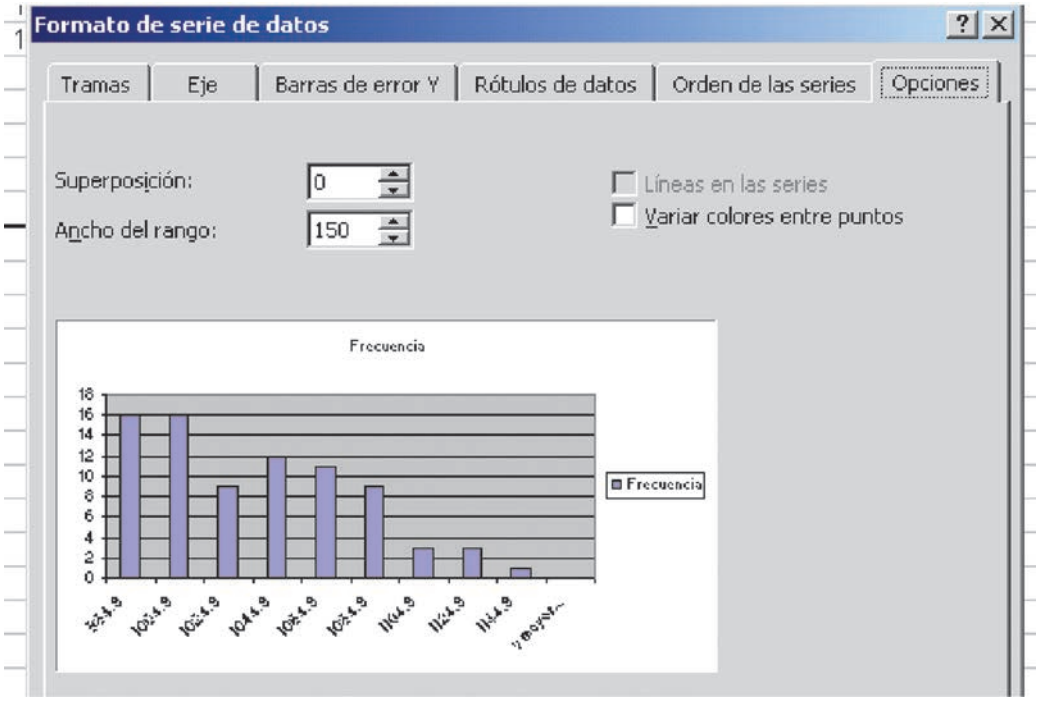


Excel no proporciona directamente el histograma sino un gráfico de barras. Para convertir este gráfico en histograma, hacer clic con botón derecho en cualquier barra del gráfico. Aparece lo siguiente:

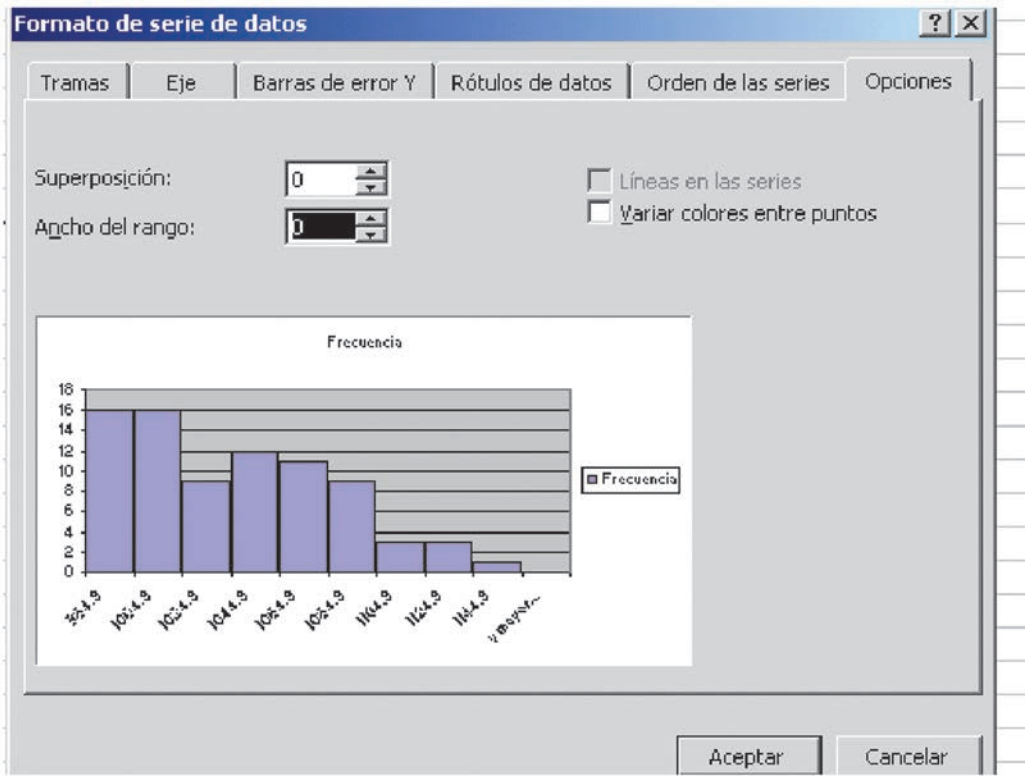


Seleccionar Formato de serie de datos.

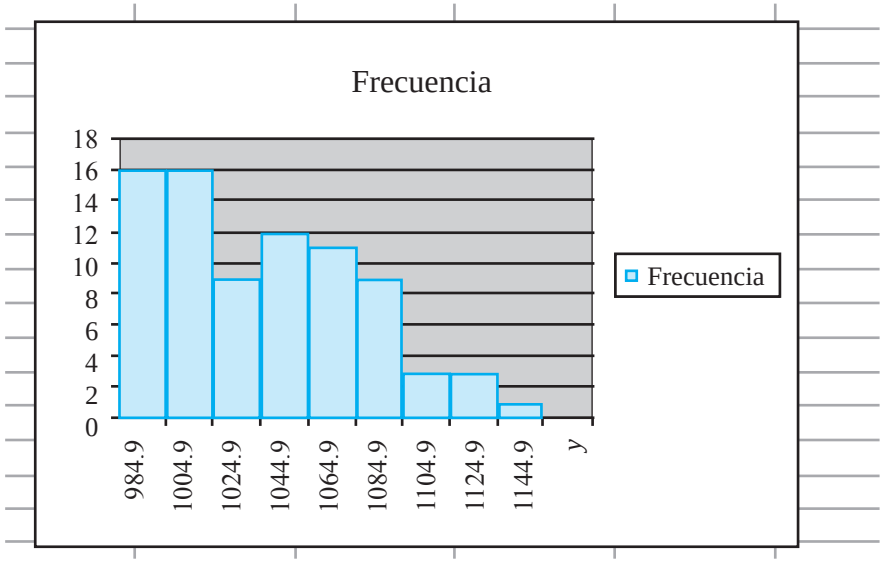
Clic en opciones:



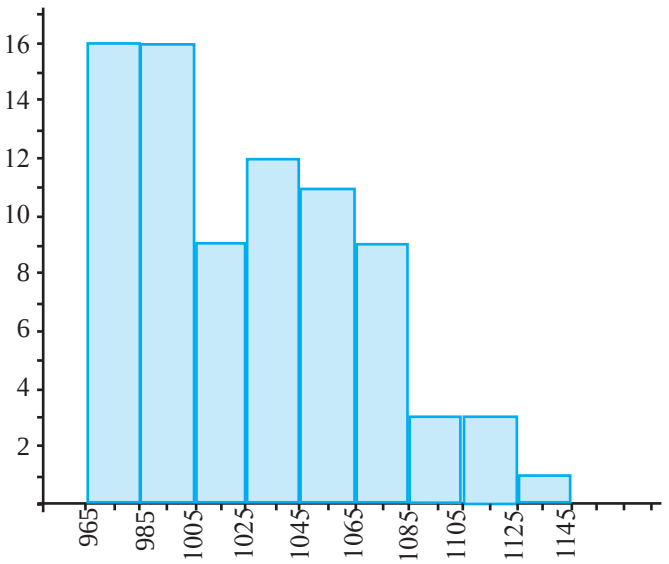
Convierte el ancho de rango en cero.



Clic en aceptar:



Reescribir los límites de intervalos:



Actividad 3.7a

Comprueba que las medidas de resumen son:

$\overline{X} = 1025.5$
$s = 42.2$
$Q_1 = 992.8$
$Med = Q_2 = 1016.5$
$Q_3 = 1053$

Interpretación de resultados

Definitivamente la forma de la distribución no es normal, más bien es segada hacia la derecha. El que la media sea mayor que la mediana confirma esta afirmación. Sin embargo, ésto tiene una posible explicación: La distribución corresponde a calificaciones de alumnos que aprobaron el examen. Es decir, faltarían datos que corresponden a alumnos con un puntaje menor que el mínimo requerido para aprobar. Entonces, pudiera pensarse que esta distribución es en forma aproximada la mitad de una normal.

Ejercicio 3.7

1. Realiza la exploración de los siguientes conjuntos de datos:

   a) El conjunto de datos adjunto está formado por observaciones del gasto de agua en regaderas (L/min) para una muestra de 130 casas:

   4.6, 12.3, 7.1, 7.0, 4.0, 9.2, 6.7, 6.9, 11.5, 5.1, 3.8, 11.2, 10.5, 14.3, 8.0, 8.8, 6.4, 5.1, 5.6, 9.6, 7.5, 7.5, 6.2, 5.8, 2.3, 3.4, 10.4, 9.8, 6.6, 3.7, 6.4, 6.0, 8.3, 6.5, 7.6, 9.3, 9.2, 7.3, 5.0, 6.3, 13.8, 6.2, 5.4, 4.8, 7.5, 6.0, 6.9, 10.8, 7.5, 6.6, 5.0, 3.3, 7.6, 3.9, 11.9, 2.2, 15.0, 7.2, 6.1, 15.3, 18.9, 7.2, 5.4, 5.5, 4.3, 9.0, 12.7, 11.3, 7.4, 5.0, 3.5, 8.2, 8.4, 7.3, 10.3, 11.9, 6.0, 5.6, 10.5, 14.6, 10.8, 15.5, 7.5, 6.4, 3.4, 5.5, 6.6, 5.9, 15.0, 9.6, 7.8, 7.0, 6.9, 9.8, 3.6, 11.9, 3.7, 5.7, 6.8, 11.3, 9.3, 9.6, 10.4, 9.3, 6.9, 9.8, 9.1, 10.6, 4.5, 6.2, 8.3, 3.2, 4.9, 5.0, 6.0, 8.2, 6.3, 9.5, 9.3, 10.4, 9.7, 5.1, 6.7, 10.2, 6.2, 8.4, 7.0, 4.8, 5.6 y 4.1.

   a) Las distancias recorridas en miles de km., antes de la primera descompostura importante en cada uno de 191 autobuses, se presentan a continuación.

   Descompostura
   0, 5, 10, 18, 17, 12, 41, 45, 46, 47, 50, 55, 56, 58, 59, 43, 40, 50, 54, 54, 56, 48
   20, 39, 39, 37, 34, 21, 24, 23, 31, 39, 27, 60, 61, 70, 64, 79, 77, 71, 63, 62, 60, 65
   67, 77, 77, 78, 78, 61, 62, 66, 61, 61, 60, 60, 61, 79, 80, 80, 82, 83, 88, 88, 89, 90
   98, 99, 90, 90, 90, 89, 80, 81, 81, 82, 83, 84, 85, 85, 97, 97, 90, 90, 90, 81, 82, 88
   99, 80, 87, 83, 100, 119, 119, 118, 118, 117, 117, 116, 111, 111, 112, 101, 101, 103
   103, 107, 108, 116, 115, 114, 119, 109, 107, 100, 109, 110, 110, 111, 116, 116, 117
   102, 102, 104, 104, 103, 114, 100, 117, 112, 112, 113, 101, 105, 109, 118, 120, 139
   121, 123, 123, 124, 124, 125, 125, 129, 139, 138, 137, 121, 123, 126, 129, 137, 137
   138, 126, 127, 123, 121, 129, 130, 133, 137, 138, 128, 138, 130, 121, 140, 142, 144
   143, 156, 157, 159, 154, 157, 159, 143, 144, 145, 157, 156, 156, 165, 178, 184, 190

2. Entra en la dirección electrónica: http://illuminations.nctm.org/LessonDetail.aspx?id=L449. Ahora realiza lo siguiente:

   a) Introduce distintos conjuntos de datos.

   b) Explora los histogramas, cambiando con el deslizador la amplitud de los intervalos.

   c) Comenta todo lo que encuentres relevante en las exploraciones.

AUTOEVALUACIÓN (UNIDAD III)

1. Deberás dominar los siguientes conceptos:

Distribución de frecuencias simples
Distribución de frecuencias agrupadas
Distribución simétrica (o normal)
Distribución sesgada a la izquierda
Distribución sesgada a la derecha
Media
Mediana
Moda
Cuartiles
Gráfico de caja
Rango
Rango intercuartílico
Desviación estándar
Gráfico de tallo y hoja
Histograma
Polígono de frecuencias
Curva de frecuencias
Ojiva
Intervalo
Amplitud de intervalo
Límites de intervalo
Marca de clase.

Agrégalos a tu diccionario.

