Unidad 8. Estadística descriptiva

Competencia de la unidad

Analizar series de datos de fenómenos de la realidad, aplicando conceptos y definiciones sobre estadística descriptiva, para tomar decisiones adecuadas en los momentos oportunos.

Relación y desarrollo

Tercer ciclo	Noveno grado	Primer año de bachillerato
Unidad 7: Gráfica de faja y circular (7°) • Gráfica de faja • Gráfica circular	**Unidad 8: Medidas de dispersión** • Dispersión • Propiedades de la desviación típica	**Unidad 8: Estadística descriptiva** • Muestreo, estadísticos y parámetros • Medidas de posición • Práctica en GeoGebra
Unidad 8: Organización y análisis de datos estadísticos (8°) • Tablas y gráficas estadísticas para variables cuantitativas • Medidas de tendencia central • Valor aproximado y dígitos significativos		

Plan de estudio de la unidad

Lección	Horas	Clases
1. Muestreo, estadísticos y parámetros	1	1. Definiciones previas
	1	2. Actividad introductoria de muestreo
	1	3. Muestreo probabilístico
	1	4. Muestreo no probabilístico
	1	5. Repaso de tablas de frecuencia
	1	6. Medidas de tendencia central
	1	7. Medidas de dispersión
	1	8. Coeficiente de variación
	1	9. Practica lo aprendido
2. Medidas de posición	1	1. Cuartiles
	1	2. Diagrama de caja y bigotes
	1	3. Análisis del diagrama de caja y bigotes
	1	4. Deciles y percentiles
	1	5. Practica lo aprendido
	2	6. Problemas de la unidad
3. Práctica en GeoGebra	1	1. Análisis estadístico
	1	Prueba de la unidad 8
	2	Prueba del cuarto periodo

18 horas clase + prueba de la unidad 8 + prueba del cuarto periodo

Puntos esenciales de cada lección

Lección 1: Muestreo, estadísticos y parámetros. Se introduce el concepto de muestreo y se repasarán los contenidos sobre medidas de tendencia central y medidas de dispersión.

Lección 2: Medidas de posición. Se dará énfasis al estudio de los cuartiles y los diagramas de caja y bigotes.

Lección 3: Práctica en GeoGebra. Se estudiará la herramienta de análisis estadístico que provee GeoGebra.

1.1 Definiciones previas

Problema inicial

A la “Feria del Libro” asistieron 1 000 personas. Por medio de una encuesta se entrevista al 15% de la población y se les pregunta acerca de: sexo, edad, género literario preferido, presupuesto para la compra de libros. Responde:

a) ¿Cuántas personas asistieron a la Feria del Libro?
b) ¿Cuántas personas fueron entrevistadas?
c) ¿Qué se les preguntó a las personas entrevistadas?

Solución

a) Asistieron un total de 1 000 personas.

b) Se entrevistó el 15% de 1 000, es decir: $1\,000 \times \frac{15}{100} = 150$.

Por lo tanto, se entrevistaron 150 personas.

c) Se les preguntó acerca del sexo, edad, género literario preferido y presupuesto para comprar libros.

Definiciones

- Se define **población** como un conjunto total de individuos, objetos o eventos que tienen las mismas características y sobre el que se está interesado en obtener conclusiones.

- Si se toma a una parte de la población con el propósito de ser estudiada, a este grupo seleccionado se le llama **muestra**.

- Al tipo de información que se desea investigar se le llama **variable estadística**. Una variable estadística es una propiedad que es susceptible a tomar diferentes valores y que pueden medirse u observarse.

Las variables estadísticas pueden clasificarse tal como muestra el esquema.

- Variable estadística
  - Cualitativa
    - Nominal
    - Ordinal
  - Cuantitativa
    - Discreta
    - Continua

- Las **variables cualitativas** expresan distintas cualidades, características o modalidades.
  - Las variables cualitativas nominales no pueden definirse mediante un orden. Por ejemplo: país, idioma, estado civil, sexo.
  - Las variables cualitativas ordinales pueden tomar distintos valores ordenados siguiendo una escala establecida. Por ejemplo: malo, regular, bueno.

Una variable cualitativa es dicotómica si toma solo dos valores. Por ejemplo: sí y no, hombre y mujer; o politómica si toma tres o más valores.

- Las **variables cuantitativas** toman valores numéricos.
  - Las variables cuantitativas discretas toman valores numéricos enteros no negativos. Por ejemplo: número de hijos.
  - Las variables cuantitativas continuas pueden tomar cualquier valor dentro de un intervalo específico de valores. Por ejemplo: el peso, altura, gastos familiares.

Lección 1

Ejemplo

A un evento sobre “El rol de la mujer en la sociedad” asisten un total de 2 000 personas de las 21 000 que habitan el municipio de Chalatenango, y de ellas se recoge la siguiente información: sexo, estatura, cantidad de hijos y cuán a menudo cocina en la casa; cuyas opciones de respuesta son: nunca, casi nunca, a veces, casi siempre o siempre. Responde los siguientes literales:

a) Identifica la población y la muestra en este evento.

b) Identifica las variables y clasifícalas.

a) La población de la cual se obtiene la muestra son las 21 000 personas que habitan el municipio de Chalatenango.

La muestra son las 2 000 personas que asistieron al evento.

b) Las variables son: sexo, edad, cantidad de hijos y periodicidad con que cocina. Al clasificarlas se obtiene:

Variables cualitativas		Variables cuantitativas	
Nominales	• Sexo	***Discretas***	• Cantidad de hijos
Ordinales	• Periodicidad con que cocina (se puede ordenar como nunca, casi nunca, ... , siempre)	***Continuas***	• Estatura

Problemas

1. Para cada una de las siguientes situaciones, identifica la población, la muestra, las variables estadísticas y su clasificación.

a) En la Biblioteca Nacional de El Salvador se desea conocer el estado de los libros de Matemática y estos se extraen de los primeros 10 estantes para categorizarlos como bueno, malo o inservible.

b) De todos los niños en edad escolar de El Salvador, se encuesta a los que están en noveno grado para conocer si les gusta la música electrónica o la instrumental.

c) En el Hospital Nacional Rosales se desea entrevistar a los pacientes que están hospitalizados por enfermedades pulmonares y saber el trato que reciben en dicho hospital.

d) En el Parque Nacional Montecristo se desea saber los años de vida que tienen todos los árboles cipreses que hay.

e) En un cine de San Salvador se entrevista a los que asisten a una película de comedia-romance para investigar si les gusta más las de romance, comedia o ambas.

El Parque Nacional Montecristo es un parque protegido que está ubicado en el municipio de Metapán, departamento de Santa Ana y tiene una extensión de 1973 hectáreas.

2. Determina si las variables estadísticas presentadas a continuación son variables cualitativas (nominales u ordinales) o variables cuantitativas (discretas o continuas).

a) El grupo sanguíneo de una persona

b) Temperatura en grados centígrados

c) Grado de escolaridad

d) Religión

e) Lugar de nacimiento

f) Número de alumnos

g) El precio de un artículo

h) Valores de la glucosa en 50 niños

i) Número de clínicas médicas por municipio

j) El ingreso mensual de un padre de familia

k) Presión arterial

l) Intensidad del dolor

Indicador de logro:

1.1 Aplica las definiciones de población, muestra y variable, y clasifica las variables entre cualitativas nominales y ordinales o cuantitativas discretas y continuas.

Secuencia:

En tercer ciclo se ha trabajado en el bloque de estadística desde séptimo grado, abordando los tipos de gráficas, luego las medidas de tendencia central (para poblaciones) y luego las medidas de dispersión, en esta unidad se trabajarán los mismos conceptos, pero dividiendo entre muestra y población.

Propósito:

En esta clase se pretende introducir los conceptos básicos sobre población, muestra, variables y tipos de variables, a partir del Problema inicial se puede introducir la parte de población y muestra, y en el Ejemplo se presenta la forma de clasificar un conjunto de variables según su tipo.

Solución de problemas:

1a) Población: los libros de matemática de la Biblioteca Nacional de El Salvador; muestra: los primeros 10 estantes; variable: estado de los libros de Matemática; tipo: cualitativa ordinal.

1b) Población: los niños en edad escolar de El Salvador; muestra: los niños de noveno grado; variable: música preferida; tipo: cualitativa nominal.

1c) Población: todos los pacientes hospitalizados; muestra: los pacientes hospitalizados por enfermedades pulmonares; variable: etapa de la enfermedad; tipo: cualitativa ordinal.

1d) Población: todos los árboles del Parque Nacional Montecristo; muestra: los árboles de ciprés; variable: años de vida; tipo: cuantitativa continua.

Para este problema, los años de vida pueden considerarse discretos, ya que es habitual mencionar años exactos, sin embargo, en la respuesta se sugiere que es una variable continua, puesto que el tiempo es continuo.

1e) Población: las personas que asisten a ver películas de comedia-romance; muestra: las personas que asistieron a ver una película de comedia-romance en el cine de San Salvador; variable: género de la película que prefiere; tipo: cualitativa nominal.

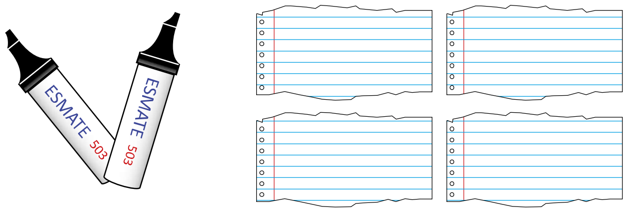
2a) Cualitativa nominal.
2b) Cuantitativa continua.
2c) Cualitativa ordinal.
2d) Cualitativa nominal.
2e) Cualitativa nominal.
2f) Cuantitativa discreta.
2g) Cuantitativa continua.
2h) Cuantitativa continua.
2i) Cuantitativa discreta.
2j) Cuantitativa continua.
2k) Cuantitativa continua.
2l) Cualitativa ordinal.

Lección 1

1.2 Actividad introductoria

Materiales

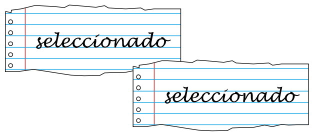
- Pedacitos de papel (uno por cada estudiante)
- Plumón



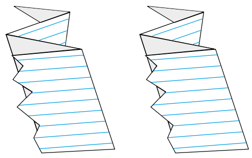
Actividad 1

Seleccionar 5 estudiantes del salón de clase utilizando el siguiente proceso:

1. Escribir en 5 de los pedacitos de papel la palabra "seleccionado".



2. Doblar todos los papelitos y entregar uno a cada estudiante.



3. Los estudiantes seleccionados son aquellos a quienes les apareció un papelito con la palabra "seleccionado".

Actividad 2

Seleccionar 5 estudiantes del salón de clase utilizando el siguiente proceso:

1. Numerar los estudiantes del 1 a N, siendo N la cantidad total de estudiantes que hay en el salón.
2. Se escoge un estudiante al azar del 1 al mayor entero m menor o igual que $\frac{N}{5}$ (se pueden usar papelitos, o el aleatorio de la calculadora , etc.)
3. El segundo seleccionado es el que tiene el número más cercano a la suma del anterior y m.
4. El tercer seleccionado es el que tiene el número del segundo sumándole m. Y así sucesivamente hasta seleccionar los 5 estudiantes, por ejemplo si el primer estudiante tiene el número a, los 5 estudiantes seleccionados serían los que tienen los número: a, $a + m$, $a + 2m$, $a + 3m$, $a + 4m$.

Por ejemplo, si $N = 40$ y $a = 3$, entonces $m = \frac{40}{5} = 8$ y los números seleccionados son: 3, 11, 19, 37, 35.

Definiciones

A la acción de seleccionar de una población de N elementos una muestra de n elementos se conoce como **muestreo**.

El tipo de muestreo en el que todos los elementos de la población tienen igual probabilidad de ser seleccionados (como en la actividad 1) se conoce como **muestreo aleatorio simple**.

Al tipo de muestreo aleatorio en el que se lista la población, y se escoge un número aleatorio menor o igual que $\frac{N}{n}$ donde N es el total de la población y n es el total de la muestra, y se seleccionan los demás sumándole al anterior $\frac{N}{n}$ se conoce como **muestreo aleatorio sistemático**.

Problemas

1. Escribe 2 formas de muestreo aleatorio simple.

2. Escribe 2 formas de muestreo aleatorio sistemático.

3. ¿Al realizar una rifa se está realizando un muestreo?