2. La siguiente distribución de frecuencias corresponde a los resultados de una encuesta realizada para contestar la pregunta: ¿cuántos libros lee usted al año?

x = número de libros leídos al año

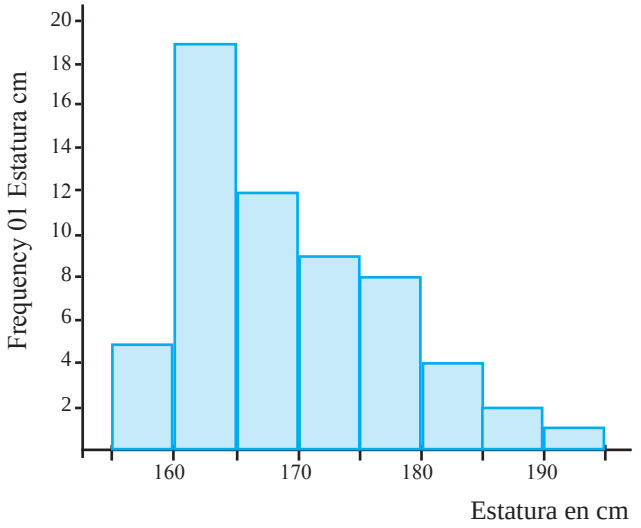
x	f
1	40
2	25
3	10
4	7
5	2
6	1
7	1

a. ¿Qué representa el «5»?
b. ¿Qué representa el «10»?
c. ¿Cuántas personas fueron encuestadas?
d. ¿Cuál es el mayor número de libros leídos por una persona encuestada?

Encuentra cada uno de los siguientes estadígrafos (muestra las fórmulas y el procedimiento)

e. Media
f. Mediana
g. Moda
h. Cuartiles
i. Varianza
j. Desviación estándar
k. Rango
l. Rango intercuartílico.

3. El siguiente histograma representa las estaturas de un grupo de estudiantes. Encuentra la respuesta para cada una de las preguntas siguientes



a. ¿Cuál es la amplitud o ancho de intervalo?
b. ¿Cuál es la marca de clase para la clase 180-185?
c. ¿Cuál es la frecuencia del intervalo (o clase) 165-170?
d. ¿Cuál es la clase que contiene a la clase modal?
e. ¿Cuál es el límite superior de la clase modal?
f. ¿Cuántos datos muestra el histograma?
g. ¿Cuál es la moda?
h. ¿Cuál es la mediana?
i. Determina la distribución de frecuencias agrupadas.

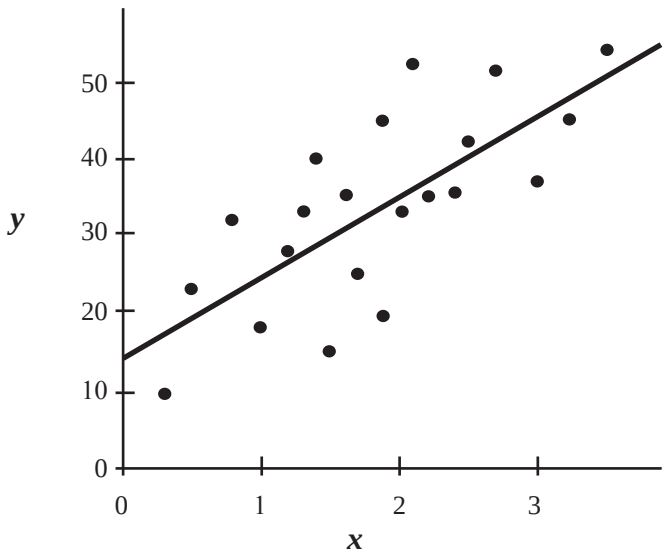
4. Los estudiantes de una preparatoria están interesados en saber qué tiempo les toma trasladarse de su casa a la escuela. Para ello, recolectaron datos con los cuales hicieron el siguiente gráfico de tallo y hoja.

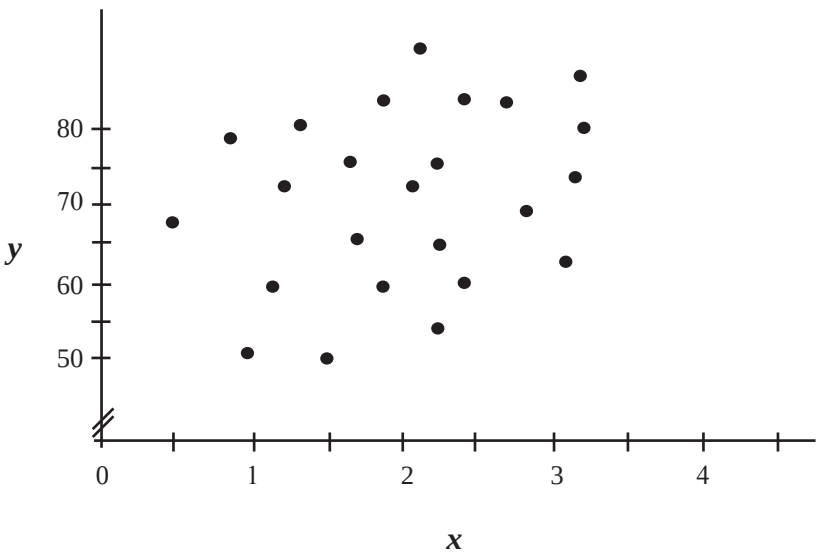
0	3 3 5 7 8 9
1	0 2 3 5 6 6 8 9
2	0 1 3 3 3 5 5 8 8
3	0 5
4	5

Con estos datos, contesta las siguiente cuestiones:

a) ¿Cuántos estudiantes hay en el grupo? ¿ En qué basas tu afirmación?
b) ¿A cuántos estudiantes les tomó menos de 15 minutos trasladarse a la escuela? ¿ En qué basas tu afirmación?
c) Escribe los tres tiempos más cortos que les tomó a estudiantes ir a la escuela..
d) Escribe los tres tiempos más largos que les tomó a estudiantes ir a la escuela.
e) ¿Cuál es el tiempo típico que toma a los estudiantes ir a la escuela? Explica tu respuesta. Puedes auxiliarte de medidas estadísticas.
f) Haz un histograma que muestre la información acerca del tiempo de tralado exhibido en el gráfico de tallo y hoja.

Exploración de datos bidimensionales





4

UNIDAD

Lección 4.1

Conceptos preliminares: Relación funcional y relación estadística, distribuciones bidimensionales, gráfico de dispersión y signo de correlación.

Objetivo: Aprende a dibujar y aplicar gráficos de dispesrsión
Empieza a comprender la naturaleza de la correlación

Actividad 15

Qué hacer

Consulta las **páginas 155-161** y resuelve:

1. El alcance de brazos, se define como la distancia existente entre los extremos de los dos brazos extendidos. A diferencia de la estatura que es más conocida por las personas, el alcance de brazos no lo es. Un fabricante de ropa deportiva, necesita saber si el conocimiento de la estatura puede decirnos algo sobre el alcande brazos. Si esto es así, a partir de las estaturas, se pueden fabricar piezas completas de trajes deportivos que incluyen pantalón y chamarra. Para tomar esta decisión, necesita contestar la siguiente pregunta: *¿Qué tan fuerte es la asociación entre estaturas y alcance de brazos? ¿Es la estatura un pronosticador útil del alcance de brazos*? Contesta esta pregunta analizando los datos adjuntos. Tu estudio debe incluir:

i) Construir un *gráfico de dispersión.*
ii) Ajustar de manera aproximada una *recta de regresión.*
iii) Explicar el tipo de *correlación.*

Estatura cm *(x)*	Alcance de brazos *cm(y)*
150	151
152	152
154	153
155	164
156	159
158	149
158	156
158	158
161	162
161	159
163	160
164	161
164	155
166	163
166	161
166	154
167	169
170	167
170	162
171	168
171	171
172	170

Hasta ahora, nuestro estudio se ha limitado a investigar el comportamiento de sólo una variable. Por ejemplo de un grupo de estudiantes estudiamos: calificaciones, estaturas, o tiempo de traslado de su casa a la escuela. En esta última unidad de nuestro curso, estudiaremos el comportamiento de dos variables simultáneamente, con el fin de determinar la posible asociación (o relación) entre ellas. Por ejemplo, ¿podría haber alguna asociación entre las calificaciones obtenidas por un estudiante y el tiempo que dedicó a estudiar?, ¿o entre su estatura y la de su papá? Este tipo de cuestiones serán abordadas en esta unidad.

Así pues, manejaremos dos variables simultáneamente. Esto nos lleva a la siguiente definición:

Datos bivariados o bidimensionales, son los valores de dos variables diferentes que se obtienen del mismo individuo poblacional.

En datos bivariados, cada elemento del colectivo tiene dos números asignados.

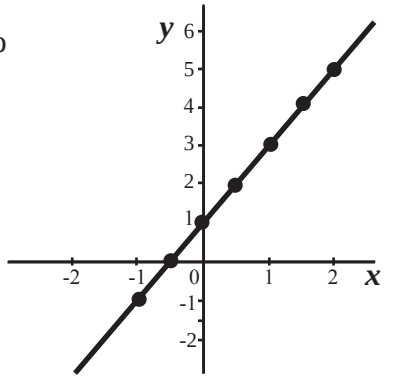
Relación funcional

Si soltamos un objeto, después de un tiempo determinado, podremos saber qué distancia ha recorrido. Esto es posible midiendo únicamente el tiempo transcurrido, puesto que existe una fórmula que nos permite calcular exactamente la distancia reccorrida en función del tiempo, a saber, $h = \frac{1}{4}gt^2$. Ésta es una relación funcional.

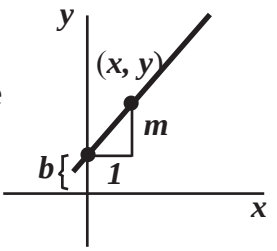
La fórmula $h = \frac{1}{4}gt^2$, expresa correlación numérica entre las variables h y t. Se dice que existe correlación puesto que una variación en el valor de h o de t, ocasiona una variación en el valor de la otra variable. Así pues, *correlación implica alguna forma de relación, asociación, función, dependencia o correspondencia.*

En tus cursos anteriores, cuando estudiaste funciones tales como $y = 2x$, estuviste frente a una **correlación perfecta,** la cual únicamente se presenta en **modelos determinísticos.**

Veamos el comportamiento de un modelo determinístico el cual se asocia a una relación funcional. Sean los conjuntos de puntos graficados en la parte derecha. Todos estos puntos se ajustan perfectamente a la recta trazada.



Recordemos que las rectas corresponden a gráficas de funciones lineales. Nos interesa reactivar cómo determinar la ecuación de una recta. Para ello, debemos tener presente lo siguiente:



- Toda recta tiene por ecuación $y = mx + b$, donde (x, y) son las coordenadas de cualquier punto de la recta; m es la pendiente de la recta y b es la ordenada en el origen

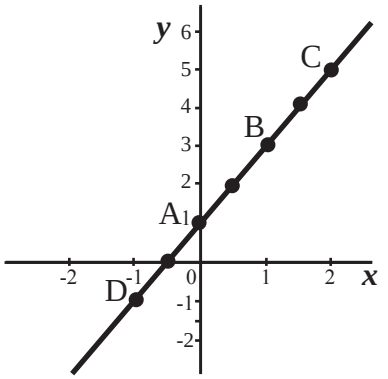
- La pendiente de una recta se determina mediante la fórmula:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

- Si conocemos dos puntos de una recta, y queremos determinar su ecuación, también podemos aplicar la fórmula

Como ejemplo, determinemos la ecuación de la recta trazada a continuación. A partir de la gráfica se pueden obtener las coordenadas de varios puntos.

x	y	Puntos
0	1	**A**(0,1)
1	3	**B**(1, 3)
2	5	**C**(2, 5)
-1	-1	**D**(-1, -1)



Fórmula: $y = mx + b$,

Datos necesarios: ***b*** y ***m.***

De la gráfica obtenemos: ***b*** **= 1**

Para determinar el valor de ***m***, sólo necesitamos las coordenadas de dos puntos de la recta. Si tomamos los puntos *A* y B, tenemos que:

A(0,1) → $x_1 = 0,\ y_1 = 1$

B(1, 3) → $x_2 = 1,\ y_2 = 3$

Fórmula:

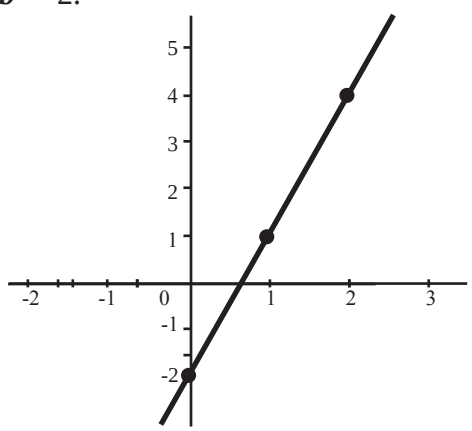
Sustituyendo en $y = mx + b$:

$$y = 2x + 1$$

Actividad 4.1 a

Resuelve:

a) Verifica que para ***x*** = 1, ***y*** = 3
b) Verifica que para ***x*** = 2, ***y*** = 5
c) Calcula el valor de ***y,*** para ***x*** = 7
d) Encuentra la ecuación de la recta con ***m*** = -3, ***b*** = 2.
e) Calcula el valor de ***x*** para ***y*** = 11.
f) Encuentra la ecuación de la recta que pasa por (3, -2) y (6, 1). Grafica la recta.
g) Encuentra la ecuación de la siguiente recta:



Relación estadística

Es evidente que sería ideal tratar siempre con relaciones funcionales, es decir utilizar ecuaciones matemáticas que nos permitieran pronosticar una cantidad exactamente en términos de otra. Pero esto, rara vez es posible. Las ciencias exactas son las que tratan, casi exclusivamente con relaciones funcionales, donde por ejemplo, a una temperatura constante la relación entre el volumen, ***V***, y la presión, ***p***, de un gas se obtiene por medio de la fórmula,

$$V = \frac{k}{p}$$, donde k es una constante.