Indicador de logro:

1.2 Planifica y aplica técnicas de muestreo aleatorio simple y sistemático.

Secuencia:

Partiendo del marco teórico definido en la clase anterior, ahora es posible determinar diferentes métodos para realizar muestras de una población.

Propósito:

Se espera que en esta actividad los estudiantes se motiven a plantear e implementar procedimientos para realizar muestreos aleatorios simples y estratificados.

Solución de problemas:

1. En este problema los estudiantes pueden mencionar varias formas, algunas pueden ser:

- Repartir números, y luego a partir de otros números doblados, extraer de una bolsa la cantidad que se requiera para la muestra, la muestra será las personas cuyos números corresponden con los números extraídos.

- Que una persona desconocida seleccione la cantidad de estudiantes que se necesita.

> De manera opcional, el docente puede mencionar que la calculadora tiene una función "aleatoria" que se puede encontrar como "Ran#" en algunas calculadoras y devuelve un número decimal aleatorio de 3 decimales.

2. En este problema los estudiantes pueden mencionar varias formas, algunas pueden ser:

- Numerar a toda la población, luego escoger un número al azar entre 1 y 10, por ejemplo 3, y luego extraer de la población todos los múltiplos de 3.

- Estratificar la población en grupos cuya cantidad de individuos coincida con la muestra requerida, y luego se seleccione al azar uno de todos los grupos.

3. Puesto que la idea de una rifa es entregar algún tipo de obsequio a un grupo de personas determinado, es claro que una rifa tiene sentido siempre y cuando la cantidad de personas a quienes se entregará el obsequio es menor a la cantidad de personas que asisten a la rifa; es por ello que una rifa equivale a seleccionar una parte de la población que asistió, sin embargo, el objetivo no es estudiar alguna característica o realizar algún estudio para establecer conclusiones acerca de algún problema, por lo tanto, una rifa no es un muestreo, por otro lado, un muestreo si se puede realizar siguiendo un procedimiento muy parecido al de una rifa.

Lección 1

1.3 Muestreo probabilístico*

Problema inicial

En un instituto se cuenta con la siguiente información de los estudiantes de bachillerato:

	Niñas	Niños
Primer año	24	12
Segundo año	9	15

Determina una forma para seleccionar una muestra de 20 estudiantes que cursen bachillerato (primero o segundo año, niñas o niños).

Solución

En este problema se tiene 4 tipos de personas para la población, se puede calcular el porcentaje de la población que le corresponde a cada uno:

Niñas de Primer año: $\frac{24}{60} \times 100 = 40\%$

Niñas de Segundo año: $\frac{9}{60} \times 100 = 15\%$

Niños de Primer año: $\frac{12}{60} \times 100 = 20\%$

Niños de Segundo año: $\frac{15}{60} \times 100 = 25\%$

Entonces para obtener la muestra de 20 estudiantes se pueden utilizar estos porcentajes:

Niñas de Primer año: 40% de 20 es 8

Niñas de Segundo año: 15% de 20 es 3

Niños de Primer año: 20% de 20 es 4

Niños de Segundo año: 25% de 20 es 5

Por lo tanto se puede extraer una muestra de 20 estudiantes de bachillerato, seleccionando 8 niñas de primer año, 4 niños de primer año, 3 niñas de segundo año y 5 niños de segundo año.

En general

El muestreo que se aplica proporcionalmente a una población que está dividida en sectores (estratos) se conoce como **muestreo estratificado**.

El muestreo que se aplica a una población donde es necesario dividir grupos y seleccionar aleatoriamente algunos de ellos se conoce como **muestreo por conglomerado**.

Se define el **muestreo probabilístico** como todos aquellos métodos de muestreo donde todos los elementos de la población tienen las mismas posibilidades de ser seleccionados para la muestra.

El muestreo aleatorio simple, el sistemático, el estratificado y el muestreo por conglomerado son todos muestreos probabilísticos.

Ejemplo

Utilizando muestreo por conglomerado determina la forma de realizar un estudio acerca de la comida preferida de las personas en el departamento de Santa Ana.

Se puede dividir la población en conglomerados: estudiantes, empleados de oficina, deportistas, profesionales independientes, etc. Y se seleccionan al azar 2 o 3 de estos conglomerados para obtener la muestra.

Problemas

Realiza un muestreo estratificado de 40 personas de una colonia que tiene los siguientes datos:

	Femenino	Masculino
Menor de edad	70	40
Mayor de edad	30	60

Indicador de logro:

1.3 Aplica el muestreo probabilístico a situaciones de la vida cotidiana.

Secuencia:

En esta clase se concluye con las técnicas de muestreo probabilístico, abordando el muestreo estratificado y el muestreo por conglomerados.

Posibles dificultades:

Para realizar el muestreo estratificado, es necesario que los estudiantes apliquen el tanto por ciento, para determinar la cantidad de personas que es necesario seleccionar de cada estrato.

Solución de problemas:

Puesto que en total la colonia tiene 200 personas, luego calculado los porcentajes para cada estrato:

Mujeres menores de edad: $\frac{70}{200} \times 100 = 35\%$

Mujeres mayores de edad: $\frac{30}{200} \times 100 = 15\%$

Hombres menores de edad: $\frac{40}{200} \times 100 = 20\%$

Hombres mayores de edad: $\frac{60}{200} \times 100 = 30\%$

Entonces para obtener la muestra de 40 personas se pueden utilizar estos porcentajes:

Mujeres menores de edad: 35% de 40 es 14

Mujeres mayores de edad: 15% de 40 es 6

Hombres menores de edad: 20% de 40 es 8

Hombres mayores de edad: 30% de 40 es 12

Por lo tanto se puede extraer una muestra de 40 personas de toda la colonia, seleccionando 14 mujeres menores de edad, 6 mujeres mayores de edad, 8 hombres menores de edad y 12 hombres mayores de edad.

1.4 Muestreo no probabilístico

Definición

En un muestreo cuando la selección depende de las características de los individuos que se desean estudiar, se dice que es un **muestreo no probabilístico**. Algunas técnicas de muestreo no probabilístico son:

- **Muestreo por conveniencia**: la selección de la muestra se hace basada en la facilidad o las características específicas del estudio.
- **Muestreo por bola de nieve**: consiste en seleccionar la muestra por medio de conocidos, es decir, se hace el estudio a personas conocidas, para que estas lo hagan a personas conocidas de ellas y así hasta llegar a la muestra que se desea. Se utiliza cuando es difícil encontrar la muestra.
- **Muestreo por cuotas**: se divide la población por grupos y se establece una cantidad de individuos de muestra por cada grupo, la cantidad de individuos por grupo se realiza de manera apreciativa.
- **Muestreo discrecional**: el muestreo se hace considerando las características y formación específicas de las personas.

Ejemplo 1

En una investigación se desea realizar una encuesta de hábitos alimenticios, y las personas con las que se tiene mayor facilidad de contacto son los miembros del equipo de baloncesto de una colonia.

En este caso se puede realizar un muestreo por conveniencia, pero puede que la muestra no refleje la tendencia de la población.

Ejemplo 2

En una investigación se desea saber sobre las costumbres y estructura de los grupos delincuenciales en El Salvador, para ello se realiza una entrevista a personas cercanas y luego estas personas la hacen a personas cercanas a ellas, hasta llegar a personas cercanas a estos grupos.

En este caso se puede realizar un muestreo por bola de nieve. Esta técnica se puede escoger para seleccionar personas que pueden no brindar información a una persona particular y es mejor selecionarla a partir de personas que los conozcan, se utiliza para estudiar temas de delincuencia, política, corrupción, etc.

Ejemplo 3

Se realiza una encuesta a 100 universitarios y 100 profesionales.

En este caso se puede realizar un muestreo por cuotas. Es una técnica parecida al muestreo por estratos, pero la asignación de la cantidad de personas por estrato no se calcula a partir de la población.

Ejemplo 4

En un salón de clase se desea seleccionar 4 estudiantes para participar en una olimpíada de matemática.

En este caso se puede realizar un muestreo discrecional, se pueden seleccionar los 4 estudiantes que tienen mejor rendimiento en matemática.

Problemas

Determina qué tipo de muestreo no probabilístico consideras más adecuado explicando las ventajas y desventajas de los tipos de muestreo para cada situación.

a) Materia favorita de un estudiante
b) Tipos de estampillas que tienen los filatelistas
c) Seguridad en cada departamento de El Salvador
d) Escoger 5 estudiantes para competir en natación

Indicador de logro:

1.4 Identifica la técnica de muestreo no probabilístico más adecuada para situaciones específicas.

Secuencia:

Luego que en las clases anteriores se abordara lo correspondiente al muestreo probabilístico, también se dedicará una clase para establecer las técnicas de muestreo no probabilístico.

Propósito:

Este tipo de técnicas normalmente se utilizan en los estudios sociales (y otras áreas humanísticas), en donde es necesario buscar estrategias diversas para estudiar los fenómenos de dichas áreas.

Solución de problemas:

a) Puesto que es una variable apreciativa de cada estudiante, en este caso podría resultar natural recolectar información del centro escolar con mayor accesibilidad, por lo cual entraría en un muestreo por conveniencia.

b) Los filatelistas (como cualquier tipo de coleccionista), suelen ser muy reservados con sus pertenencias, por lo cual se volvería difícil que se pueda investigar esto de manera personal, y es muy probable que sea necesario aplicar la técnica de la bola de nieve, para lograr captar información de diferentes filatelistas.

c) Para obtener el punto de vista de diferentes partes de la población, podría ser útil realizar un muestreo por cuotas, en el cual los grupos que se establecen pertenecerían a departamentos diferentes.

d) Puesto que las personas seleccionadas serán sometidos a una prueba sobre sus capacidades en natación, resulta conveniente un muestreo discrecional, en el cual se seleccionen las personas cuyas características favorezcan sus resultados.

Lección 1

1.5 Repaso de tablas de frecuencia

Problema inicial

En una comunidad del área metropolitana de San Salvador se pregunta la edad a jóvenes menores de 21 años, obteniendo la siguiente información:

a) Organiza la información en una tabla de distribución de frecuencias en 4 clases de 3 en 3, iniciando en 9 y terminando en 21.
b) Elabora una tabla de distribución de frecuencias y calcula la media aritmética (μ), la moda y la mediana para estos datos agrupados.
c) Calcula la varianza (σ^2) y la desviación típica (o desviación estándar σ).

Edades					
9	14	15	14	19	16
11	18	9	12	20	12
12	11	10	19	14	13
15	12	11	18	11	16
14	16	17	12	13	17

Solución

a)

Edades	Cantidad de jóvenes (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \mu$	$(Pm - \mu)^2$	$f(Pm - \mu)^2$
9 a 12	7	10.5	73.5	–4	16	112
12 a 15	11	13.5	148.5	–1	1	11
15 a 18	7	16.5	115.5	2	4	28
18 a 21	5	19.5	97.5	5	25	125
TOTAL	**30**					

Solución de a)

En cada clase se cumple que los datos contados en la frecuencia son mayores o iguales al límite inferior y menores al límite superior, excepto en la última clase, donde es menor o igual al límite superior.

b) Para calcular la media aritmética se suman los valores de la columna $f \times Pm$ y se divide por el total de datos.

$$\mu = \frac{\sum f \times Pm}{n} = \frac{73.5 + 148.5 + 115.5 + 97.5}{30} = 14.5$$

La clase que contiene la mediana es la segunda (12 a 15) porque en ella se encuentran los datos 15 y 16.

$$\text{Mediana} = \frac{12 + 15}{2} = 13.5$$

La clase con la mayor frecuencia es la segunda (12 a 15).

$$\text{Moda} = \frac{12 + 15}{2} = 13.5$$

c) Para la varianza se suman los valores de la columna $f(Pm - \mu)^2$ y se divide por el total de datos.

$$\sigma^2 = \frac{\sum f(Pm - \mu)^2}{n} = \frac{112 + 11 + 28 + 125}{30} = 9.2$$

$$\sigma = \sqrt{\frac{\sum f(Pm - \mu)^2}{n}} = \sqrt{9.2} \approx 3.03$$

Problemas

Las velocidades que registra un policía de tránsito en la carretera de Los Chorros están en la tabla de la derecha, realiza lo siguiente:

a) Organiza la información en una tabla de distribución de frecuencias en 4 clases de 20 en 20, iniciando en 40 y terminando en 120.
b) Calcula la media aritmética, la moda y la mediana.
c) Calcula la varianza y la desviación típica.