Sin embargo, para las ciencias sociales y humanas debemos buscar la llamada **relación estadística**. Por ejemplo, se aprecia una relación estadística entre las siguientes variables:

Variables: *calificaciones-tiempo de estudio*

Se podría conjeturar que entre más estudies, obtendrías mejores calificaciones.

Variables: *estaturas padres-estaturas hijos*

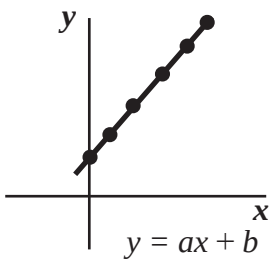
Se podría suponer que papás altos tendrán hijos altos.

El estudio que realizarás en esta unidad, te permitirá pasar de la simple conjetura, a una argumentación sólida basada en datos. Básicamente son dos las preguntas que debemos contestar sobre este tipo de cuestiones:

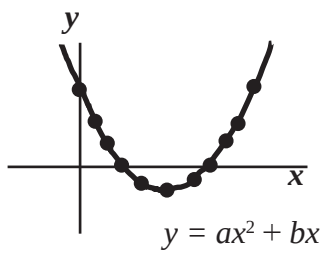
i) ¿Hay una relación entre la variable x y la variable y?
ii) ¿Qué tan fuerte es la relación entre la la variable x y la variable y?

Distribución bidimensional y gráfico de dispersión

En fenómenos que llevan directamente a una relación funcional, los puntos que representan los valores de las variables, se ajustan perfectamente a una recta o una curva.

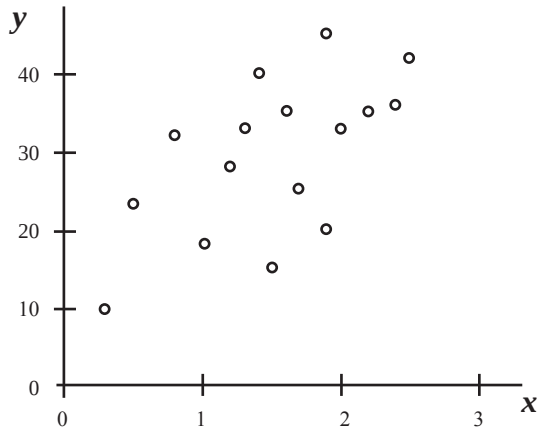


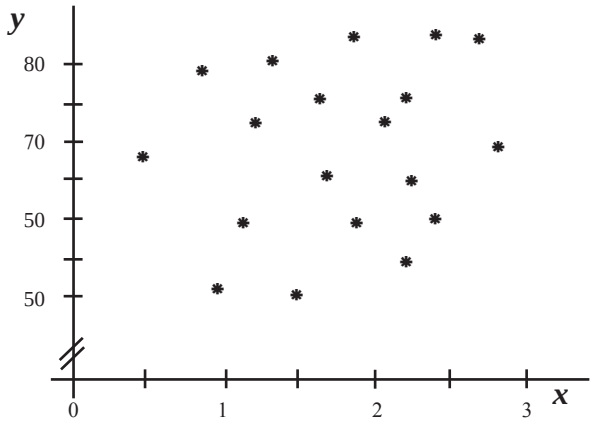
$y = ax + b$



$y = ax^2 + bx + c$

En cambio, en fenómenos que implican una relación estadística, los datos observados originan una «*nube de puntos*», como los mostrados a continuación:





Estos gráficos reciben el nombre de gráficos de dispersión y se utilizan para representar distribuciones bidimensionales.

Distribución bidimensional de una muestra de tamaño n, es un conjunto de datos originados al tomar dos medidas a cada individuo de la muestra , es decir, se investigan qué valores toman, en ellos, dos variables, X e Y. Se obtienen así un conjunto de pares de valores:

$$(x_1, y_1); (x_2, y_2); (x_3, y_3); ...; (x_i, y_i); ...; (x_n, y_n);$$

Gráfico de dispersión, es una representación gráfica para datos de dos variables (bivariados), en la que cada par de datos (x_i, y_i) es representado por un punto de coordenadas (x_i, y_i), en un sistema de ejes coordenados.

Ejemplo La tabla siguiente muestra las calificaciones obtenidas en los exámenes finales de matemáticas, mecánica e inglés por 12 estudiantes de un grupo de segundo de preparatoria.

Alumno	Calificación Matemáticas	Calificación Mecánica	Calificación Inglés
1	3	4	4
2	4	5	7
3	5	4	8
4	6	5	7
5	6	7	4
6	9	8	4
7	7	6	3
8	8	5	9
9	10	9	6
10	10	10	9
11	8	8	3
12	10	8	3

¿Existe alguna relación entre:

- *Calificaciones de matemáticas y calificaciones de mecánica*

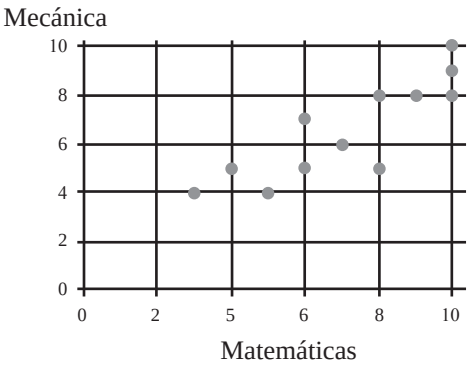
- *Calificaciones de matemáticas y calificaciones de inglés?*

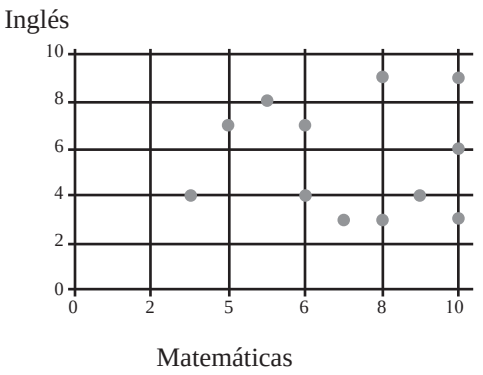
Analiza cuidadosamente los datos presentados y llega a una conclusión. compártela con tus compañeros.

Ejemplo (Cont.)

A continuación construimos un gráfico de dispersión para las dos distribuciones bidimensionales señaladas:

-Calificaciones de matemáticas y calificaciones de mecánica
-Calificaciones de matemáticas y calificaciones de inglés?



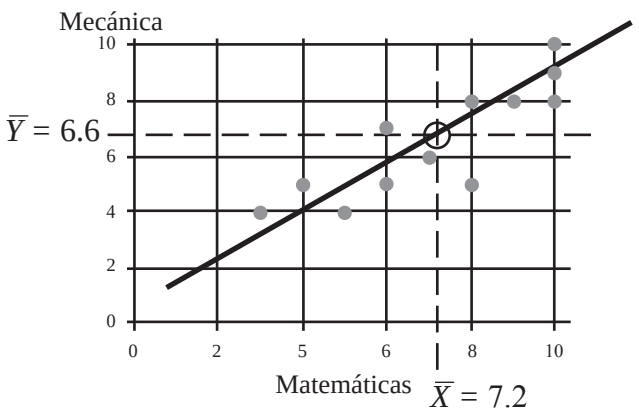


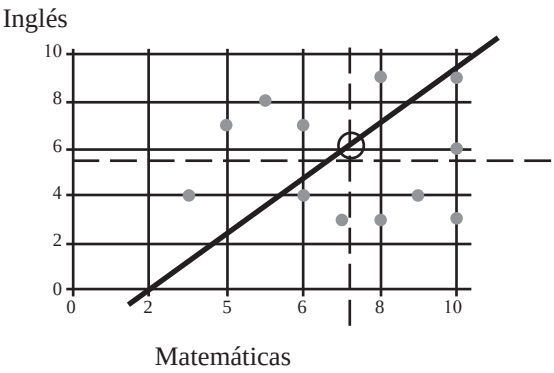
Los gráficos nos permiten observar que en la primera distribución (matemáticas-mecánica), los puntos están más alineados y, por tanto, **la correlación** o (relación) entre las variables **es más fuerte**. En la segunda se ve una **correlación débil**.

La fuerza o debilidad de una correlación se percibe mejor si trazamos una recta que se ajuste lo mejor posible a la nube de puntos. Esta recta se llama **recta de regresión.**

En un primer momento, estas rectas pueden trazarse a ojo, intentando que crucen por el «centro» de las nube de puntos». Para lograr esto, ayuda mucho el pasar la recta por el punto $(\bar{x}, \bar{y})$, es decir, por el punto con abscisa igual a la media de la primer variable y como ordenada la media de la segunda variable.

Gráficos de dispersión con la recta de regresión.





Resuelve

a) Calcula la media de las calificaciones en matemáticas. Llámala $\overline{x}$.

b) Calcula la media de las calificaciones en mecánica. Llámala $\overline{y}$.

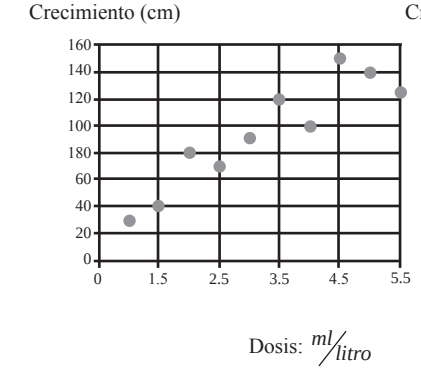
c) Calcula la media de las calificaciones en inglés. También llámala $\overline{y}$.

Signos de correlación

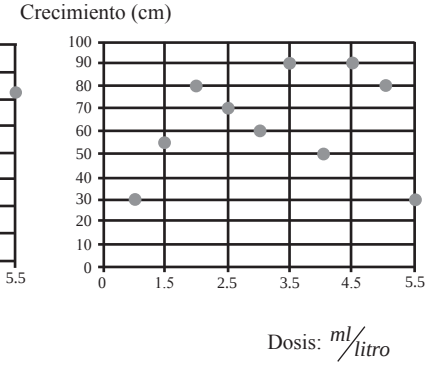
Consideremos los resultados de un experimento.

Con la finalidad de probar la efectividad biológica, medida por el crecimiento de la planta de melón, se suministraron en cada una de diez zonas de invernadero una dosis distinta de un producto A y se midió el crecimiento de la planta al cabo de un mes. De manera idéntica se repite el experimento para dos productos B y C. Los resultados se muestran gráficamente:

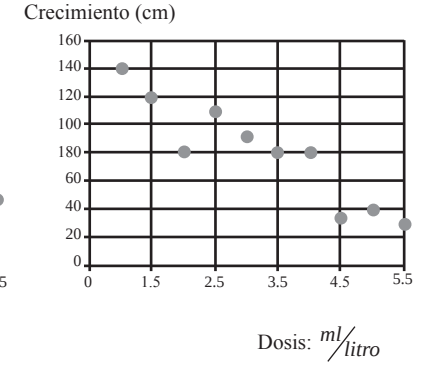
Producto A



Producto B



Producto C



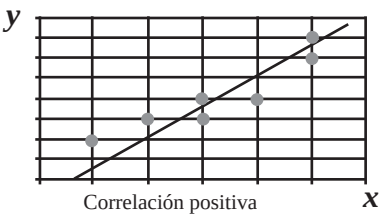
A partir del análisis de los gráficos, puede concluirse que el producto *A* favorece el crecimiento de las plantas, el *B* no influye y *C* es perjudicial.

La **correlación** de la primer gráfica es **positiva** y de la última **negativa**. Es decir, el signo de la correlación es igual al signo de la pendiente de la recta de regresión correspondiente.

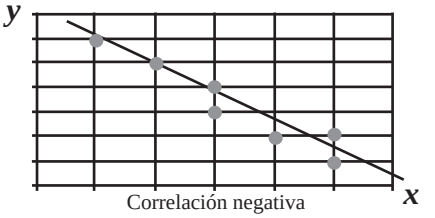
En el segundo gráfico, la nube de puntos no muestra un patrón claro que nos permita establecer una tendencia, y esto sugiere que no hay correlación entre las variables (se dice que son **no correlacionadas**).

En resumen:

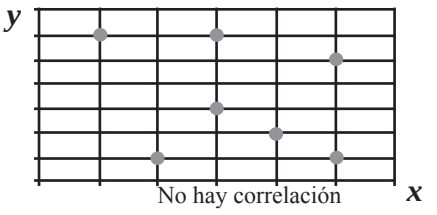
- Correlación positiva:
  *si aumenta **x**-aumenta **y***



- Correlación negativa:
  *si aumenta **x**-disminuye **y***



- No hay correlación :
  si no hay una tendencia clara



Ejercicio 4.1

1. La tabla siguiente muestra los resultados obtenidos por los equipos de futbol de primera división en el torneo de clausura 2008.