Velocidad en km/h						
60	65	40	80	80	90	45
70	100	70	50	80	55	118
75	65	90	85	70	100	55
110	70	95	70	80	115	100

Indicador de logro:

1.5 Elabora tablas de frecuencia y calcula las medidas de tendencia central y de dispersión para datos agrupados.

Secuencia:

Una vez establecidos los conceptos sobre muestreo y algunas técnicas probabilísticas y no probabilísticas, se hará un repaso sobre las tablas de frecuencia, además de la forma para estimar las medidas de tendencia central y de dispersión.

Propósito:

En esta clase se puede abordar el Problema inicial como un ejemplo (para recordar) y luego trabajar con la parte de Problemas.

Solución de problemas:

a)

Velocidad en km/h	Frecuencia (f)
40 a 60	5
60 a 80	9
80 a 100	8
100 a 120	6
TOTAL	**28**

b)

Velocidad en km/h	Frecuencia (f)	Punto medio (Pm)	$f \times Pm$
40 a 60	5	50	250
60 a 80	9	70	630
80 a 100	8	90	720
100 a 120	6	110	660
TOTAL	**28**		

Sugerir a los estudiantes aproximar hasta las decimas cuando complete los datos en las tablas para el cálculo de la media y la desviación típica.

$$\mu = \frac{\sum f \times Pm}{n} = \frac{250 + 630 + 720 + 660}{28} \approx 80.7 \text{ (km/h)}$$

Puesto que la mediana está entre el dato 14 y 15, y estos caen en clases diferentes, se puede considerar que Mediana = 80.

La clase con la mayor frecuencia es la segunda (60 a 80), entonces, Moda = $\frac{60 + 80}{2} = 70$.

c)

Velocidad en km/h	Frecuencia (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \mu$	$(Pm - \mu)^2$	$f(Pm - \mu)^2$
40 a 60	5	50	250	–30.7	942.5	4 712.5
60 a 80	9	70	630	–10.7	114.5	1 030.5
80 a 100	8	90	720	9.3	86.5	692
100 a 120	6	110	660	29.3	858.5	5 151
TOTAL	**28**					

$$\sigma^2 = \frac{\sum f(Pm - \mu)^2}{n} \approx \frac{4\,712.5 + 1\,030.5 + 692 + 5\,151}{28} \approx 413.8$$

$$\sigma = \sqrt{\frac{\sum f(Pm - \mu)^2}{n}} \approx \sqrt{413.8} \approx 20.3 \text{ (km/h)}$$

1.6 Medidas de tendencia central

Problema inicial

De una empresa se tiene el registro de ventas del último mes en sus 30 sucursales y un registro de 20 de sus 30 sucursales, las cuales se muestran a continuación:

Ventas	Cantidad de sucursales (f)
De \$1,000 a \$2,000	5
De \$2,000 a \$3,000	11
De \$3,000 a \$4,000	8
De \$4,000 a \$5,000	6
TOTAL	**30**

Ventas	Cantidad de sucursales (f)
De \$1,000 a \$2,000	3
De \$2, 000 a \$3,000	7
De \$3,000 a \$4,000	6
De \$4,000 a \$5,000	4
TOTAL	**20**

a) Determina la media aritmética, mediana y moda para las 30 sucursales.

b) Determina la media aritmética, mediana y moda para las 20 sucursales.

Solución

a)

Ventas	Cantidad de sucursales (f)	Punto medio (Pm)	$f \times Pm$
De \$1,000 a \$2,000	5	1,500	7,500
De \$2,000 a \$3,000	11	2,500	27,500
De \$3,000 a \$4,000	8	3,500	28,000
De \$4,000 a \$5,000	6	4,500	27,000
TOTAL	**30**		90,000

$$\text{Media} = \frac{90{,}000}{30} = 3{,}000$$

$$\text{Mediana} = \frac{2{,}000 + 3{,}000}{2} = 2{,}500$$

$$\text{Moda} = \frac{2{,}000 + 3{,}000}{2} = 2{,}500$$

b)

Ventas	Cantidad de sucursales (f)	Punto medio (Pm)	$f \times Pm$
De \$1,000 a \$2,000	3	1,500	4,500
De \$2,000 a \$3,000	7	2,500	17,500
De \$3,000 a \$4,000	6	3,500	21,000
De \$4,000 a \$5,000	4	4,500	18,000
TOTAL	**20**		61,000

$$\text{Media} = \frac{61{,}000}{20} = 3{,}050$$

$$\text{Mediana} = 3{,}000$$

$$\text{Moda} = \frac{2{,}000 + 3{,}000}{2} = 2{,}500$$

Definición

Las medidas de tendencia central referentes a una población (tal como las 30 sucursales) se conocen como **parámetros** de la población y a menudo se denotan:

Media Poblacional = μ Mediana Poblacional = Me Moda Poblacional = Mo

Las medidas de tendencia central referentes a una muestra (tal como las 20 sucursales) se conocen como **estadísticos (o estadígrafos)** y se denotan:

Media Muestral = $\bar{x}$ Mediana Muestral = $\tilde{x}$ Moda Muestral = $\hat{x}$

Problemas

Considerando la población del salón de clase, recopila la información sobre el tiempo que se tardan tus compañeros en llegar desde la casa a la escuela, luego realiza una muestra aleatoria del 60% de la población y calcula todas las medidas de tendencia central, tanto para la población como para la muestra.

Indicador de logro:

1.6 Calcula las medidas de tendencia central para una muestra y una población en datos agrupados.

Secuencia:

Ahora que se ha recordado la manera de ordenar datos en una tabla de distribución de frecuencias, se pretende diferenciar entre las medidas de tendencia central de una población y de una muestra.

Propósito:

En esta clase hay que dar énfasis a las diferencias de la notación entre las medidas de tendencia central de una población y de una muestra, además de hacer notar que dependiendo de la representatividad de la muestra, puede que los valores de estas medidas difieran mucho o poco.

Solución de problemas:

En este problema la intención es recoger los datos particulares de cada salón, y hacer un ejercicio de recolección, organización, cálculo, análisis e interpretación de los datos, para hacer la recolección; el docente puede considerar alguna de las siguientes opciones:

- Preguntar a cada estudiante por fila el tiempo que tarda en llegar de la casa a la escuela, y anotarlo en la pizarra.
- Pasar una página en blanco para que anoten el tiempo, y luego fotocopiarla para cada estudiante.
- Recoger la información un día antes, y preparar el material digitado.
- Que un estudiante de cada fila recoja la información de la fila correspondiente, y luego se comparta con todos los estudiantes.

Luego, con los datos recolectados, los estudiantes realizan el cálculo de las medidas de tendencia central. Y para establecer la muestra se puede utilizar cualquiera de los métodos vistos en el muestreo probabilístico, se recomienda que sea un muestreo aleatorio simple.

Finalmente se calculan las medidas de tendencia central para la muestra extraída, el docente debe verificar que se esté utilizando la notación adecuada para cada caso. Después de haber realizado el cálculo, se puede hacer el análisis comparativo entre las medidas de tendencia central de la población y la muestra.

Lección 1

1.7 Medidas de dispersión

Problema inicial

Con los datos de las ventas de las sucursales, calcula la varianza y la desviación típica, para cada tabla.

Ventas	Cantidad de sucursales (f)
De \$1,000 a \$2,000	5
De \$2,000 a \$3,000	11
De \$3,000 a \$4,000	8
De \$4,000 a \$5,000	6
TOTAL	**30**

Ventas	Cantidad de sucursales (f)
De \$1,000 a \$2,000	3
De \$2,000 a \$3,000	7
De \$3,000 a \$4,000	6
De \$4,000 a \$5,000	4
TOTAL	**20**

Solución

Para calcular la varianza y la desviación típica de una muestra no se divide por n sino por $n-1$, para que la estimación tenga menor sesgo.

Ventas	Cantidad de sucursales (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \mu$	$(Pm - \mu)^2$	$f(Pm - \mu)^2$
De \$1, 000 a \$2, 000	5	1,500	7,500	–1,500	2,250, 000	11,250,000
De \$2, 000 a \$3, 000	11	2,500	27,500	–500	250 000	2,750,000
De \$3, 000 a \$4, 000	8	3,500	28,000	500	250 000	2,000,000
De \$4, 000 a \$5, 000	6	4,500	27,000	1,500	2,250, 000	13,500,000
TOTAL	**30**					29,500,000

Varianza $= \dfrac{\sum f(Pm-\mu)^2}{30} \approx 983,333.3$ Desviación $= \sqrt{\dfrac{\sum f(Pm-\mu)^2}{30}} = \sqrt{983,333.3} \approx 991.63$

Ventas	Cantidad de sucursales (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \bar{x}$	$(Pm - \bar{x})^2$	$f(Pm - \bar{x})^2$
De \$1,000 a \$2,000	3	1,500	4,500	–1,550	2,402,500	7,207,500
De \$2,000 a \$3,000	7	2,500	17,500	–550	302,500	2,117,500
De \$3,000 a \$4,000	6	3,500	21,000	450	202,500	1,215,000
De \$4,000 a \$5,000	4	4,500	18,000	1,450	2,102,500	8,410,000
TOTAL	**20**					18,950,000

Varianza $= \dfrac{\sum f(Pm-\bar{x})^2}{19} \approx 997,368$ Desviación $= \sqrt{\dfrac{\sum f(Pm-\bar{x})^2}{19}} = \sqrt{997,368} \approx 998.68$

Conclusión

Para una población, la varianza se denota por σ^2 y la desviación típica se denota por σ. Y se calcula de la siguiente manera:

$$\sigma^2 = \frac{\sum f(Pm-\mu)^2}{N} \qquad \sigma = \sqrt{\frac{\sum f(Pm-\mu)^2}{N}}$$

Para una muestra la varianza se denota por s^2 y la desviación típica se denota por s. Y se calcula de la siguiente manera:

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

Problemas

Considerando la información recopilada en la clase anterior sobre el tiempo que se tardan tus compañeros en llegar desde la casa a la escuela, tanto en la muestra como en la población calcula la varianza y la desviación típica.

Unidad 8

Indicador de logro:

1.7 Calcula las medidas de dispersión para una muestra y una población en datos agrupados.

Secuencia:

Después de abordar las medidas de tendencia central para muestras y poblaciones, se aborda lo correspondiente a las medidas de dispersión de los datos, de igual manera tanto para muestras como para poblaciones.

Propósito:

En esta clase, al igual que en la anterior, hay que dar énfasis a las diferencias de la notación entre las medidas de dispersión de una población y de una muestra, así como que en este caso es necesario hacer un ajuste en el cálculo de la varianza y la desviación típica muestral.

Solución de problemas:

En este problema la intención es continuar estudiando los datos recopilados en la clase anterior, haciendo el análisis de la dispersión de estos, tanto para la muestra como para la población, el docente debe verificar que se esté utilizando la notación adecuada para cada caso. Después de haber realizado el cálculo, se puede hacer el análisis comparativo entre las medidas de dispersión de la población y la muestra.

El docente puede mencionar que para el caso de la muestra, se divide por $n - 1$, para evitar el sesgo de la muestra respecto de la población, o en otras palabras, para que los valores de la varianza y la desviación típica se aproximen más a los valores poblacionales de estas medidas. Si el docente considera conveniente, puede realizar el ejercicio de calcular las medidas de dispersión de la muestra con n y con $n - 1$, y luego comparar cuál es más aproximado al los valores de la población. Esta corrección sobre el sesgo es conocida como corrección de Bessel.

Lección 1

1.8 Coeficiente de variación*

Problema inicial

La siguiente tabla muestra la media y la desviación típica de la estatura de personas del sexo masculino con edades de 5 y 17 años de una población para el año 2016:

Edad	Media	Desviación típica
5 años	110.4	4.74
17 años	170.7	5.81

a) ¿Se puede comparar la magnitud de dispersión de las dos poblaciones solo con la desviación típica?
b) Para ambas poblaciones calcula el cociente:
(desviación típica) ÷ (media)
Luego compáralos.

Solución

a) Aunque el grupo de 17 años tiene mayor desviación típica, no se puede decir que este grupo tiene mayor dispersión, porque la media también es mayor.
b) Población de 5 años: 4.74 ÷ 110.4 ≈ 0.043 Población de 17 años: 5.81 ÷ 170.7 ≈ 0.034

Este valor puede utilizarse para determinar que la estatura de la población de 17 años tiene menor dispersión que la de 5 años.

Definición

Se define el **coeficiente de variación** como el porcentaje de la desviación típica s y la media aritmética $\overline{x}$ de un conjunto de datos, se denota por CV, y se calcula: $CV = \frac{s}{\overline{x}}(100)$.