Clasificación (C)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Partidos ganados (G)	11	10	8	7	7	6	5	6	5	5	5	6	5	3	4	2	3	2
Partidos empatados (E)	3	6	4	5	5	7	8	5	7	7	6	6	6	8	5	8	5	7
Partidos perdidos (P)	3	1	5	5	5	4	4	6	5	5	6	5	6	6	8	7	9	8
Goles a favor (F)	42	34	23	28	21	21	21	25	25	19	21	22	22	18	21	15	17	26
Goles en contra (GC)	23	19	17	22	24	17	22	27	19	20	20	24	29	21	27	26	30	33

Estudiar la correlación entre:

a) La variable C y G.
b) La variable C y E.
c) La variable C y P.
d) Entre G y F.
e) Entre P y GC.

2. La tabla de la derecha, muestra el efecto de la fecha de siembra en la producción de maíz. Estudiar la correlación entre fecha de siembra y pérdida de rendimiento.

Fecha de siembra	Pérdida de rendimiento (Ton/hect)
15-sept	1.96
20-sept	1.62
25-sept	1.31
30-sept	1.03
05-oct	0.772
10-oct	0.548
15-oct	0.353
20-oct	0.187
25-oct	0
30-oct	0
05-nov	0
10-no	0
15-nov	0
20-nov	0
25-nov	0
30-nov	0
05-dic	0.021
10-dic	0.15
15-dic	0.309
20-dic	0.497
25-dic	0.714
30-dic	0.96
05-enero	1.29
10- enero	1.6
15- enero	1.94

Lección 4.2 Medición de la correlación: el *CRCC*

Objetivo: Comprende de manera intuitiva cómo se mide la correlación a través del Coeficiente de Razón de Conteo de Cuadrantes (***CRCC***).

Actividad 16

Qué hacer

Consulta las **páginas 163 a 165** y resuelve.

1) La tabla siguiente muestra las estaturas de 8 padres y sus respectivos hijos a la edad de 12 años.

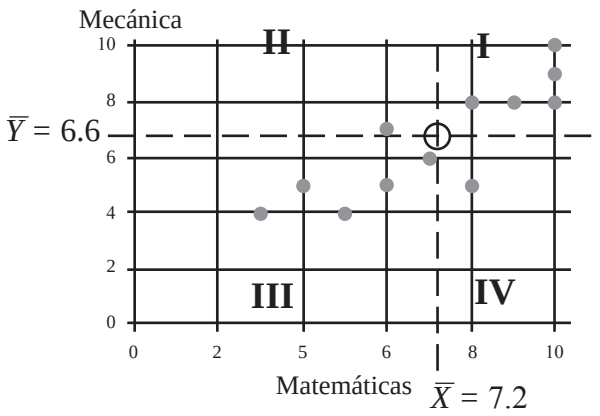
Estatura de **padre *x***	1.80	1.76	1.63	1.79	1.68	1.79	1.71	1.84
Estatura de **hijo *y***	1.60	1.46	1.40	1.49	1.50	1.48	1.43	1.64

Calcula El *CRCC* e interprétalo.

Medir la fuerza de la correlación o asociación entre dos variables es un concepto estadístico importante. Esta fuerza se mide con el coeficiente de correlación de Pearson. Antes de estudiar este coeficiente, estudiaremos una manera aproximada de medir la correlación.

Medición de la correlación

Para comprender la naturaleza de la medición de la correlación, repetiremos el gráfico de dispersión correspondiente a las calificaciones en matemáticas y mecánica., incluyendo una línea vertical trazada a través de la calificación media en matemáticas ($\overline{Y} = 6.6$) y una línea horizontal trazada a través de la media de calificación en mecánica ($\overline{X} = 7.2$).



Las dos líneas dividen el gráfico de dispersión en cuatro regiones (o cuadrantes). En nuestro ejemplo, la región superior derecha (cuadrante I) contiene puntos de alumnos con calificación tanto en matemáticas como en mecánica más altas que el promedio. La región superior izquierda (Cuadrante II) contiene puntos que corresponden a alumnos con calificación en matemáticas abajo del promedio y calificación en mecánica más alto que el promedio. La región inferior izquierda (cuadrante III) contiene puntos que corresponden a alumnos con calificación tanto en matemáticas como en mecánica abajo del promedio. La región inferior derecha (cuadrante 4) contiene puntos que corresponden a alumnos con calificación en matemáticas más altas que el promedio y calificación en mecánica abajo del promedio.

Se observa que la mayoría de los puntos en el gráfico de dispersión están ya sea en el cuadrante I o en el cuadrante III. Esto es, la mayoría de los alumnos con calificación en matemáticas más alta que el promedio también tiene calificación en mecánica más alta que el promedio (cuadrante I), y la mayoría de los alumnos con calificación en matemáticas abajo del promedio también tiene calificación en mecánica abajo del promedio (cuadrante III). Un alumno tiene calificación en matemáticas abajo del promedio con calificación en mecánica mayor que el promedio (cuadrante II) y también un alumno tiene calificación en matemáticas mayor que el promedio y calificación en mecánica abajo del promedio (cuadrante IV).

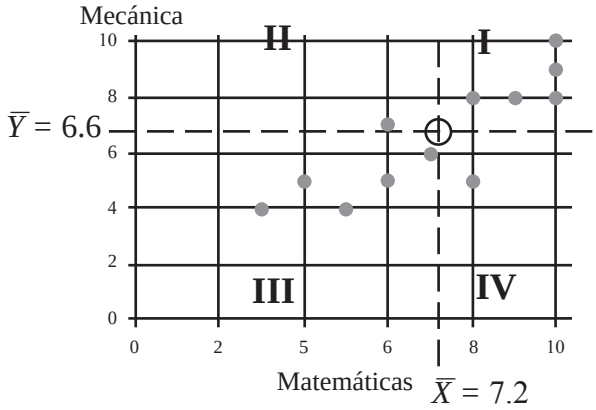
Estos resultados nos permiten hacer la siguiente afirmación:

De manera análoga, ***asociación o correlación negativa*** entre dos variables ocurre cuando valores abajo del promedio de una variable tienden a ocurrir con valores arriba del promedio de la otra y cuando valores arriba del promedio de una variable tienden a ocurrir con valores abajo del promedio de la otra.

Estas ideas nos llevan directamente al Coeficiente de razón de Conteo de Cuadrantes.

Coeficente de Razón de Conteo de Cuadrantes (*CRCC*)

Un coeficiente de correlación es una cantidad que mide la dirección y fuerza de una asociación entre variables. Volvamos al ejemplo previo sobre las variables calificaciones:



Se observa que puntos en el cuadrante I y III contribuyen a la correlación postiva entre calificación en matemáticas y calificación en mecánica, y hay un total de 10 puntos en esos dos cuadrantes. Puntos en los cuadrantes II y IV no están contribuyendo a la correlación positiva entre calificación en matemáticas y calificación en mecánica, y hay un total de 2 puntos en esos dos cuadrantes. Por esta razón, el Coeficiente Razón de Conteo de Cuadrante (CRCC), se define de la siguiente manera:

Coeficiente de Razón de Conteo de Cuadrantes:

$$CRCC = \frac{\left(\begin{array}{l}\textit{Número de puntos}\\ \textit{en los cuadrantes I y III}\end{array}\right) - \left(\begin{array}{l}\textit{Número de puntos}\\ \textit{en los cuadrantes II y IV}\end{array}\right)}{\textit{Total de puntos en los cuatro cuadrantes}}$$

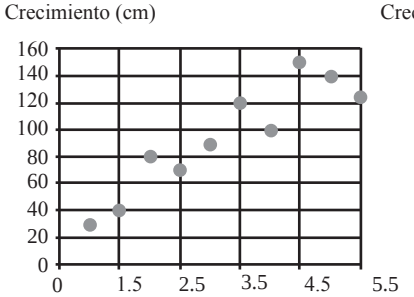
El *CRCC* entre las calificaciones en matemáticas y mecánica es:

$$CRCC = \frac{10-2}{12} = \frac{8}{12} = 0.67$$

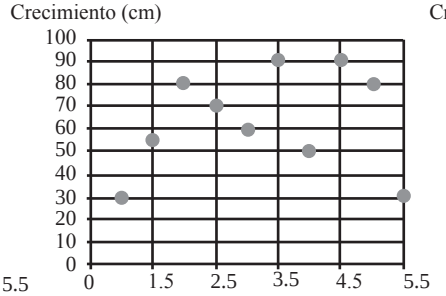
Actividad 4.2 a

A continuación se repiten los gráficos de dispersión relativos a los experimentos para determinar la relación entre dosis de tres productos y crecimiento de las plantas de melón.

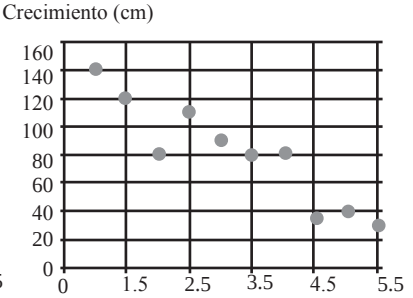
Producto A



Producto B



Producto C



Calcula el *CRCC* para cada gráfica. Compara su valor con la correlación que se presenta entre las variables en cada caso. ¿Qué concluyes?

A partir de la definción del *CRCC* pueden obtenerse las siguientes propiedades:

- El *CRCC* es menor que 1.
- El *CRCC* está siempre entre -1 y 1.

Ejercicio 4.2

En el ejercicio 4.1 analizaste la tabla siguiente que muestra los resultados obtenidos por los equipos de futbol de primera división en el torneo de clausura 2008.

Clasificación (*C*)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Partidos ganados (*G*)	11	10	8	7	7	6	5	6	5	5	5	6	5	3	4	2	3	2
Partidos empatados (*E*)	3	6	4	5	5	7	8	5	7	7	6	6	6	8	5	8	5	7
Partidos perdidos (*P*)	3	1	5	5	5	4	4	6	5	5	6	5	6	6	8	7	9	8
Goles a favor (*F*)	42	34	23	28	21	21	21	25	25	19	21	22	22	18	21	15	17	26
Goles en contra (*GC*)	23	19	17	22	24	17	22	27	19	20	20	24	29	21	27	26	30	33

Incorporando el cálculo del *CRCC* vuelve a estudiar la correlación entre:

a) La variable *C* y *G*.
b) La variable *C* y *E*.
c) La variable *C* y *P*.
d) Entre *G* y *F*
e) Entre *P* y *GC*.

Lección 4.3 Medición de la correlación: el *Coeficiente de correlación de Pearson.*

Objetivo: Comprende, calcula e interpreta el coeficiente de correlación de Pearson.

Actividad 12

Qué hacer

Consulta las **páginas 167 a 174** y resuelve.

1) La tabla siguiente muestra las estaturas de 8 padres y sus respectivos hijos a la edad de 12 años.

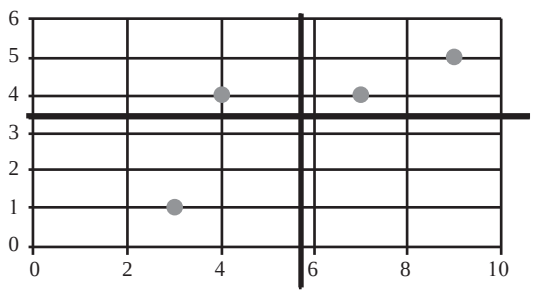
Estatura de **padre *x***	1.80	1.76	1.63	1.79	1.68	1.79	1.71	1.84
Estatura de **hijo *y***	1.60	1.46	1.40	1.49	1.50	1.48	1.43	1.64

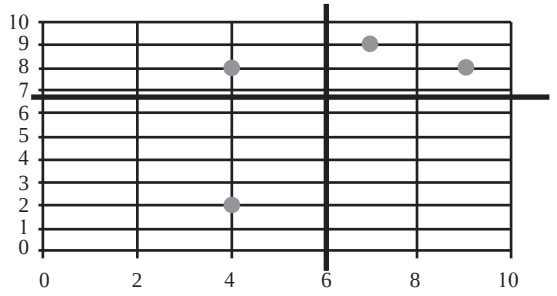
Calcula el *Coeficiente de correlación* e interprétalo.

El *CRCC* es una medida de la fuerza de correlación basada sólo en el número de puntos que caen en cada cuadrante. Sin embargo, la posición de los puntos con respecto a los valores promedios, tiene gran influencia en la correlación.

Por ejemplo, observa los siguientes gráficos de dispersión y contesta:

- ♦ Observando los gráficos, ¿cuál presenta menor correlación?
- ♦ ¿Cuál distribución tiene un *CRCC* menor?





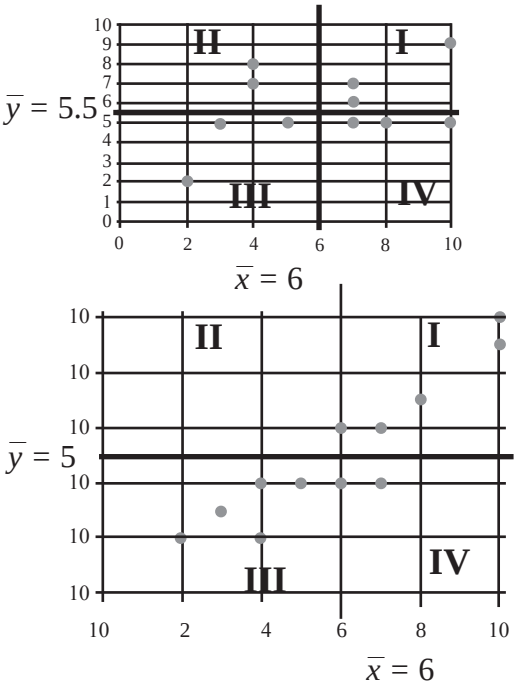
Debes observar que el gráfico de la izquierda presenta mayor correlación que el de la derecha. Sin embargo, ambas distribuciones tienen un *CRCC* igual a 0.5. Debemos entonces, buscar otra forma de medir la correlación entre dos variables.

La covarianza

En la búsqueda de un nuevo coeficiente de correlación, surge el concepto estadístico denominado **covarianza.** Para llegar a su definición, sigue el siguiente análisis:

1) Para cada gráfico de dispersión de la derecha, calcula $(x_i - \bar{x})$ y $(\bar{y}_i - \bar{y})$ y para todos los puntos de cada uno de los cuadrantes.

¿Cuál gráfico presenta mayor correlación?



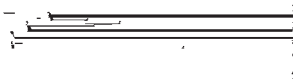
Debes comprobar los signos que aparecen en la tabla siguiente:

	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})$ $(y_i - \bar{y})$
En el cuadrante I	+	+	+
En el cuadrante II	-	+	-
En el cuadrante III	-	-	+
En el cuadrante IV	+	-	-

Si trazas la recta de regresión, puedes verificar que cuanto más fuerte es la correlación, más apretados está los puntos en torno a la recta, y más puntos hay en los cuadrantes I y III y menos en los cuadrantes II y IV.

2) Si tenemos en cuenta que en los cuadrantes I y III el producto $(x_i - \bar{x})$ $(y_i - \bar{y})$ es positivo, mientras que en los otros dos cuadrantes es negativo, deducimos que, cuanto más fuerte es la correlación, más puntos hay para los que el producto $(x_i - \bar{x})$ $(y_i - \bar{y})$ es positivo y menos puntos para los que el producto es negativo.

Entonces, cuanto mayor es la correlación, mayor es la suma



De manera semejante puede verificarse que para correlación negativa, la suma

es negativa y tanto más grande sea en valor absoluto, más estrecha será la relación entre ambas variables.

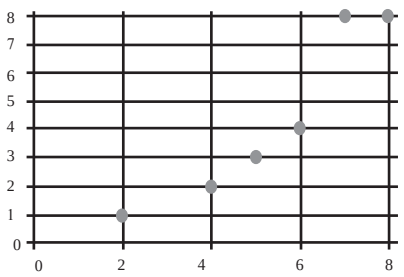
Ahora bien, para efecto de poder comparar distribuciones de tamaños diferentes, la suma se divide entre ***n -1*** cuando se trata de una muestra (tal y como se hizo para la varianza). De esta manera se obtiene la medida estadística llamada covarianza:

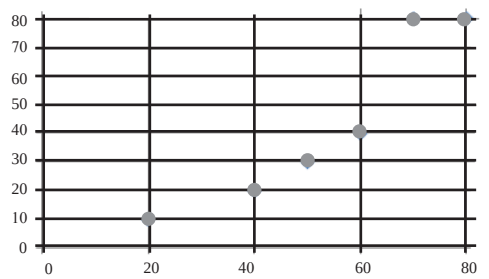
$$\text{Covarianza} = \frac{\quad}{n-1}$$

Al igual que la desviación media fue el camino para llegar a la desviación estándar, la covarianza cobra importancia porque nos lleva al coeficiente de correlación de Pearson.

El coeficiente de correlación de Pearson

El coeficiente de correlación surge a partir de un inconveniente de la covarianza, el cual se muestra a continuación.





Los gráficos de dispersión son idénticos, con excepción en la unidad utilizada en los ejes. La covarianza del primer gráfico es 6.07 y, la del segundo 607. El cambio de escala influye en la covarianza. Este problema se corrige dividiendo la covarianza entre el producto de las desviación estándar de las variables ***x*** e ***y.*** De esta manera, surge el coeficiente de correlación denotado con la letra ***r:***

$$\text{Coeficiente de correlación} = r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{(n-1)\,s_x\,s_y} \qquad \textbf{(A)}$$

s_x: *desviación estándar de la* variable *x*

s_y: *desviación estándar de la* variable *y*

La fórmula anterior, que nos permite calcular el coeficiente de correlación , requiere de cálculos poco prácticos, por lo que a continuación procederemos a desarrollar una fórmula equivalente de fácil manejo.

La nueva fórmula para r, surge a partir de las siguientes consideraciones:

- Como ya es conocido, la definición de desviación estándar produce las siguientes fórmulas:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad ; \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \qquad \textbf{(B)}$$

Sustituyendo (B) en (A):

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \textbf{(C)}$$

- Ahora, transformaremos la expresión del numerador y la de los dos radicales (a partir de este momento, para simplificar la notación, omitiremos los subíndices):

$$\begin{aligned}
\Sigma(x-\bar{x})^2 &= \Sigma\left(x^2-2x\bar{x}+\bar{x}^2\right)\\
&= \Sigma x^2 - 2\bar{x}\Sigma x + \Sigma\bar{x}^2\\
&= \Sigma x^2 - 2\bar{x}\frac{\Sigma x}{n}n + n\bar{x}^2\\
&= \Sigma x^2 - 2\bar{x}^2 n + n\bar{x}^2\\
&= \Sigma x^2 - n\bar{x}^2\\
&= \Sigma x^2 - n\left(\frac{\Sigma x}{n}\right)^2\\
&= \Sigma x^2 - \frac{(\Sigma x)^2}{n}
\end{aligned}$$

Recuerda que esta última expresión recibe el nombre de «suma de cuadrados de x» y se denota con $SC(x)$.

Entonces,

$$CS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \qquad \textbf{(D)}$$

Por un simple cambio de variable, definimos «la suma de cuadrados de y»

$$CS(y) = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \qquad \textbf{(E)}$$

De manera análoga, puede demostrarse la equivalencia entre las siguientes expresiones:

A la segunda expresión le llamaremos «suma de productos ***xy***.

$$SP(xy) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} \qquad \textbf{(F)}$$

Sustituyendo (D), (E) y (F) en (C):

$$r = \frac{SP\,(xy)}{\sqrt{SP(x)}\,\sqrt{SC(y)}} = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}\,\sqrt{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}}}$$

Esta expresión aparentemente compleja, sólo necesita los siguientes cálculos:

x	y	x^2	y^2	xy
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
Σx	Σy	Σx^2	Σy^2	Σxy

Ejemplo Vamos a retomar el ejemplo previo que trata sobre las variables calificaciones.

Alumno	Calificación Matemáticas	Calificación Mecánica	Calificación Inglés
1	3	4	4
2	4	5	7
3	5	4	8
4	6	5	7
5	6	7	4
6	9	8	4
7	7	6	3
8	8	5	9
9	10	9	6
10	10	10	9
11	8	8	3
12	10	8	3

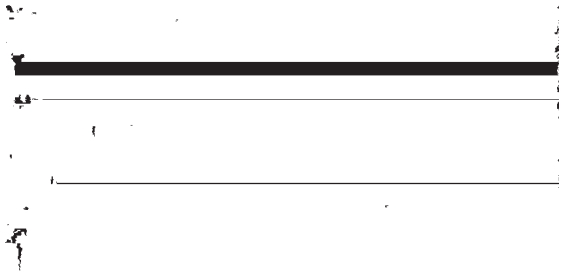
Nos interesa el coeficiente de correlación entre las variables calificaciones en matemáticas y calificaciones en mecánica.

Solución

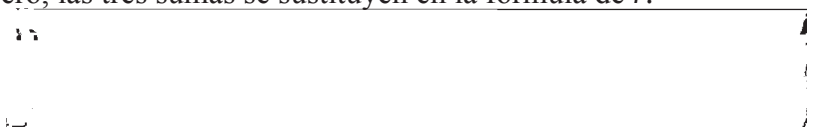
Primero, es necesario elaborar una tabla ampliada, enumerando tanto los valores de x como de y, así como x^2, y^2 y xy que son necesarios en la fórmula

Calificación Matemáticas (x)	x^2	Calificación Mecánica (y)	y^2	xy
3	9	4	16	12
4	16	5	25	20
5	25	4	16	20
6	36	5	25	30
6	36	7	49	42
9	81	8	64	72
7	49	6	36	42
8	64	5	25	40
10	100	9	81	90
10	100	10	100	100
8	64	8	64	64
10	100	8	64	80
Sumas	$\Sigma x = 85$	$\Sigma x^2 = 680$	$\Sigma y = 79$	$\Sigma y^2 = 565$	$\Sigma xy = 612$

Segundo, se calculan *SC(x), SC(y) y SP(xy):*



Tercero, las tres sumas se sustituyen en la fórmula de *r:*



Actividad 4.3 a

Calcula el coeficiente de correlación para las variables calificación en matemáticas y calificación en inglés.

Propiedades del coeficiente de correlación

Los valores del coeficiente de correlación lineal ayudan a responder la pregunta ¿existe una correlación lineal entre las dos variables en consideración?

Por ejemplo, sabíamos de antemano, que la relación entre las calificaciones en matemáticas y las de mecánica era alta, y el coeficiente de correlación fue $r = 0.86$. También sabíamos que la relación entre calificaciones en matemáticas e inglés era baja y en la actividad (4.3 a) debiste encontra que $r = -0.02$. Estos resultados nos permiten afirmar que: cuando el valor de r es cercano a cero, se concluye que hay poca correlación lineal o que no hay correlación lineal. A medida que el valor calculado de r cambia de 0 a +1 ó -1, la correlación lineal se hace cada vez más fuerte entre las dos variables.

En resumen:

- El valor de r siempre oscila entre -1 y 1.
- Cuando r es -1 ó 1 la relación es funcional. Por tanto, en este caso el valor de una variable se obtiene con seguridad, a partir de la otra.
- A medida que el coeficiente de correlación se acerca a 1 ó -1, la nube de puntos se hace más estrecha en torno a la recta de regresión.

Ejercicio 4.3

En el ejercicio 4.2 analizaste la tabla siguiente que muestra los resultados obtenidos por los equipos de futbol de primera división en el torneo de clausura 2008.

Clasificación (C)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Partidos ganados (G)	11	10	8	7	7	6	5	6	5	5	5	6	5	3	4	2	3	2
Partidos empatados (E)	3	6	4	5	5	7	8	5	7	7	6	6	6	8	5	8	5	7
Partidos perdidos (P)	3	1	5	5	5	4	4	6	5	5	6	5	6	6	8	7	9	8
Goles a favor (F)	42	34	23	28	21	21	21	25	25	19	21	22	22	18	21	15	17	26
Goles en contra (GC)	23	19	17	22	24	17	22	27	19	20	20	24	29	21	27	26	30	33

Incorporando el cálculo del *Coeficiente de correlación lineal,* vuelve a estudiar la correlación entre:

a) La variable C y G.
b) La variable C y E.
c) La variable C y P.
d) Entre G y F
e) Entre P y GC.

Lección 4.4 Regresión lineal

Objetivo: Desarrolla destrezas para calcular ecuaciones de rectas de regresión

Actividad 18

Qué hacer

Consulta las **páginas 175 a 183** y resuelve.

1) La tabla siguiente muestra las estaturas de 8 padres y sus respectivos hijos a la edad de 12 años.

Estatura de **padre** x	1.80	1.76	1.63	1.79	1.68	1.79	1.71	1.84
Estatura de **hijo** y	1.60	1.46	1.40	1.49	1.50	1.48	1.43	1.64

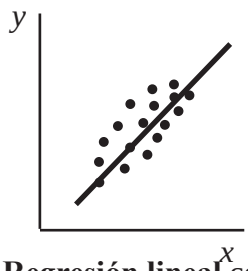
i) Construir un diagrama de dispersión.
ii) Calcula la recta de regresión.

El coeficiente de correlación mide la intensidad de una relación lineal, pero no dice nada sobre la relación matemática que hay entre las dos variables. El coeficiente de correlación no ayuda a predecir el valor de la variable *y*.

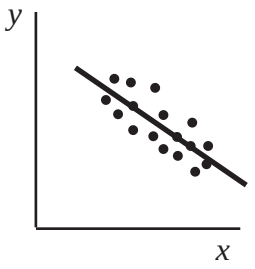
Análisis de regresión se refiere al estudio de relaciones entre variables. Si la nube de puntos en un gráfico de dispersión tiene una forma lineal, la línea recta puede ser un modelo realista de la relación entre las variables bajo estudio. El análisis de regresión encuentra la ecuación de la recta que describe mejor la relación entre las dos variables. Una aplicación de esta ecuación es hacer predicciones. Hay muchas situaciones en las que se aplican estas predicciones; por ejemplo, es de interés predecir cuál será la población de un país en el futuro, o la producción de maíz, o la distancia necesaria para detener un automóvil conociendo su velocidad.