El coeficiente de variación se utiliza para comparar *la magnitud* de la dispersión de los datos de diferentes poblaciones, cuando la diferencia de las medias es grande (si la diferencia entre las medias es poca, o las medias son iguales se puede utilizar la desviación típica para comparar); por lo general cuando la media es grande, la desviación típica tiende a aumentar.

El porcentaje del coeficiente de variación también se utiliza para determinar la confiabilidad de la media aritmética de un conjunto de datos, en general para determinar la confiabilidad se puede usar la siguiente tabla como parámetro:

Valor de CV	Representatividad de la media
0% – 10%	Media altamente representativa
10% – 20%	Media bastante representativa
20% – 30%	Media con representatividad
30% – 40%	Media con representatividad dudosa
40% o más	Media no representativa

Ejemplo

¿Cómo es la dispersión de las notas de un exámen si se califica con base 10 respecto de calificarlo base 100?

El CV es igual para ambos casos, puesto que base 100 tanto la media como la desviación típica es 10 veces la media y la desviación típica de calificarlo base 10, por lo tanto la variación de los datos es igual.

Problemas

Los siguientes datos son sobre la cantidad de productos lácteos en malas condiciones que se han encontrado en 4 marcas diferentes. Determina la media de qué marca es más confiable.

Marca 1: $\overline{x}$ = 14, s = 3 Marca 2: $\overline{x}$ = 17, s = 2 Marca 3: $\overline{x}$ = 12, s = 5 Marca 4: $\overline{x}$ = 15, s = 1

Indicador de logro:

1.8 Utiliza el coeficiente de variación para analizar la representatividad de la media aritmética en series de datos diferentes.

Secuencia:

Después de haber trabajado con las medidas de tendencia central y las de dispersión, es posible introducir lo correspondiente al coeficiente de variación, lo cual puede servir de parámetro para determinar la representatividad de la media aritmética.

Propósito:

El coeficiente de variación puede resultar una herramienta muy útil para comparar resultados entre dos series de datos y tener un parámetro de la representatividad de la media, está claro que mientras más representativa es la media, es debido a que los datos se aproximan más a ella, y por lo tanto, la dispersión de los mismos tenderá a ser menor.

Solución de problemas:

Para determinar la confiabilidad, primero sería adecuado calcular los coeficientes de variación de cada marca:

Marca 1: $CV = \frac{s}{\overline{x}}(100) = (3 \div 14)(100) = 21.4\%$

Marca 2: $CV = \frac{s}{\overline{x}}(100) = (2 \div 17)(100) = 11.8\%$

Marca 3: $CV = \frac{s}{\overline{x}}(100) = (5 \div 12)(100) = 41.7\%$

Marca 4: $CV = \frac{s}{\overline{x}}(100) = (1 \div 15)(100) = 6.7\%$

A partir de lo cual resulta claro que la marca 4 posee la media más representativa, luego está la marca 2, sin embargo, la media de la marca 2 es mayor que la de la marca 4, por lo cual se esperaría que la marca 4 sería la más confiable, puesto que sus patrones para medir la cantidad de productos en malas condiciones ha presentado datos más regulares.

En esta parte el docente puede mencionar que si analizamos únicamente la media, está claro que la mejor sería la marca 3, sin embargo posee mucha dispersión en sus datos, lo cual indica que pueden estar pasando varias cosas, o bien que la marca mejoró mucho o bien que empeoró mucho, y por eso no se puede confiar en la media y se necesitaría más información para declinarse por esta opción; por otro lado al menos la marca 4 da la información más certera, además que su media tampoco difiere en gran medida de la media de la marca 3.

Lección 1

1.9 Practica lo aprendido

1. En cada una de las siguientes situaciones, identifica la población, la muestra, las variables estadísticas y cómo se clasifican.
   a) Para determinar el impacto de una política educativa se realiza una prueba diagnóstica a 40 escuelas de todo el país.
   b) Se desea saber la calidad de un producto lácteo y para ello se realiza una prueba a una unidad del producto en cada supermercado del país.
   c) Para establecer el ingreso promedio que tiene una persona en el área rural de El Salvador se realiza una encuesta a 20 personas de cada área rural del país.

2. Determina si las variables estadísticas presentadas a continuación son: variables cualitativas (nominales u ordinales) o variables cuantitativas (discretas o continuas).
   a) Cantidad de hermanos
   b) Relación con sus padres (mala, regular, buena)
   c) Resultados de un examen de matemática
   d) Marca de jabón preferida

3. Clasifica las siguientes estrategias de muestreo como aleatorio simple o aleatorio sistemático.
   a) Numerar la población del 1 al 10 (se repite al finalizar) y luego escoger un número, de modo que todos los que tengan ese número serán parte de la muestra.
   b) Numerar la población y eliminar 3 números y el cuarto es escogido, luego otros 3 y el cuarto es escogido, y así sucesivamente hasta abarcar toda la población.
   c) Hacer grupos en una población y seleccionar 2 de esos grupos al azar para que sean la muestra.
   d) Numerar la población, tirar un dado y seleccionar la persona de la población con ese número, luego tirarlo de nuevo y sumárselo al resultado anterior para seleccionar la otra persona y así sucesivamente.

4. Realiza una muestra de 30 estudiantes de una empresa que dispone de los siguientes datos:

	Femenino	Masculino
Estudiantes	35	15
Profesionales	25	25

5. Determina qué tipo de muestreo no probabilístico consideras más adecuado para cada situación.
   a) Investigación social para una tarea de seminario
   b) Forma de distribución de sustancias ilícitas.
   c) Personas que presentan mayor irritabilidad al conducir
   d) Estudiantes que participan en atletismo.

6. En un estacionamiento de centro comercial se calcula el tiempo promedio que permanece un carro estacionado, y se obtienen los siguientes datos para todo el estacionamiento y para los primeros 30 puestos:

Tiempo	Cantidad de carros
De 0 a 1 hora	12
De 1 a 2 horas	30
De 2 a 3 horas	32
De 3 a 4 horas	16
TOTAL	**90**

Tiempo	Cantidad de carros
De 0 a 1 hora	4
De 1 a 2 horas	10
De 2 a 3 horas	11
De 3 a 4 horas	5
TOTAL	**30**

   a) Calcula media, mediana, moda, varianza y desviación típica tanto para la población como para la muestra.
   b) Considerando que el tiempo promedio que las personas permanecen en el centro comercial es 2 horas con una desviación típica de 0.8 horas, ¿qué promedio es más confiable?

Unidad 8

215

Indicador de logro:

1.9 Resuelve problemas correspondientes a estadísticos y parámetros en muestras y poblaciones.

Solución de problemas:

1a) Población: las escuelas de todo el país; muestra: las 40 escuelas en las que se realiza la prueba diagnóstica; variable: impacto de la política educativa; tipo: cualitativa ordinal (en ocaciones también puede asignarse un valor cuantitativo, sin embargo por las condiciones de la situación, sería más adecuada una escala como aceptable, regular, no aceptable, etc.).

1b) Población: el producto lácteo de todos los supermercados del país; muestra: el conjunto de los productos extraídos de cada supermercado; variable: calidad del producto lácteo; tipo: cualitativa ordinal.

1c) Población: personas en el área rural del país; muestra: las 20 personas de cada área rural del país; variable: ingreso promedio; tipo: cuantitativa continua.

2a) Cuantitativa discreta. **2b)** Cualitativa ordinal. **2c)** Cuantitativa continua. **2d)** Cualitativa nominal.

3a) Muestreo aleatorio sistemático. **3b)** Muestreo aleatorio simple.

3c) Muestreo aleatorio sistemático. **3d)** Muestreo aleatorio simple.

4. Mujeres estudiantes: $\frac{35}{100} \times 100 = 35\%$ Hombres estudiantes: $\frac{15}{100} \times 100 = 15\%$

Mujeres profesionales: $\frac{25}{100} \times 100 = 25\%$ Hombres profesionales: $\frac{25}{100} \times 100 = 25\%$

Entonces para obtener la muestra de 30 personas se pueden utilizar estos porcentajes:

Mujeres estudiantes: 35% de 30 es 11 Hombres estudiantes: 15% de 30 es 5

Mujeres profesionales: 25% de 30 es 7 Hombres profesionales: 25% de 30 es 7

Por lo tanto se puede extraer una muestra de 30 personas de toda la colonia, seleccionando 11 mujeres estudiantes, 7 mujeres profesionales, 5 hombres estudiantes y 7 hombres profesionales.

En este problema en particular, los valores para las muestras no son enteros, y los valores son 10.5, 7.5, 4.5 y 7.5, sin embargo, deben aproximarse a un valor entero teniendo el cuidado de no sobrepasar la muestra. También se podría aproximar siempre al entero mayor y sobrepasar la muestra a 32 personas.

5a) Muestreo por conveniencia, puesto que se hará la investigación a personas del entorno cercano.

5b) Muestreo por bola de nieve, puesto que se requiere información que no se puede obtener de forma directa, tanto por seguridad como por confidencialidad.

5c) Muestreo por estratos, puesto que sería conveniente analizar conductores con diferentes características.

5d) Muestreo discrecional, dado que es preferible que participen las personas con más actitudes para el atletismo.

6a) $\mu = \frac{6 + 45 + 80 + 56}{90} \approx 2.08$, Me = 2.5, Mo = 2.5, $\sigma^2 \approx \frac{31.2 + 12 + 6.4 + 32}{90} \approx 0.91$, $\sigma \approx \sqrt{0.91} \approx 0.95$.

$\bar{x} = \frac{2 + 15 + 27.5 + 17.5}{30} \approx 2.07$, $\tilde{x} = 2.5$, $\hat{x} = 2.5$, $s^2 \approx \frac{10.4 + 4 + 2.2 + 10}{29} \approx 0.92$, $s \approx \sqrt{0.92} \approx 0.96$.

6b) Para este caso se pueden comparar los coeficientes de variación de ambos datos:

$CV_1 = \frac{s}{\bar{x}}(100) \approx (0.95 \div 2.08)(100) \approx 45.7\%$ $CV_2 = \frac{s}{\bar{x}}(100) \approx (0.8 \div 2)(100) = 40\%$

Por lo tanto, es un poco más representativa la media del tiempo que permanecen en el centro comercial, sin embargo, ninguna de las medias es representativa de la variable correspondiente.

Lección 2 Medidas de posición

2.1 Cuartiles

Problema inicial

Al finalizar el año escolar el profesor cuenta las inasistencias de sus estudiantes y obtiene los siguientes datos:

4, 5, 6, 2, 4, 8, 10, 11, 13, 12, 11, 10, 12, 6, 7, 9, 8, 13, 14, 15

a) ¿Cuál es la mediana del conjunto de datos?
b) ¿Cuál es la mediana de la primera mitad del conjunto de datos ordenados de menor a mayor?
c) ¿Cuál es la mediana de la segunda mitad del conjunto de datos ordenados de menor a mayor?

Solución

a) Se ordenan los datos y se calcula la mediana:

2, 4, 4, 5, 6, 6, 7, 8, 8, 9, 10, 10, 11, 11, 12, 12, 13, 13, 14, 15

La mediana es $\dfrac{9+10}{2} = 9.5$.

b) Considerando la primera mitad del conjunto de datos del literal a):

2, 4, 4, 5, 6, 6, 7, 8, 8, 9

La mediana es $\dfrac{6+6}{2} = 6$.

Si el total de datos fuera impar ($2n + 1$) para b) se tendrían que considerar los primeros n datos y para c) los últimos n datos, es decir, sin considerar el dato que coincide con la mediana.

c) Considerando la segunda mitad del conjunto de datos del literal a):

10, 10, 11, 11, 12, 12, 13, 13, 14, 15

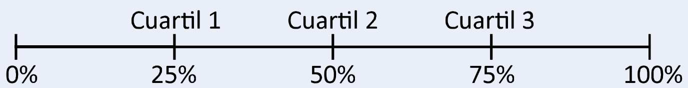
La mediana es $\dfrac{12+12}{2} = 12$.

Esto significa que el 25% de los estudiantes del salón tiene a lo sumo 6 inasistencias, el 50% de los estudiantes tiene a lo sumo 9.5 inasistencias y el 75% tiene a lo sumo 12 inasistencias.

Definición

Los valores de una variable que dividen al conjunto de datos en cuatro partes con igual cantidad de datos se conoce como **cuartiles**.

El cuartil 1 concentra al 25% de los datos menores a este, el cuartil 2 concentra al 50% de los datos menores a él, el cuartil 3 concentra al 75% de los datos menores a él.



La diferencia entre el cuartil 3 y el cuartil 1 (C3 – C1) se conoce como **rango intercuartílico**, denotado por **RI**.

Problemas

Determina y analiza los 3 cuartiles de los siguientes conjuntos de datos obtenidos de la cantidad de personas reportadas con dengue en cada mes por algunos centros asistenciales.

a) 3, 1, 4, 12, 10, 9, 11, 7, 12, 16, 3
b) 6, 3, 7, 8, 10, 15, 8, 12, 17, 2
c) 1, 3, 2, 5, 10, 14, 15, 13, 10, 5, 9, 3, 8
d) 4, 2, 5, 7, 10, 16, 12, 9, 14, 8, 5, 1

Indicador de logro:

2.1 Determina y analiza los cuartiles de una serie de datos simple.

Secuencia:

Luego de trabajar con la parte de muestreo, se abarcarán otro tipo de medidas para describir una serie de datos, las medidas de posición ayudan a describir la forma aproximada de los datos.

Propósito:

Para resolver el Problema inicial se espera que los estudiantes apliquen el concepto de la mediana, para encontrar todos los valores requeridos.

Solución de problemas:

a) 1, 3, 3, 4, 7, 9, 10, 11, 12, 12, 16
El cuartil 1 es 3, el cuartil 2 es 9 y el cuartil 3 es 12.

b) 2, 3, 6, 7, 8, 8, 10, 12, 15, 17
El cuartil 1 es 6, el cuartil 2 es 8 y el cuartil 3 es 12.

c) 1, 2, 3, 3, 5, 5, 8, 9, 10, 10, 13, 14, 15
El cuartil 1 es 3, el cuartil 2 es 8 y el cuartil 3 es 11.5.

d) 1, 2, 4, 5, 5, 7, 8, 9, 10, 12, 14, 16
El cuartil 1 es 4.5, el cuartil 2 es 7.5 y el cuartil 3 es 11.

En cada literal, la forma de calcular los cuartiles es diferente.

El análisis que pueden hacer en cada caso es determinar el valor para el cual el 25%, 50% o 75% de los datos son inferiores a él; así por ejemplo, el 25% de los datos a) y c) se concentran de 1 hasta el valor 3, sin embargo, en b) se concentran de 2 hasta 6, con lo cual se puede pensar que el 25% menor de los datos tiene mayor rango que el 25% menor de las demás series de datos.

2.2 Diagrama de caja y bigotes

Problema inicial

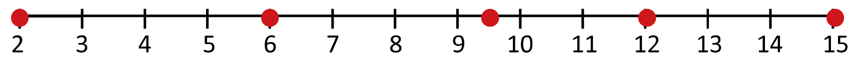
Considerando los datos de las inasistencias de los estudiantes que recolectó el profesor:

4, 5, 6, 2, 4, 8, 10, 11, 13, 12, 11, 10, 12, 6, 7, 9, 8, 13, 14, 15

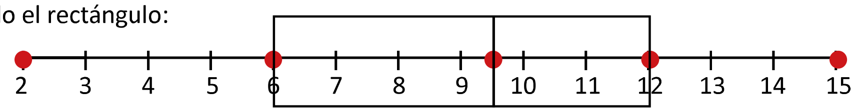
a) Ubica en una recta numérica el valor mínimo, el cuartil 1, el cuartil 2, el cuartil 3 y el valor máximo.
b) Dibuja un rectángulo que cubra desde el cuartil 1 hasta el cuartil 3.

Solución

a) El valor mínimo es 2, el cuartil 1 es 6, el cuartil 2 es 9.5, el cuartil 3 es 12 y el valor máximo es 15; entonces al ubicarlos en una recta se tiene:

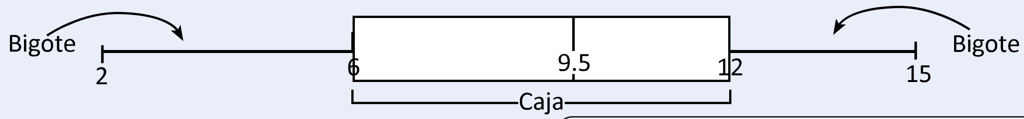


b) Dibujando el rectángulo:



Definición

El diagrama elaborado se conoce como **diagrama de caja y bigotes**.

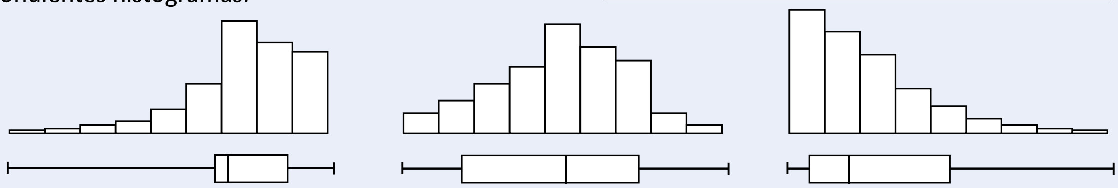


Si el bigote de la izquierda es más corto que el de la derecha significa que el 25% de los datos menores tienen menor rango que el 25% de los datos mayores, así mismo si una caja es angosta significa que el 50% central de los datos tienen rango más estrecho.

La forma del diagrama de caja puede orientar para la descripción de la forma de distribución de los datos, observa los diagramas de caja siguientes y sus correspondientes histogramas.

Estadísticamente, para construir el diagrama de caja se suele utilizar como parámetro el valor de 1.5 veces el rango intercuartílico y los bigotes se representan por la primera y última observación que queda dentro del rango de C1 – 1.5(RI) hasta C3 + 1.5(RI). Esta construcción ayuda a la identificación de datos atípicos en un conjunto de datos.

A un histograma le corresponde un diagrama de caja, pero un diagrama de caja puede corresponder a dos o más histogramas.

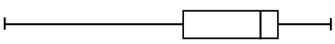
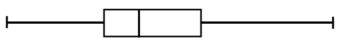
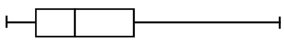


Problemas

1. Elabora y analiza el diagrama de caja de los datos de las personas reportadas con dengue.

   a) 3, 1, 4, 12, 10, 9, 11, 7, 12, 16, 3
   b) 6, 3, 7, 8, 10, 15, 8, 12, 17, 2
   c) 1, 3, 2, 5, 10, 14, 15, 13, 10, 5, 9, 3, 8
   d) 4, 2, 5, 7, 10, 16, 12, 9, 14, 8, 5, 1

2. Observa los siguientes diagramas de caja y estima la forma en que se distribuyen los datos.

   a)  b)  c) 

Indicador de logro:

2.2 Elabora y analiza el diagrama de caja y bigotes de una serie de datos simple.

Secuencia:

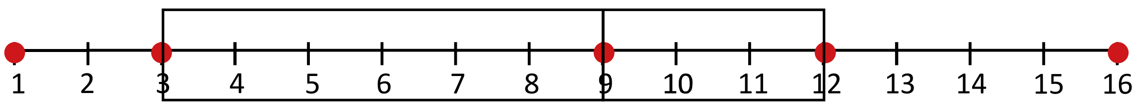
Puesto que en la clase anterior se estableció la definición de los cuartiles, en esta clase se procederá a analizar la gráfica de caja y bigotes, cuyo fundamento principal son los cuartiles.

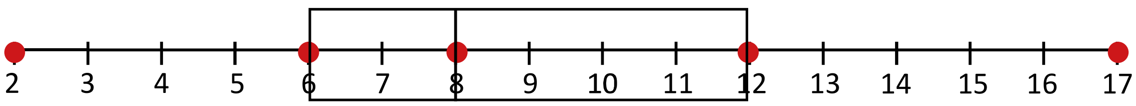
Propósito:

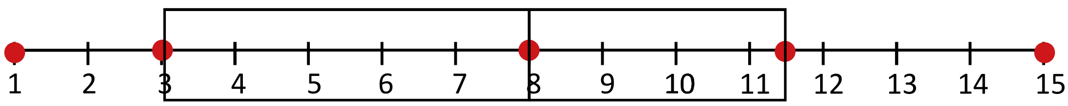
Está claro que a partir de un histograma sí se puede inferir la forma del diagrama de caja y bigotes, sin embargo en esta clase, cuando se pide el histograma que puede estar representado por un diagrama, el objetivo es que el estudiante imagine las posibilidades que puede tener el histograma para cada caso.

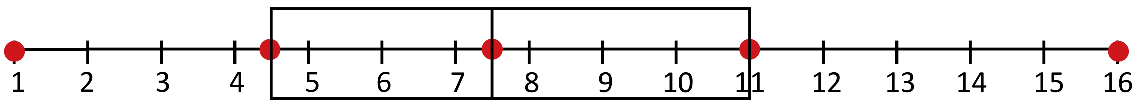
Solución de problemas:

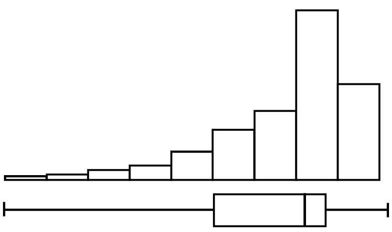
1. Las series de datos son las mismas de la clase anterior, por lo cual ya se tienen calculados los cuartiles, y solamente falta realizar el diagrama.

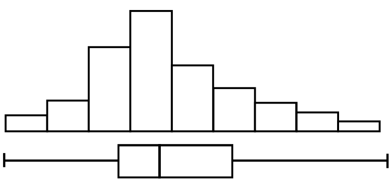
a)


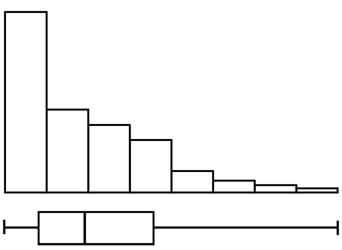
b)


c)


d)


2a)


2b)


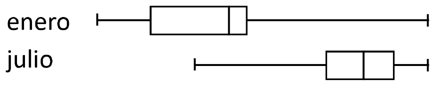
2c)


Todos estos histogramas marcan una forma posible general, lo cual no significa que necesariamente sean los que están asociados al diagrama de caja y bigotes dado, únicamente se puede analizar cierta tendencia en la forma de los datos.

2.3 Análisis del diagrama de caja y bigotes*

Problema inicial

A continuación se presentan dos diagramas de caja correspondientes a las ventas registradas por día durante dos meses diferentes, analiza y luego responde:

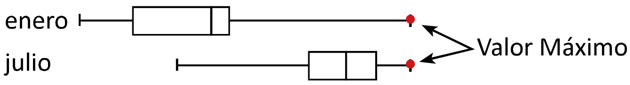


Nota que la cantidad de días es igual para ambos meses.

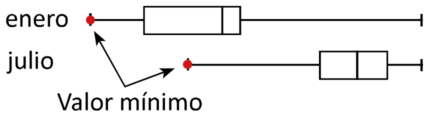
a) ¿En qué mes se dio la mejor venta?
b) ¿En qué mes sucedió la peor venta?
c) Determina cómo fue la variabilidad entre los cuartiles de ambos meses.
d) ¿En qué mes hubo mayor cantidad de ventas?

Solución

a) La mejor venta fue igual en ambos meses, puesto que el valor máximo en ambos diagramas es el mismo.

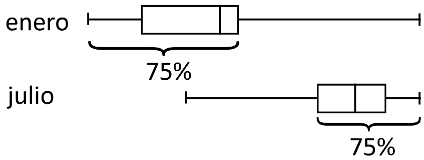


b) La peor venta sucedió en enero, puesto que el valor mínimo del diagrama 1 es menor que el valor mínimo del diagrama 2.



c) El 25% de ventas más bajas, enero tuvo poca variabilidad (rango estrecho) y se concentra en ventas bajas, al contrario de julio, cuyo 25% de ventas más bajas tiene mayor rango y alcanza ventas más altas que enero, por otro lado el rango intercuartílico de enero es mayor que el rango intercuartílico de julio, lo cual significa que hubo más variabilidad en este rango en enero que en julio; finalmente, al analizar el 25% de ventas mayores se determina una variabilidad muy grande en enero viniendo de valores muy bajos, sin embargo en julio el rango es pequeño y se concentra en ventas muy altas.

d) Al menos el 75% de las ventas más bajas de enero están por debajo del cuartil 1 de julio, por esta razón se puede considerar que las ventas de julio fueron mejores que las de enero.