Se busca expresar la relación matemática entre dos variables mediante una expresión algebraica. Dependiendo de la forma que adopta la nube de puntos, se establece un modelo general de varios posibles.

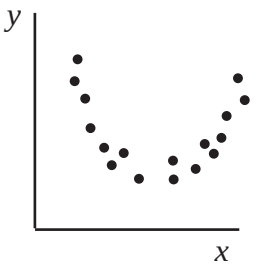
Las figura siguientes muestran patrones de datos de dos variables que parecen tener alguna relación entre sí.



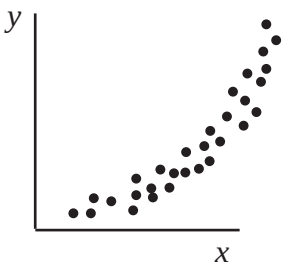
Regresión lineal con pendiente positiva.
Modelo: $\hat{y} = b_0 + b_1 x$



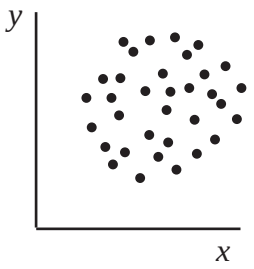
Regresión lineal con pendiente negativa
Modelo: $\hat{y} = b_0 + b_1 x$



Regresión cuadrática
Modelo: $\hat{y} = a + bx + cx^2$



Regresión exponencial
Modelo: $\hat{y} = a(\ b^x\ )$



Variables que no están relacionadas.

En este curso sólo estudiarás la recta de regresión, válida cuando la nube de puntos adopta una forma aproximadamente rectilínea.

Recta de regresión

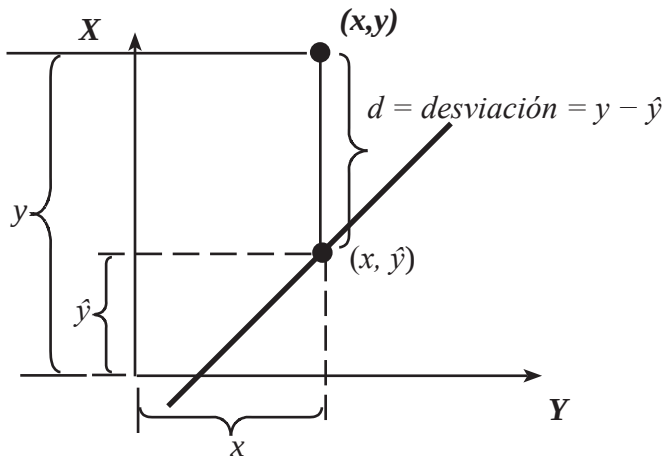
Ya hemos aprendido que la recta de regresión es una recta trazada entre los puntos de una distribución bidimensional. Ahora, se trata de buscar la recta que se ajuste mejor a la nube de puntos. Recordemos que las rectas en el plano están definidas por ecuaciones lineales con dos incógnitas, las cuales son de la forma $y = mx + b$. Esta manera de presentar las ecuaciones lineales es comúnmente usada en otras ramas de las matemáticas. Sin embargo, en estadística la que más se usa es $y = a + bx$ o bien $y = ax + b$.

En este texto usaremos $y = ax + b$.

Entonces, ***a*** representará a la pendiente de la recta y *b* seguirá representando la ordenada en el origen.

El método de los mínimos cuadrados

Ahora estudiaremos un método que nos permite obtener la recta que mejor se ajuste a los valores de una distribución bidimensional. Este método se llama el «***método de los mínimos cuadrados***». La recta de mínimos cuadrados cruza a través del centro (en algún sentido) de la nube de puntos. La figura siguiente muestra las notaciones que utilizaremos.



- (x,y) son las coordenadas de la observación real.
- $(x, \hat{y})$ son las coordenadas del punto que es considerado en la recta de regresión en lugar de (x,y).
- $\hat{y}$ (que se lee «*y* sombrero») representa el valor estimado de y que corresponde a un valor particular de x.
- d representa la desviación en la dirección «*y*» entre el punto del gráfico de dispersión y la recta de regresión.

Por tanto,

$$d = y - \hat{y}$$

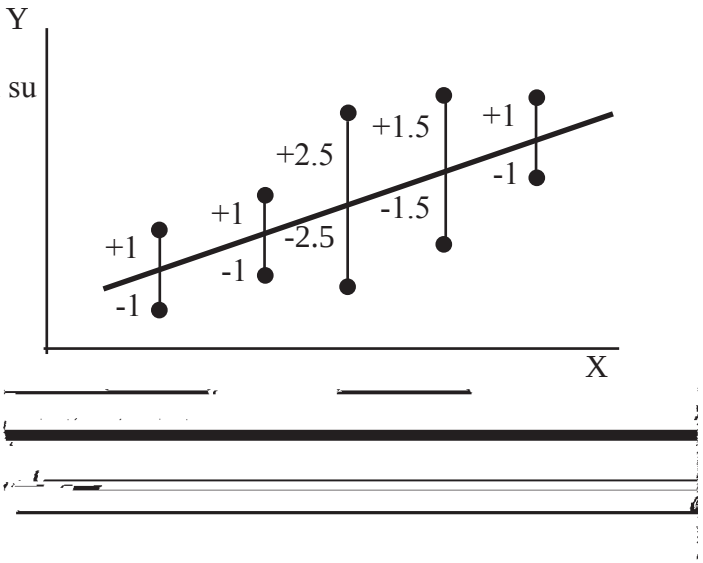
Obsérvese que d es positivo cuando el punto (x,y) está por arriba de la recta y negativo cuando está por abajo de la recta.

En este tema que nos acupa de regresión, *dispersión* es la variación alrededor de la recta de mínimos cuadrados (como si midiera la desviación estándar de las desviaciones definidas por $d = y - \hat{y}$).

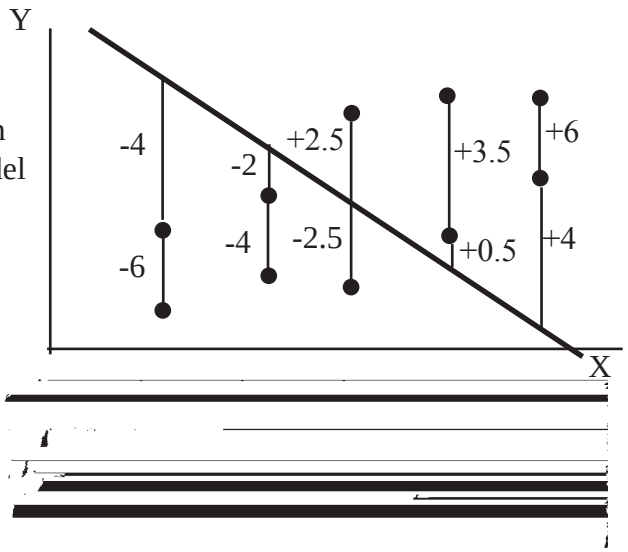
El problema de encontrar la ecuación de la recta de mejor ajuste, se convierte en encontrar una recta que produzca las menores desviaciones. Debido a que estas desviaciones pueden ser negativas, el problema cambia a buscar una recta que cause la suma mínima de los desvíos al cuadrado.

Estudia los siguientes gráficos de dispersión que muestran como la suma $\Sigma\,(y - \hat{y})^2$ efectivamente cambia al cambiar de recta.

Gráfico de dispersión con su recta del mejor ajuste



Mismo gráfico de dispersión con una recta que no es la del mejor ajuste



Se requiere pues, encontrar los valores de las constantes ***a*** (*pendiente*) y ***b*** *(ordenada en el origen)* tales que $\Sigma\,(y - \hat{y})^2$ sea lo más pequeña posible.

Los valores de estas constantes, que satisfacen el criterio de mínimos cuadrados, se encuentran aplicando las fórmulas siguientes:

Pendiente:

$$a = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

Pero, ya conocemos fórmulas equivalentes para estas expresiones:

Suma de productos xy: [redacted]

Suma de cuadrados de x: [redacted]

Entonces:

$$\text{Pendiente: } a = \frac{\text{Suma de productos } xy}{\text{Suma de cuadrados de } x} = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

Ordenada en el origen:

$$b = \frac{(\text{Suma de } y) - (\text{pendiente})\,(\text{suma de } x)}{n} = \frac{\Sigma y - (a \cdot \Sigma x)}{n}$$

La recta del mejor ajuste siempre pasa por el punto $(\bar{x}, \bar{y})$. Entonces, una vez conocido a, podemos sustituir $\bar{x}$ y $\bar{y}$ en $y = ax + b$ y despejar b. Sin embargo, se recomienda calcular b con la expresión anterior y usar la sustitución de $\bar{x}$ y $\bar{y}$ como comprobación del procedimiento seguido.

Ejemplo

A continuación determinaremos la recta de regresión entre las variables calificaciones en matemáticas y calificaciones en mecánica.

Solución

Ecuación buscada: $\hat{y} = ax + b$

En donde: [redacted]

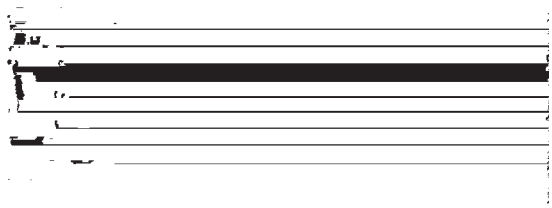
Las cantidades que necesitamos para estas expresiones, ya fueron calculadas con el coeficiente de correlación. Ver página 173.

$SP\,(xy) = 45.8$

$SC\,(x) = 63.7$

$\Sigma y = 79$

$\Sigma x = 86$

Por lo que: 

Sustituyendo estos valores en $\hat{y} = ax + b$, obtenemos:

$$\hat{y} = 0.72x + 1.42$$

Esta es la recta de mínimos cuadrados entre las variables calificaciones en matemáticas y calificaciones en mecánica.

Comprobación: las coordenadas  deben satisfacer la ecuación:

$$\hat{y} = 0.72x + 1.42$$

$$6.58 = 0.72(7.17) + 1.42$$

$$6.58 = 5.16 + 1.42$$

$6.58 = 6.58$ Correcto

La recta de regresión para hacer predicciones

La recta de regresión nos proporciona de manera aproximada el valor esperado de *y,* para un cierto valor de *x,* o viceversa, A estos valores se les llama estimaciones.

Por ejemplo, hemos encontrado que la recta de regresión entre las calificaciones en matemáticas y en mecánica es $\hat{y} = 0.72x + 1.42$.

Si quisiéramos estimar la calificación en mecánica de un estudiante que obtuvo 10 en matemáticas, sustituimos $x = 10$ en la ecuación:

$$\hat{y} = 0.72(10) + 1.42 = 8.62$$

En los datos originales puede apreciarse que para un 10 en matemáticas, existen dos calificaciones para mecánica de 9 y 10. Así que, no debemos esperar que el valor estimado ocurra exactamente; en vez de eso, $\hat{y}$ es el

número promedio de «calificación en mecánica» que debemos esperar de todos los estudiantes que obtuvieron 10 en matemáticas.

Actividad 4.4 a

Determina la recta de regresión entre las variables caficaciones en matemáticas y calificaciones en inglés. Comenta los resultados.

Series de tiempo

Los gráficos de dispersión también se pueden utilizar para obtener gráficamente la relación de una variable numérica en el tiempo. Estos gráficos se llaman series de tiempo.

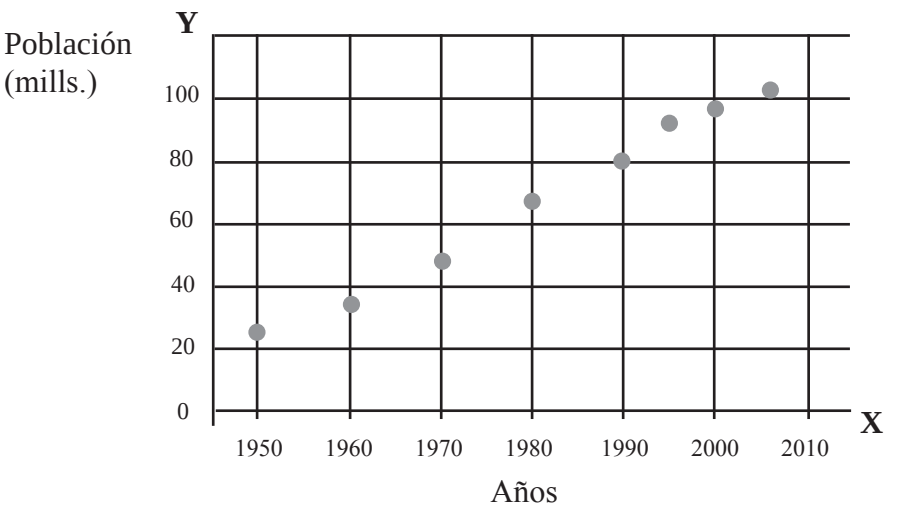
Series de tiempo, son distribuciones de pares **(*x*, *y*)** en los cuales ***x*** corresponde a la variable tiempo, que se expresa en periodos que pueden ser años, meses, días u otra unidad adecuada al tipo de problema que se esté trabando.

El método de los mínimos cuadrados, se puede utilizar para describir la tendencia de una serie de tiempo.

Ejemplo

La población en México, ha crecido considerablemente en los últimos años como podemos apreciar en la siguiente tabla.