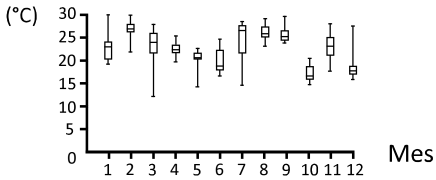
Conclusión

El diagrama de caja brinda información muy valiosa para la comparación de datos valorando la dispersión que pueda existir entre ellos y es una herramienta muy útil para describir los datos de manera certera.

Problemas

Analiza los siguientes diagramas de caja de la temperatura de El Salvador durante los 12 meses del año, luego responde las preguntas.

a) ¿En qué mes variaron más las temperaturas?
b) ¿En qué mes variaron menos las temperaturas?



Indicador de logro:

2.3 Compara los diagramas de caja y bigotes de series de datos en una misma escala.

Secuencia:

Luego de haber introducido el diagrama de caja para una serie de datos, se procurará enfatizar en el análisis y la interpretación de la representación de varios diagramas de caja y bigotes para momentos diferentes, y lograr obtener conclusiones a partir de ellos.

Propósito:

Se espera que los estudiantes logren identificar qué información pueden extraer al comparar dos o más gráficos, para lograr establecer conclusiones y tomar decisiones en función de las conclusiones establecidas a partir de los datos estadísticos.

Solución de problemas:

a) A partir de los gráficos y del rango que se puede apreciar en los diagramas de caja, se puede visualizar que en el mes 3 (marzo) el 25% de las temperaturas más bajas, varió bastante, de manera parecida el mes 12 (diciembre) el 25% de las temperaturas más altas tuvieron valores muy variables (o al menos un valor muy grande) sin embargo, durante este mes se logra apreciar que al menos el 75% de las temperaturas más bajas no tuvo un rango muy grande, y por lo tanto, su variación se aprecia un poco menor que la del mes 3.

b) Durante el mes 4 se logra visualizar el diagrama con comportamiento más regular respecto de los demás, valorando su forma en todos los cuartiles y su rango, además, a pesar que hay otros meses como el 8, 9 o el 10 cuyo rango es parecido, pero en ellos se marca alguna mayor variabilidad entre algunos de sus cuartiles, por lo tanto, el mes que se aprecia con menor variación es el 4 (abril).

Lección 2

2.4 Deciles y percentiles

Problema inicial

Al finalizar el año escolar el profesor cuenta las inasistencias de sus estudiantes y obtiene los siguientes datos:

4, 5, 6, 2, 4, 8, 10, 11, 13, 12, 11, 10, 12, 6, 7, 9, 8, 13, 14, 15

a) ¿Cuál fue el número máximo de inasistencias que tuvo el 20% de los estudiantes con menos inasistencias?
b) ¿Cuál fue el número máximo de inasistencias que tuvo el 60% de los estudiantes con menos inasistencias?
c) ¿Cuál fue el número máximo de inasistencias que tuvo el 5% de los estudiantes con menos inasistencias?
d) ¿Cuál fue el número máximo de inasistencias que tuvo el 95% de los estudiantes con menos inasistencias?

Solución

Se ordenan los datos y se dividen en 20 partes iguales (cada una equivale a un 5%).

2, 4, 4, 5, 6, 6, 7, 8, 8, 9, 10, 10, 11, 11, 12, 12, 13, 13, 14, 15

5% 20% 60% 95%

a) El 20% de estudiantes con menos inasistencias tiene a lo sumo 5 inasistencias.
b) El 60% de estudiantes con menos inasistencias tiene a lo sumo 10 inasistencias.
c) El 5% de estudiantes con menos inasistencias tiene a lo sumo 2 inasistencias.
c) El 95% de estudiantes con menos inasistencias tiene a lo sumo 14 inasistencias.

Definición

Los valores de una variable que dividen al conjunto de datos en diez partes con igual cantidad de datos cada una se conoce como **deciles**.

Cada decil concentra 10% más de los datos que el anterior, el primero concentra el 10% de los datos, el segundo el 20% de los datos y así sucesivamente hasta el decil 9 que concentra el 90% de los datos.

D1 D2 D3 D4 D5 D6 D7 D8 D9

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Los valores de una variable que dividen al conjunto de datos en cien partes con igual cantidad de datos cada una, se conoce como **percentiles**.

Cada percentil concentra 1% más de los datos que el anterior, el primero concentra el 1% de los datos, el segundo el 2% de los datos y así sucesivamente hasta el percentil 99 que concentra el 99% de los datos.

Para calcular el decil d de un total de n datos, se ordenan los datos de menor a mayor y se busca el dato cuya posición se aproxime más al valor $d\frac{n}{10}$. Análogamente para calcular el valor del percentil p, se busca el dato cuya posición se aproxima más al valor $p\frac{n}{100}$.

Problemas

Calcula los deciles indicados en cada serie de datos, luego analiza la información que proveen estos datos.

a) 6, 9, 2, 10, 1, 7, 8, 2, 7, 5, 11, 12, 9, 5, 3, 10, 12, 7, 4, 8. Deciles 3, 5 y 7.
b) 4, 6, 10, 15, 13, 7, 9, 5, 7, 7, 12, 14, 10, 9, 6, 11. Deciles 4, 6 y 9.

Indicador de logro:

2.4 Determina y analiza los deciles y percentiles de una serie de datos simple.

Secuencia:

Para finalizar esta lección se estudian las últimas medidas de posición, los deciles y percentiles, con lo cual se puede realizar un análisis más detallado de la forma que tiene la serie de datos.

Propósito:

En esta clase se introducen tanto deciles como percentiles en el Problema inicial, sin embargo esto únicamente es para introducir la Definición, puesto que para la serie de datos del Problema inicial no tiene mucho sentido calcular percentiles (muchos se repetirán); por esta razón es que en los Problemas no se refiere a percentiles, sino únicamente a deciles.

Solución de problemas:

a) Ordenando los datos: 1, 2, 2, 3, 4, 5, 5, 6, 7, 7, 7, 8, 8, 9, 9, 10, 10, 11, 12, 12.

Decil 3: $d\,\frac{n}{10}$ = 3 × 2 = 6, por lo tanto, es el valor en la sexta posición, es decir, 5.

Decil 5: $d\,\frac{n}{10}$ = 5 × 2 = 10, por lo tanto, es el valor en la décima posición, es decir, 7.

Decil 7: $d\,\frac{n}{10}$ = 7 × 2 = 14, por lo tanto, es el valor en la décimo cuarta posición, es decir, 9.

b) Ordenando los datos: 4, 5, 6, 6, 7, 7, 7, 9, 9, 10, 10, 11, 12, 13, 14, 15.

Decil 4: $d\,\frac{n}{10}$ = 4 × 1.6 = 6.4, por lo tanto, es el valor en la sexta posición, es decir, 7.

Decil 6: $d\,\frac{n}{10}$ = 6 × 1.6 = 9.6, por lo tanto, es el valor en la décima posición, es decir, 10.

Decil 9: $d\,\frac{n}{10}$ = 9 × 1.6 = 14.4, por lo tanto, es el valor en la décimo cuarta posición, es decir, 13.

Lección 2

2.5 Practica lo aprendido

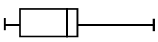
1. Determina los 3 cuartiles de los siguientes conjuntos de datos obtenidos de la cantidad de personas nuevas que son matriculadas en el Centro de Rehabilitación de Ciegos "Eugenia Dueñas". Luego analiza la información que brinda cada cuartil.

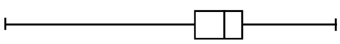
   a) 5, 10, 8, 6, 3, 2, 8, 12, 5, 1, 7, 9, 4

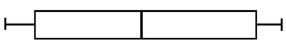
   b) 3, 2, 5, 9, 10,15, 7, 9, 12, 10, 3, 1

2. Elabora y analiza el diagrama de caja de los datos de las personas matriculadas en el Centro de Rehabilitación de Ciegos "Eugenia Dueñas" (numeral 1).

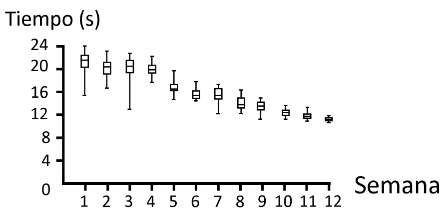
3. Observa los siguientes diagramas de caja y estima la forma en que se distribuyen los datos usando histogramas.

   a) 

   b) 

   c) 

4. Analiza los siguientes diagramas de caja del desempeño de un atleta en el tiempo que tarda para recorrer 100 metros planos durante 12 semanas de entrenamiento.



   a) ¿En qué semana se obtuvo el mejor rendimiento?

   b) ¿En qué semana tuvo el peor rendimiento?

   c) ¿En qué semana marcó el mejor tiempo?

   d) ¿Cómo fue el desempeño en la semana 7?

   e) ¿Qué conclusiones puedes sacar del entrenamiento del atleta?

5. Calcula los deciles indicados en cada serie de datos, luego analiza la información que proveen estos datos.

   a) 10, 6, 7, 11, 13, 8, 9, 5, 9, 10, 12, 12, 7, 9, 11, 15, 4, 6. Deciles 2, 4 y 8.

   b) 10, 5, 7, 11, 8, 9, 12, 7, 6, 10, 9, 8, 14, 13, 9, 11, 5. Deciles 3, 7 y 9.

   c) 8, 5, 4, 2, 1, 7, 3, 9, 10, 9, 8, 6, 2, 11, 3, 14, 11, 8, 13, 10, 6, 12, 10, 4, 3. Deciles 1, 5 y 7.

Indicador de logro:

2.5 Resuelve problemas correspondientes a las medidas de posición.

Solución de problemas:

1a) 1, 2, 3, 4, 5, 5, 6, 7, 8, 8, 9, 10, 12

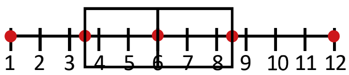
El cuartil 1 es 3.5, el cuartil 2 es 6 y el cuartil 3 es 8.5.

1b) 1, 2, 3, 3, 5, 7, 9, 9, 10, 10, 12, 15

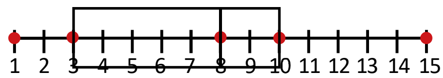
El cuartil 1 es 3, el cuartil 2 es 8 y el cuartil 3 es 10.

El análisis es semejante al realizado en los problemas de la clase 2.1.

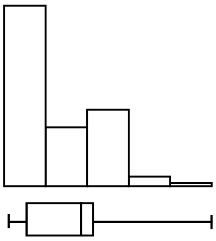
2a)



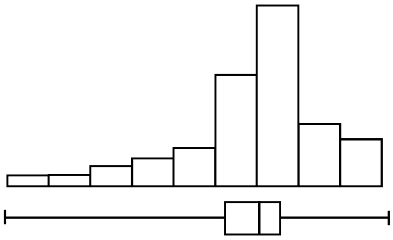
2b)



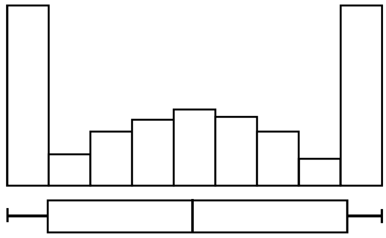
3a)



3b)



3c)



4a) Se podría asegurar que fue durante la semana 12 por dos razones; primero: porque obtuvo uno de los tiempos más bajos, y segundo: porque la mayoría de tiempos están concentrados en un rango muy pequeño.

4b) Durante la semana 4 se observa concentración de tiempos muy altos, aunque en las semanas de la 1 a la 3 también se concentra en tiempos altos, sin embargo, se registró al menos algunos tiempos más bajos que durante la semana 4, por lo tanto, se aprecia que su peor rendimiento es durante la semana 4.

4c) El mejor tiempo fue marcado durante la semana 12.

4d) El 25% de los tiempos más bajos son bastante aceptables, sin embargo, en general no es muy bueno el desempeño de la semana 7, puesto que el 75% de los tiempos resultaron un poco altos, en especial porque la semana anterior tuvo un mejor desempeño.

4e) En general, se marca una tendencia a mejorar semana tras semana (excepto particularidades), por lo cual se puede pensar que fue un buen entrenamiento; por otro lado, cabe destacar que aunque hubo semanas que alcanzó buenos tiempos, esto no indica que el atleta está preparado para garantizar un buen tiempo, sino que debe intentarse concentrar los resultados en un rango pequeño, tal como sucede en la semana 12.