Años *(X)*	1950 (0)	1960 (1)	1970 (2)	1980 (3)	1990 (4)	1995 (4.5)	2000 (5)	2006 (5.6)
Población en millones *(Y)*	25.8	34.9	48.2	66.8	81.2	93.0	97.4	103.1



Detemina una recta de regresión, y apartir de ella, estima la población en el 2020.

Ejemplo (Cont.)

Solución

En el manejo de series de tiempo, la variable X por lo general tendrá valores grandes. Para evitar trabajar con estos valores y lograr que la ecuación tenga valores pequeños, se acostumbra asignar al primer año (en este caso 1950) el valor $\boldsymbol{x} = 0$, al año siguiente (1960) el valor $\boldsymbol{x} = 1$, y así sucesivamente hasta el año 2006 que le corresponde $\boldsymbol{x} = 5.6$.

	Años (x)	x^2	**Población** (y)	y^2	xy
	0	0	25.8	665.64	0
	1	1	34.9	1218.01	34.9
	2	4	48.2	2323.24	96.4
	3	9	66.8	4462.24	200.4
	4	16	81.2	6593.44	324.8
	4.5	20.25	93.0	8649.00	418.5
	5	25	97.4	9486.76	487.0
	5.6	31.36	103.1	10629.61	577.36
Sumas	$\Sigma x = 25.1$	$\Sigma x^2 = 106.61$	$\Sigma y = 549.4$	$\Sigma y^2 = 44027.94$	$\Sigma xy = 2139.36$

La ecuación buscada es de la forma:

$$\hat{y} = ax + b$$

En donde:

$$a = \frac{SP(xy)}{SC(x)} = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} = \frac{2139.36 - \dfrac{(25.1)(549.4)}{8}}{106.61 - \dfrac{(25.1)^2}{8}} = \frac{415.62}{27.86} = 14.92$$

$$b = \frac{\Sigma y - (a \bullet \Sigma x)}{n} = \frac{549.4 - (14.92)(25.1)}{8} = \frac{174.91}{8} = 21.86$$

Sustituyendo estos valoes en $\hat{y} = ax + b$:

$$\hat{y} = 14.92x + 21.86$$

Comprobación:

La ecuación debe satisfacer las coordenas $(\bar{x}, \bar{y})$:

$68.68 = (14.92)(3.14) + 21.86$

$68.68 = 46.85 + 21.86$

$68.68 \approx 68.72$ Correcto

Para pronosticar la población en el año 2020, necesitamos determinar primero el valor de x que corresponde a ese año. Para ello, debemos tener en cuenta que cada unidad equivale a diez años. Entonces, si al año 2000 le corresponde $x = 5$, y al 2010 $x = 6$, al año 2020 le corresponderá $x = 7$.

Sustituyendo x = 7 en

$\hat{y} = 14.92x + 21.86$

$=14.92(7) + 21.86 = 104.44 + 21.86 = 126.3$

De continuar la tendencia actual, para el año 2020 se espera una población de 126.3 millones. Sin embargo, debemos ser conscientes que esta proyección está basada en un modelo lineal, y la población realmente podría seguir un modelo distinto tal vez exponencial.

Actividad 4.4 b

La siguiente tabla, muestra la variabilidad de las precipitaciones anuales en el estado de Sinaloa en el periodo de 1996 a 2007.

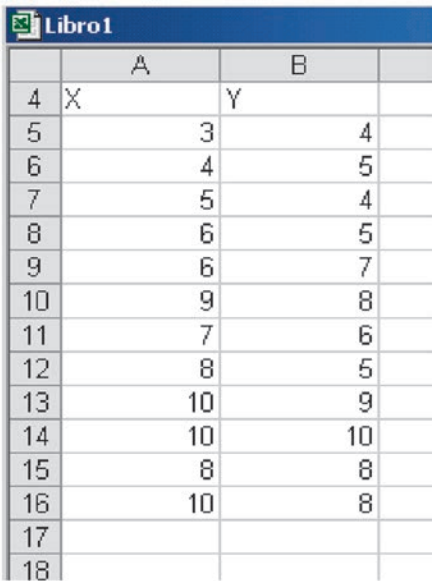
Años *(X)*	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Precipitación en mililitros de agua *(Y)*	731	718	683	571	783	684	554	634	932	522	767	616

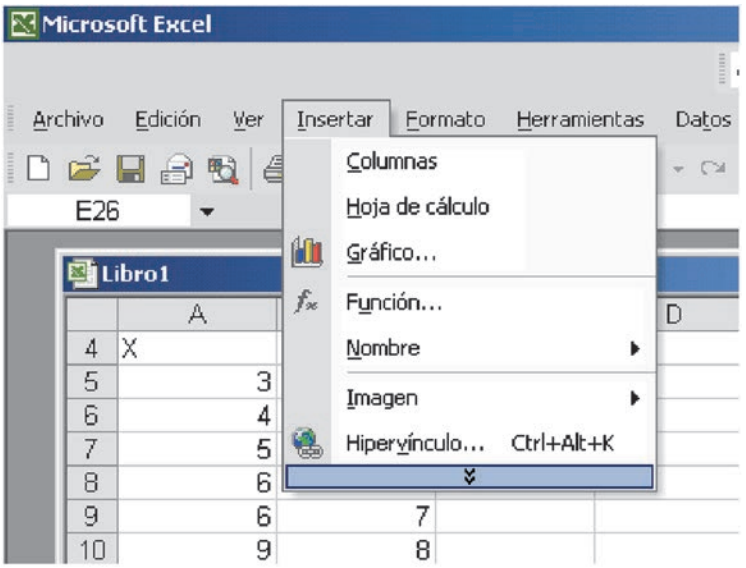
a) Dibujar la serie de tiempo.
b) Calcular la recta de regresión.
c) Estimar la precipitación para el año 2010.

Correlación y regresión con Excel

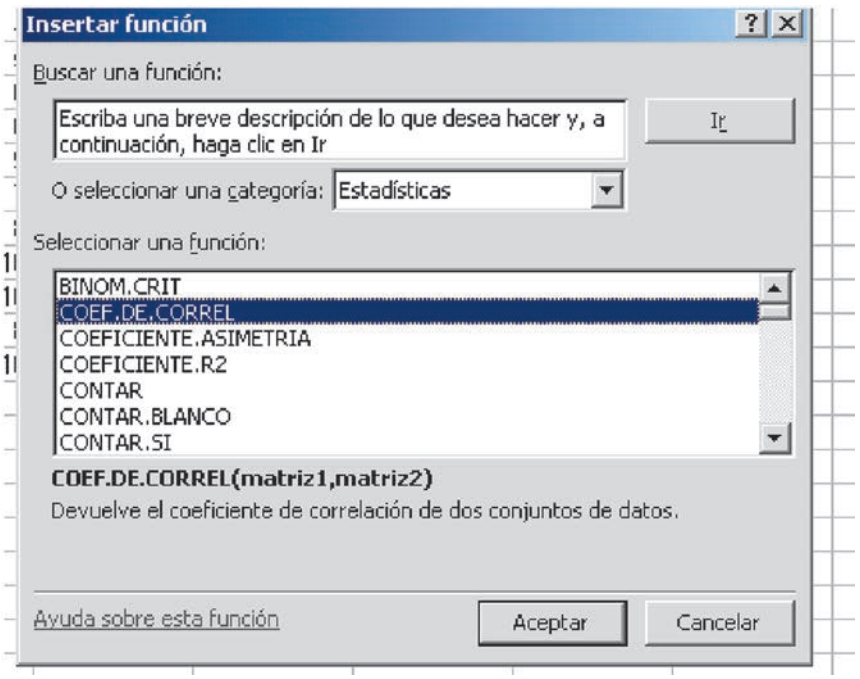
La hoja de cálculo Excel, proporciona directamente el coeficiente de correlación y la recta de regresión. El procedimiento lo explicaremos con el ejemplo de las variables calificación en matemáticas y calificación en mecánica.

Primero. Abrir el programa excel, capturar los datos, Clic en ***insertar*** y seleccionar ***Función*** f_x.

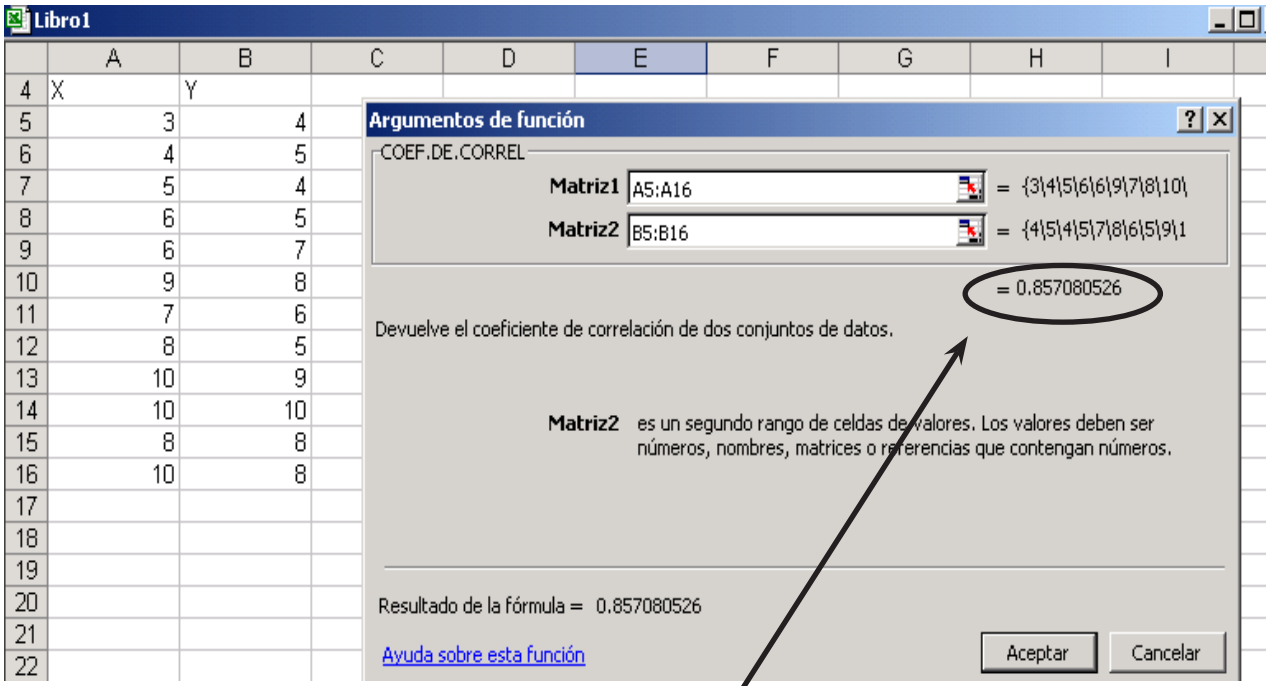




Segundo. Aparece el siguiente cuadro. Donde se indica seleccionar categoría, seleccionar ***estadísticas, y, en el cuadro inferior seleccionar COEF. DE CORREL:***

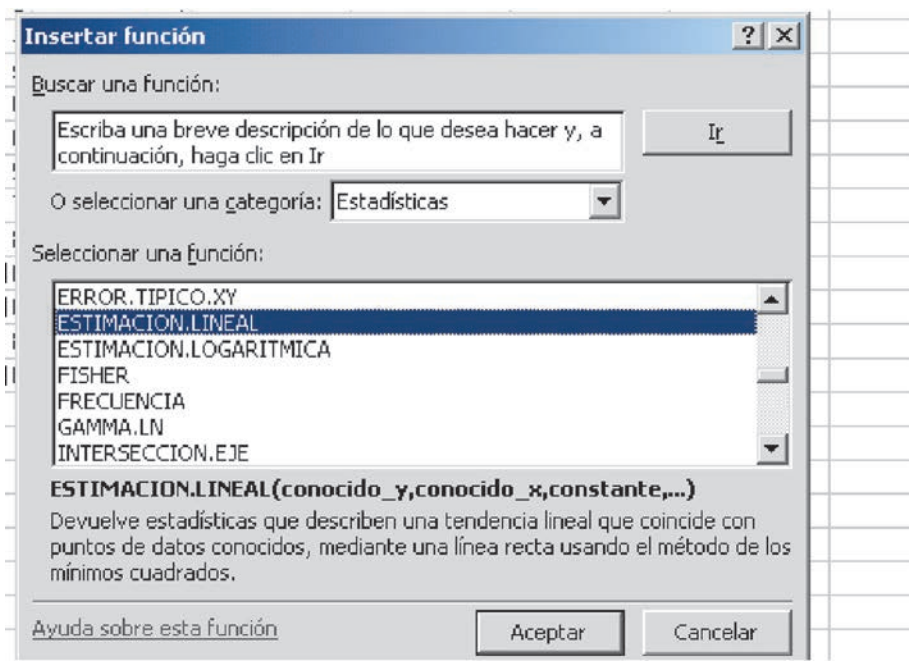


Tercero. Clic en aceptar y capturar matriz de datos.

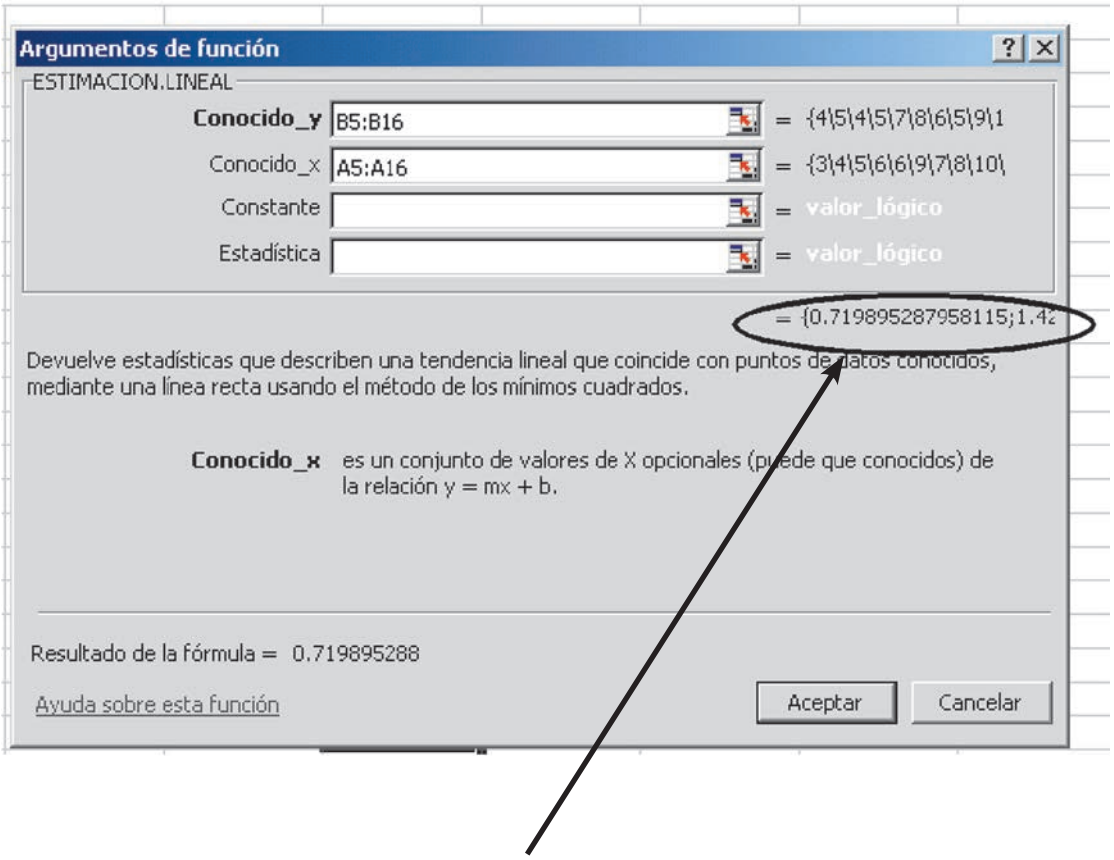


Aparece directamente el coeficiente de correlación ***r = 0.857***

Si queremos la ***recta de regresión***, en el paso dos elegimos ***ESTIMACIÓN LINEAL*** en vez de COEF. DE CORREL:



Clic en aceptar, introducir las celdas en donde están los valores *y*, después los valoress *x*, y automáticamente aparecen los valores de la pendiente ***(a)*** y de la ordenada en el origen ***(b).***



Valor de la pendiente: ***a*** = *0.7198*

Valor de la ordenada en el origen: ***b*** = *1.42*

Por lo tanto, la ecuación de la recta de regresión es: ***y = 0.72x + 1.42***

Ejercicio 4.4

1. La tabla siguiente muestra los resultados obtenidos por los equipos de futbol de primera división en el torneo de clausura 2008.

Clasificación (C)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Partidos ganados (G)	11	10	8	7	7	6	5	6	5	5	5	6	5	3	4	2	3	2
Partidos empatados (E)	3	6	4	5	5	7	8	5	7	7	6	6	6	8	5	8	5	7
Partidos perdidos (P)	3	1	5	5	5	4	4	6	5	5	6	5	6	6	8	7	9	8

Determinar la recta de regresión de la variable G en función de la variable C.

2. La tabla siguiente muestra los gastos en publicidad y las correspondientes ventas, de dos empresas. Estudia la correlación entre los gastos en publicidad y ventas de cada una.

	Gastos en Publicidadad (en millones de pesos)	1	2	3	4	5	6
Empresa A	Ventas (en millones de pesos)	10	17	30	28	39	47
Empresa B	Ventas (en millones de pesos)	10	12	19	22	25	30

3. Con los datos del ejercicio 2, obtener la recta de regresión de las ventas obtenidas en función del gasto en publicidad para cada una de las empresas.

4. La tabla siguiente muestra la producción nacional de papa (en miles de toneladas) durante el período 1994-2005.

Años *(X)*	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Producción (miles de toneladas) *(Y)*	1170	1269	1287	1320	1284	1490	1629	1635	1483	1664	1509	1653

Determinar la recta de regresión, y a partir de ella estimar la producción de papa en 2010.

AUTOEVALUACIÓN (UNIDAD IV)

1. Deberás comprender los siguientes términos:

Datos bivariados	El *CRCC*
Relación funcional	El Coeficiente de Correlación de Pearson
Relación estadística	Covarianza
Distribución bidimensional	Análisis de Regresión
Gráfico de dispersión	Métodos de los mínimos cuadrados
Análisis de correlación	Series de Tiempo.
Signos de correlación	

2. Responde «verdadero» si la afirmación es siempre cierta. Si la afirmación no siempre es verdad, reemplaza las palabras en negritas con las palabras que hagan que la afirmación sea siempre vredadera.
   a. El coeficiente de correlación lineal se usa para determinar la **ecuación que representa** la relación enre dos variables.
   b. La **pendiente** de la recta de regresión representa la cantidad de cambio que se espera ocurra en *y* cuadnom *x* crece en una unidad.
   c. Cuando el valor calculado de *r* es positivo, el valor calculado de la pendiente es **negativo.**
   d. Los coeficientes de correlación varían entre **0 y +1**.
   e. La recta del mejor ajuste se usa para predecir el **valor promedio** de *y* que puede esperarse que ocurra en un valor dado de *x*.

3. A partir de la siguiente tabla ampliada encuentra lo siguiente:

x	y	x^2	xy	y^2
2	6	4	12	36
3	5	9	15	25
3	7	9	21	49
4	7	16	28	49
5	7	25	35	49
5	9	25	45	81
6	8	36	48	64
28	49	124	204	353

   a. $SC(x)$
   b. $SC(y)$
   c. El coeficiente de correlación lineal, *r.*
   e. La pendiente.
   f. La ordenada en el origen.
   g. La ecuación de la recta del mejor ajuste.

Bibliografía

Robert Johnson, Patricia Kuby. *Estadística Elemental,* tercera edición. Thomson. México 2004.

John E. Freund. *Estadística elemental,* octava edición. Pearson. México. 1992.

American Statistical Association. A Curriculum Framework for Prek-12 Statistics Education.

http: //education.uncc.edu/droyster/PMET/GAISE/GAISE%20PreK-12.pdf

José Alfredo Juárez, Armando Flórez, Arturo Ylé, José Alberto Alvarado. *Estadística y probabilidad.* DGEP-UAS. 2002.

ALEA, Grupo portugués que promueve la enseñanza de la estadística.

http: //alea-estp.ine.pt/html/nocoes/html.

Sharon L. Lohr. *Muestreo: Diseño y Análisis*. Thomson. México. 2000..

Des Raj. *Teoría del muestreo.* Fondo de cultura económica. México. 1980.

Miguel de Guzmán, José Colera y Adela Salvador. *Matemáticas III*. Ediciones Anaya. Madrid 1988.

Carmen Batanero. *Didáctica de la estadística.* Departamento de Didáctica de la Estadística. Universidad de Granada España.

Guillermo Pastor. *Estadística básica.* Trillas. Edición Conalep. México. 1998.

Lincoyán Portus Govinden. *Curso práctico de estadística,* seguna edición. Mc Graww Hill. Bogotá, Colombia. 1999.

Ma. José Asencio, José A. Romero y Estrella de Vicente. *Estadística.* Mc Graw Hill. España. 1999.

Haroldo Elorza. *Estadística para las ciencias sociales y del comportamiento.* Segunda edición. Oxford. México, 2001.

Paulo Afonso Lopes. *Probabilidad y Estadística.* Prentice Hall. Colombia. 2000.

Fred *aprende estadística básica*. Trillas, México. 1979.

Allan J. Rossman, Beth L. Chance. Workshop Statistics, *Discovery with Data and Fathom.* Key Curriculum Press. United States of America. 2001.

Gudmund R. Iversen, Mary Gergen, Statistics, *The Conceptual Approach.* Springer. New York. 1997.

Christopher J. Wid, George A. F. Seber. *Chance Encounters, A first Course in Data Analyssis and Infernece*. John Wiley & Sons, Inc. United States of America. 1999.

Dennis Hurley Phee. *Probabilidad y estadística 4.* CECSA. México. 1980.

Gildaberto Bonilla. *Métodos prácticos de inferencia estadística.* Trillas. México. 1991.

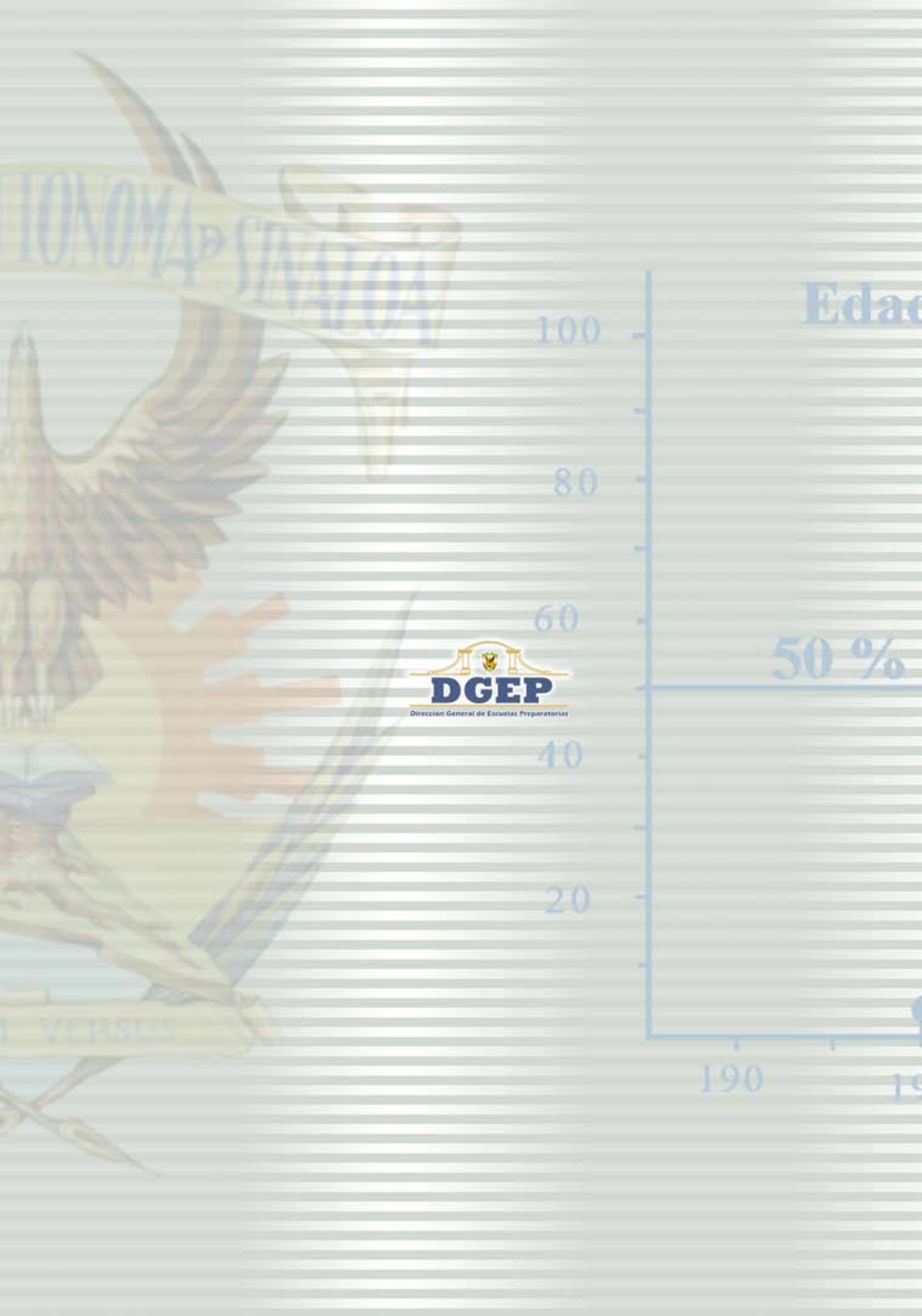
Estadística

exploración de datos

de José Alfredo Juárez Duarte, Arturo Ylé Martínez, Armando Flórez Arco y Santiago Inzunsa Cázarez

Se terminó de imprimir en el mes de agosto de 2012
en los talleres gráficos de Servicios Editoriales Once Ríos,
calle Río Usumacinta 821 , Col. Industrial Bravo.
Tel. 01(667)712-2950. Culiacán, Sin.

Esta obra consta de 9 000 ejemplares.


Edad
100
80
60
50 %
40
20
190