5a) Ordenando los datos: 4, 5, 6, 6, 7, 7, 8, 9, 9, 9, 10, 10, 11, 11, 12, 12, 13, 15.

Decil 2: $d\,\frac{n}{10} = 2 \times 1.8 = 3.6$, por lo tanto, es el valor en la cuarta posición, es decir, 6.

Decil 4: $d\,\frac{n}{10} = 4 \times 1.8 = 7.2$, por lo tanto, es el valor en la séptima posición, es decir, 8.

Decil 8: $d\,\frac{n}{10} = 8 \times 1.8 = 14.4$, por lo tanto, es el valor en la décimo cuarta posición, es decir, 11.

5b) Al ordenar los datos: 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 10, 10, 11, 11, 12, 13, 14.

El decil 3 es 7, el decil 7 es 10, y el decil 9 es 12.

5c) Al ordenar los datos: 1, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 8, 8, 8, 9, 9, 10, 10, 10, 11, 11, 12, 13, 14.

El decil 1 es 2, el decil 5 es 8, y el decil 7 es 10.

2.6 Problemas de la unidad

Se realiza el control de calidad de un tipo de laptop, cuyo objetivo es evaluar el tiempo de duración de la carga de la computadora, para ello se toma una muestra de 50 computadoras y se registran los siguientes datos:

Duración de la batería en horas	Cantidad de laptops
0 a 2	10
2 a 4	15
4 a 6	9
6 a 8	9
8 a 10	5
10 a 12	2

a) Identifica la variable a estudiar y clasifícala.

b) ¿Qué tipo de muestreo es el más adecuado para este control de calidad?

c) ¿Cuánto tiempo dura en promedio la batería de una laptop de este tipo?

d) ¿Cuánto tiempo es más frecuente que dure la batería de una laptop de este tipo?

e) ¿Cuál es el valor de la mediana de este conjunto de datos?

f) Calcula la varianza y la desviación típica de esta muestra.

g) Calcula el coeficiente de variación de estos datos.

h) ¿Cómo es la representatividad de la media para el conjunto de datos?

i) La siguiente información es acerca de la duración de la batería de otro tipo de laptop:

Duración de la batería en horas	Cantidad de laptops
0 a 2	8
2 a 4	17
4 a 6	13
6 a 8	8
8 a 10	3
10 a 12	1

Si se necesita comprar una computadora en la que se requiera la mejor duración de la batería, ¿qué tipo de laptop sería más adecuado comprar? ¿por qué? Compara entre la laptop inicial y la mencionada en el literal i.

Indicador de logro:

2.6 Resuelve problemas correspondientes a estadística descriptiva.

Solución de problemas:

a) La variable es el tiempo de duración de la batería y es de tipo cuantitativa continua.

b) Puesto que solo es un tipo de laptop, lo mejor y más sencillo sería un muestreo aleatorio simple.

Duración de la batería en horas	Cantidad de laptops (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \bar{x}$	$(Pm - \bar{x})^2$	$f(Pm - \bar{x})^2$
0 a 2	10	1	10	–3.6	12.96	129.6
2 a 4	15	3	45	–1.6	2.56	38.4
4 a 6	9	5	45	0.4	0.16	1.44
6 a 8	9	7	63	2.4	5.76	51.84
8 a 10	5	9	45	4.4	19.36	96.8
10 a 12	2	11	22	6.4	40.96	81.92
TOTAL	**50**					

c) $\bar{x} = \dfrac{10 + 45 + 45 + 63 + 45 + 22}{50} = 4.6.$ **d)** $\hat{x} = 3$ **e)** $\tilde{x} = 4$

f) $s^2 = \dfrac{129.6 + 38.4 + 1.44 + 51.84 + 96.8 + 81.92}{49} \approx 8.16,\ s \approx \sqrt{8.16} \approx 2.86.$

g) $CV = \frac{s}{\bar{x}}(100) \approx (2.86 \div 4.6)(100) \approx 62.2\%$

Se considera conveniente asignar a la mediana el valor límite de la clase ya que la posición del elemento medio de la muestra al ordenarse como serie simple es 25.5, es decir entre la clase 2 y 3; sin embargo, también puede tomarse el valor del punto medio de la clase mediana (50 ÷ 2 = 25).

h) A partir del coeficiente de variación, se puede concluir que la media es no representativa.

Para determinar la representatividad de la media, puede recomendar a sus estudiantes revisar la tabla de la clase 1.8.

i) Calculando la media y desviación típica:

Duración de la batería en horas	Cantidad de laptops (f)	Punto medio (Pm)	$f \times Pm$	$Pm - \bar{x}$	$(Pm - \bar{x})^2$	$f(Pm - \bar{x})^2$
0 a 2	8	1	8	–3.36	11.29	90.32
2 a 4	17	3	51	–1.36	1.85	31.45
4 a 6	13	5	65	0.64	0.41	5.33
6 a 8	8	7	56	2.64	6.97	55.76
8 a 10	3	9	27	4.64	21.53	64.59
10 a 12	1	11	11	6.64	44.09	44.09
TOTAL	**50**					

$\bar{x} = \dfrac{8 + 51 + 65 + 56 + 27 + 11}{50} = 4.36,\quad s^2 \approx 5.95,\ s \approx \sqrt{5.95} \approx 2.44,\ CV \approx (2.44 \div 4.36)(100) \approx 55.96\%.$

En este caso se vuelve un poco difícil determinar cuál tipo de laptop sería más adecuada, puesto que aunque la media del segundo tipo de laptop es más representativa (por una diferencia mínima), también la duración promedio es un poco más baja, entonces para este caso con los datos calculados se determina que comprar una u otra sería casi lo mismo.

2.7 Problemas de la unidad

A continuación se presentan los datos obtenidos por semana durante las últimas 8 semanas del rendimiento de un atleta de salto largo, en el cuadro se registra la longitud saltada en metros:

Semana 1	3.5	3.8	3.7	3.8	3.9	3.7	4.0
Semana 2	3.8	4.2	4.3	4.2	4.4	4.6	4.6
Semana 3	4.5	4.8	4.7	4.9	4.9	5.3	5.2
Semana 4	5.2	5.5	5.7	5.6	5.8	5.9	6.0
Semana 5	5.8	6.3	6.5	6.8	6.8	6.9	6.8
Semana 6	6.7	6.9	7.0	6.5	6.8	7.0	7.1
Semana 7	6.9	7.2	7.3	7.2	7.4	7.1	7.5
Semana 8	7.4	7.7	7.8	7.6	7.9	7.8	7.9

a) Realiza una aproximación de los cuartiles para los datos de cada semana.

b) Construye el diagrama de caja y bigotes para cada semana.

c) Realiza un diagrama que compare el desempeño del atleta durante las 8 semanas mediante los diagramas de caja y bigotes.

d) ¿En qué semana se obtuvo el mejor rendimiento?

e) ¿En qué semana tuvo el peor rendimiento?

f) ¿En qué semana marcó el mejor salto?

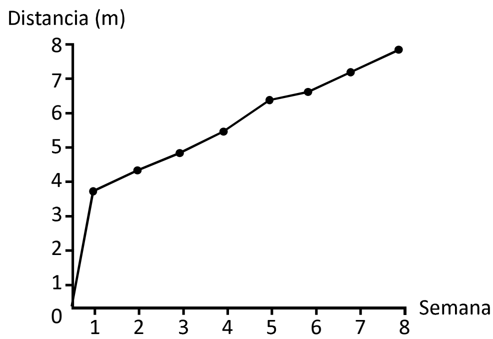
g) ¿Cómo fue el desempeño en la semana 7?

h) ¿A partir de qué semana se puede asegurar con mayor probabilidad que el atleta puede realizar un salto de al menos 5 metros?

i) ¿Se puede pensar que después del entrenamiento de la semana 1, el atleta era capaz de saltar al menos 4 metros? ¿por qué?

j) ¿Qué conclusiones puedes sacar del entrenamiento del atleta?

k) El siguiente gráfico ha sido elaborado con los promedios de cada semana, establece las ventajas entre el diagrama elaborado en el literal c) y el diagrama presentado a continuación.



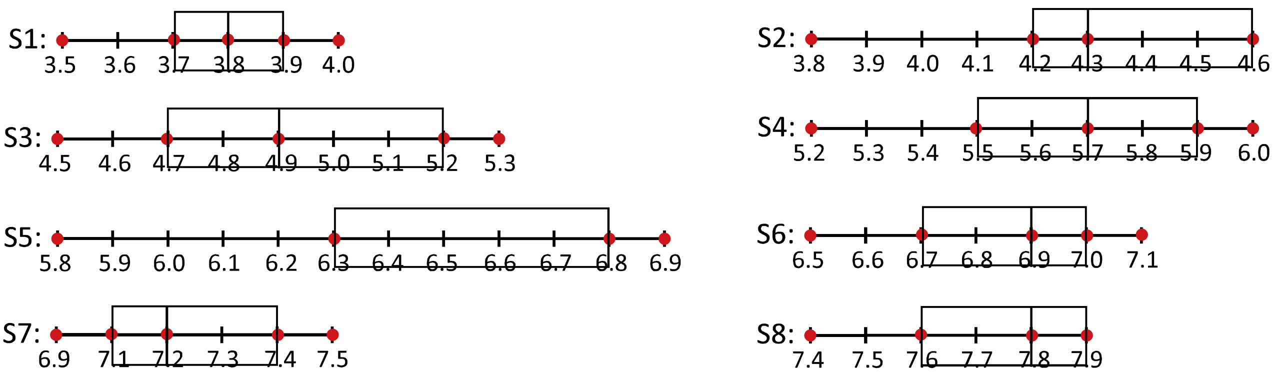
Indicador de logro:

2.7 Resuelve problemas correspondientes a estadística descriptiva.

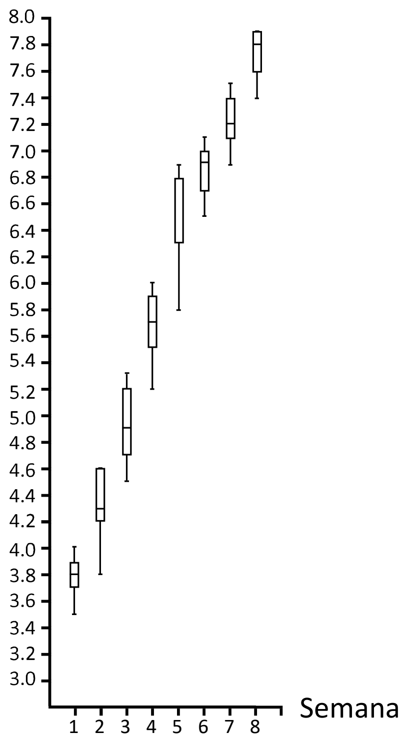
Solución de problemas:

a) Semana 1: C1 = 3.7, C2 = 3.8, C3 = 3.9; semana 2: C1 = 4.2, C2 = 4.3, C3 = 4.6; semana 3: C1 = 4.7, C2 = 4.9, C3 = 5.2; semana 4: C1 = 5.5, C2 = 5.7, C3 = 5.9; semana 5: C1 = 6.3, C2 = 6.8, C3 = 6.8; semana 6: C1 = 6.7, C2 = 6.9, C3 = 7.0; semana 7: C1 = 7.1, C2 = 7.2, C3 = 7.4; semana 8: C = 7.6, C2 = 7.8, C3 = 7.9.

b)



c) Distancia (m)



d) El mejor rendimiento fue durante la semana 8, puesto que el rango entre el cuartil 1 y el dato superior es muy poco.

e) Durante la semana 3 el 75% de los datos se distribuyeron en un rango muy grande, muy parecido a la semana 4 y 5.

f) Durante la semana 8.

g) Fue bastante regular, casi todos los cuartiles dividen los datos en rangos muy parecidos, y el atleta presentó saltos de longitud muy aceptable.

h) A partir de la semana 3 poco menos del 50% de los datos sobrepasan los 5 metros, sin embargo, a partir de la semana 4 con total seguridad los sobrepasa.

i) No, porque a pesar que logró esa marca en los saltos de esa semana, casi el 100% de los datos está por debajo de esa marca, por lo tanto, sería casi imposible que el atleta lo lograra en esa semana.

j) Fue un entrenamiento muy bueno, que presentó grandes márgenes de mejora entre cada semana, y que a partir de la semana 6 se reguló un poco y sus marcas fueron muy buenas.

k) En el gráfico de los promedios únicamente se puede apreciar la tendencia de mejora, pero no se logra saber el comportamiento de los resultados en cada semana, las semanas con mayor estabilidad y simetría en los datos. Por lo tanto, el gráfico con diagramas de caja puede ser utilizado para extraer información con más detalle, incluso se puede tener una idea sobre la dispersión a partir del rango de los datos, entre otras ventajas.

3.1 Práctica en GeoGebra: análisis estadístico

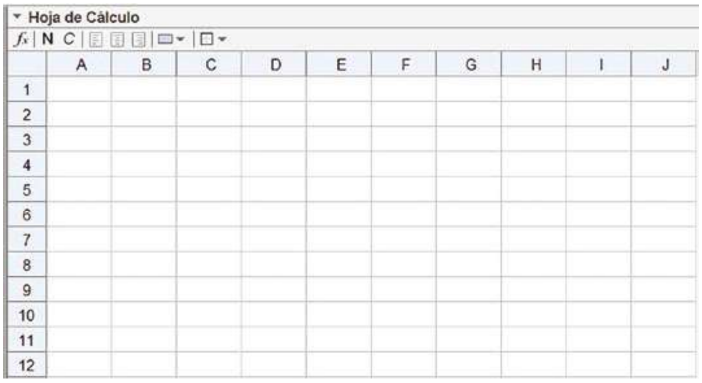
Para esta práctica se utilizarán los recursos de GeoGebra para realizar análisis estadístico de una variable, y construir diagramas de caja y bigotes sobre las situaciones planteadas en la unidad. Para ello sigue los pasos indicados en la parte de Práctica y construye los diagramas y el análisis necesario. Luego trabaja en GeoGebra la parte Actividades que está al final de esta práctica.

Práctica

Retomando los datos del problema de la clase 1.5:

Velocidad en Km/h						
60	65	40	80	80	90	45
70	100	70	50	80	55	120
75	65	90	85	70	100	55
110	70	95	70	80	115	100

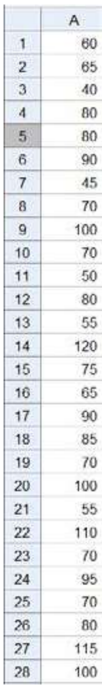
1. Se utilizará la vista de hoja de cálculo, para ello puedes utilizar el menú que se abre al iniciar GeoGebra en la opción Hoja de Cálculo, o bien, desde el menú vista dando click en la opción Hoja de Cálculo, y se abrirá una ventana como la que se muestra abajo.



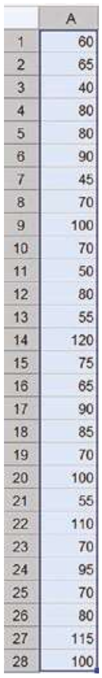
2. Ingresa los datos de la tabla de velocidades uno por uno en la columna A de la vista de Hoja de Cálculo.

3. Manteniendo presionado el clic izquierdo, selecciona todos los datos que se ingresaron, estos quedarán sombreados en un color azul suave.

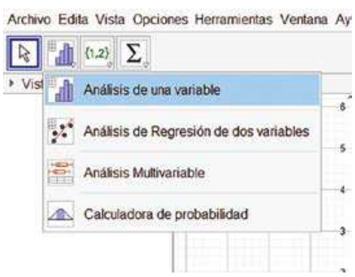
4. Ahora utiliza el botón Análisis de una variable, luego se abrirá un cuadro de texto llamado "Fuente de datos", en ella aparecerán los datos seleccionados en el paso anterior, presiona Analiza y se mostrará una gráfica como la que se muestra a continuación.

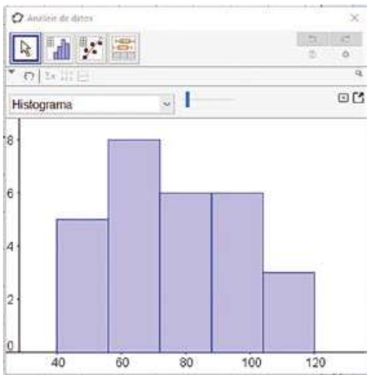


	A
1	60
2	65
3	40
4	80
5	80
6	90
7	45
8	70
9	100
10	70
11	50
12	80
13	55
14	120
15	75
16	65
17	90
18	85
19	70
20	100
21	55
22	110
23	70
24	95
25	70
26	80
27	115
28	100

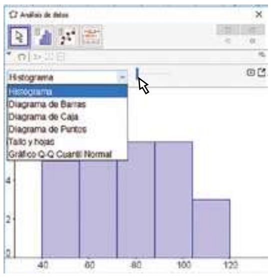


	A
1	60
2	65
3	40
4	80
5	80
6	90
7	45
8	70
9	100
10	70
11	50
12	80
13	55
14	120
15	75
16	65
17	90
18	85
19	70
20	100
21	55
22	110
23	70
24	95
25	70
26	80
27	115
28	100

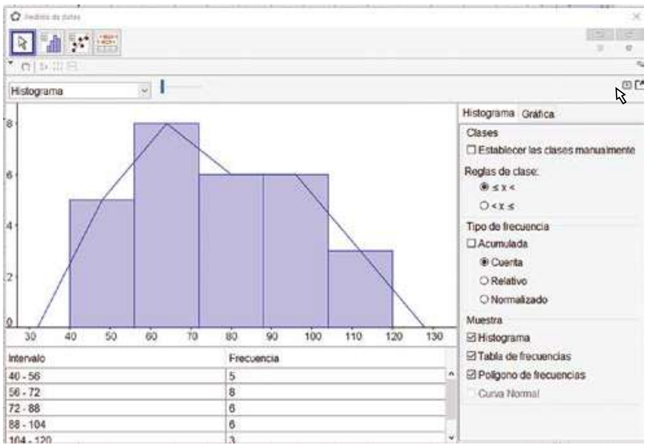




5. Al expandir las opciones, puedes notar que la gráfica que se presenta es un histograma, y que es posible seleccionar diferentes tipos de gráficos estadísticos, entre los cuáles están el diagrama de barras (estudiado en educación básica), histograma (estudiado en 8° grado), diagrama de caja (estudiado en esta unidad) y otros diagramas que no se han estudiado por el momento.

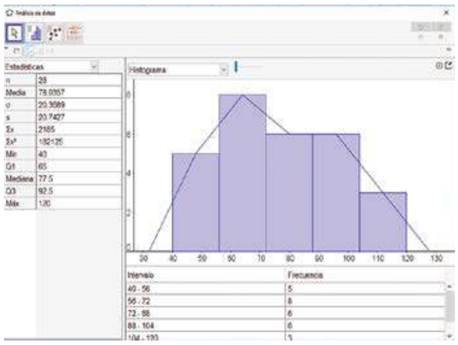


6. En la parte superior derecha puedes extender las opciones de la vista, en ella marca las opciones de tabla de frecuencias (la cual puedes construir manualmente) que creará una tabla de distribución de frecuencias de forma automática; y marca la opción polígono de frecuencias, y se obtendrá el siguiente resultado.



7. Finalmente selecciona la opción Estadísticas, ubicado en la parte superior izquierda, con ícono de un símbolo de sumatorio, así se obtendrán algunos estadísticos como la media, la desviación estándar (muestral y poblacional), cuartiles, mediana, mínimo, máximo, etc.

8. Comprueba la resolución de este problema, verificando tu respuesta y luego corrige si es necesario.



Actividades

Utiliza la herramienta de la hoja de cálculo de GeoGebra para resolver el problema de la clase 2.7 acerca de problemas de la unidad, para ello, como se requiere comparar datos por cada semana, ingresa en cada columna los datos de una semana, por ejemplo, los datos de la semana 1 en la columna A, la semana 2 en la columna B, y así sucesivamente hasta llegar a la semana 8 en la columna H. Luego selecciona todos los datos, y utiliza la opción de análisis multivariante.

Indicador de logro:

3.1 Utiliza un software matemático para realizar el análisis estadístico descriptivo de una serie de datos.

Secuencia:

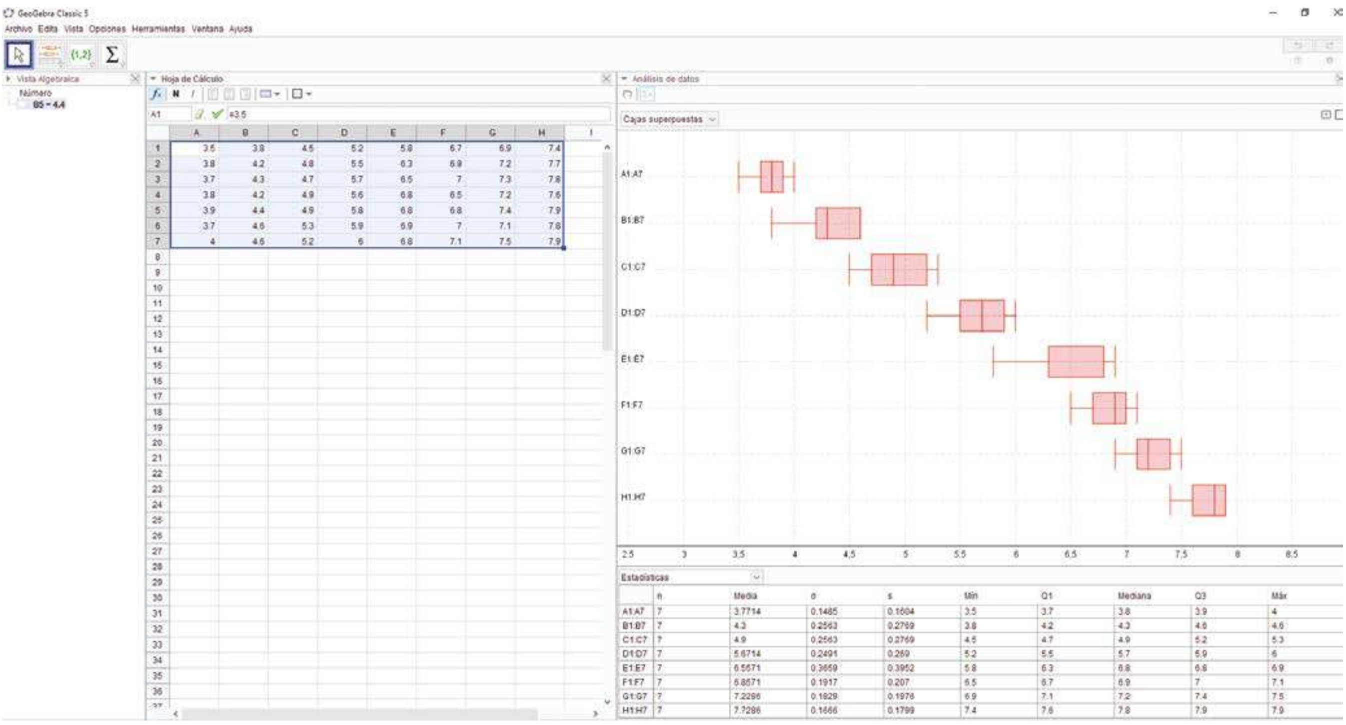
Luego de haber desarrollado todos los contenidos sobre los conceptos básicos de estadística descriptiva, se utilizará GeoGebra para hacer y realizar el análisis estadístico de series de datos.

Propósito:

Utilizar software y contextualizar la forma de realizar análisis estadístico, en donde no es prioritario el cálculo, sino la interpretación y análisis que se pueda realizar de los cálculos realizados por el software.

Solución de problemas:

Siguiendo las indicaciones del problema, se obtendrá el siguiente resultado en GeoGebra:



Con GeoGebra es muy sencillo generar la gráfica que se realizó en la clase 2.7, y con lo cual el énfasis puede ser en el análisis y la correcta interpretación de los resultados. En esta parte hay que destacar que, actualmente todos los análisis estadísticos utilizan software, debido a la eficiencia que esto supone, además de aspectos como reducción de errores de cálculo, y el hecho de que para trabajar con bases de datos gigantes (obtenidas de censos o estudios muestrales a gran escala) es humanamente imposible realizar el cálculo de los parámetros y estadísticos correspondientes. Finalmente, cabe destacar que Geogebra presenta muchas herramientas para trabajar diferentes áreas de la matemática, y es suficiente para realizar análisis estadístico con intención didáctica, sin embargo, no es un software adecuado para trabajar bases de datos gigantes; para ello se suelen usar diferentes softwares estadísticos como: R (libre), Stata (comercial), SPSS (comercial), etc.