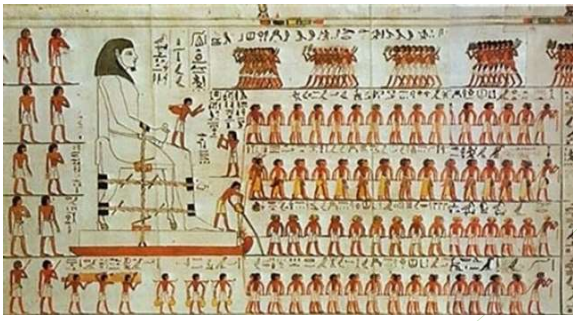
Estadística descriptiva

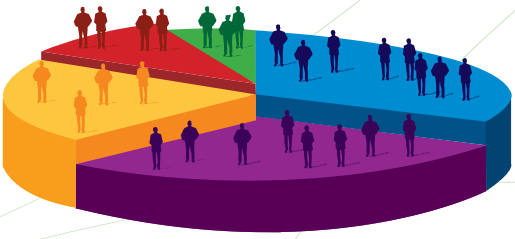
Unidad 8

El registro de la información ha sido una tarea que se ha dado durante toda la historia, esto debido a la importancia de mantener un registro sobre datos como: la cantidad de nacimientos, la tasa de mortalidad, las ventas, las deudas, entre otros. Los primeros aportes sobre este tipo de información datan desde aproximadamente el año 3050 a.C., asimismo se encuentran registros en la biblia y en la cultura china de hace unos 40 siglos, además se considera que los mayas poseían un gran dominio de estos conocimientos, los cuales utilizaban para resolver problemas referentes a sus actividades cotidianas.

Es hasta el siglo XIX que se alcanza un desarrollo pleno de los métodos utilizados para el estudio de las características de un conjunto de datos, con la publicación del primer censo con estas descripciones por el Ministro del Interior francés Chaptal y se van sistematizando a lo largo del desarrollo de este siglo. La aplicación de la estadística descriptiva siempre ha sido muy útil especialmente en aspectos demográficos, para extraer las características esenciales de los datos, analizarlos y dar conclusiones.



Registros estadísticos de los egipcios en la antigüedad.



Los estudios demográficos dieron origen a la estadística descriptiva.

Algunas temáticas que se desarrollarán durante esta unidad son la idea de muestreo, las medidas de tendencia central y dispersión para muestras y poblaciones, coeficiente de variación y medidas de posición, con especial énfasis en el diagrama de caja y bigotes; luego se presenta una práctica en **GeoGebra** para afianzar los aprendizajes con el uso correcto del recurso tecnológico.

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.

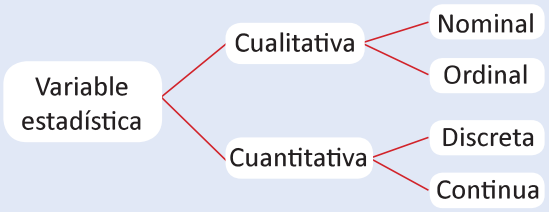
Definiciones

- 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.



- 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.

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.

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.

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.

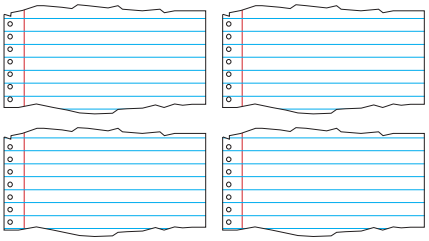
2. Determina si las variables estadísticas presentadas a continuación son variables cualitativas (nominales u ordinales) o variables cuantitativas (discretas o continuas).

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

1.2 Actividad introductoria

Materiales

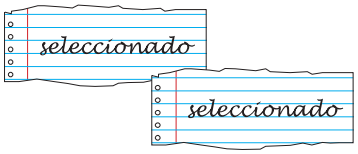
- Pedacitos de papel (uno por cada estudiante)
- Plumón



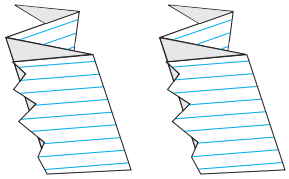
Actividad 1

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?

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: $\frac{24}{60} \times 100 = 40\%$

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

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

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

Niños de Primer año: 20% de 20 es 4

Niñas de Segundo año: 15% de 20 es 3

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

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

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

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.

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

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

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

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

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.

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 $\bar{x}$ de un conjunto de datos, se denota por CV, y se calcula: $CV = \frac{s}{\bar{x}}(100)$.

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

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?

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:

$$\underbrace{2, 4, 4, 5, 6, 6, 7, 8, 8, 9}, \underbrace{10, 10, 11, 11, 12, 12, 13, 13, 14, 15}$$

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

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

$$\underbrace{2, 4, 4, 5, 6}, \underbrace{6, 7, 8, 8, 9}$$

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

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):

$$\underbrace{10, 10, 11, 11, 12}, \underbrace{12, 13, 13, 14, 15}$$

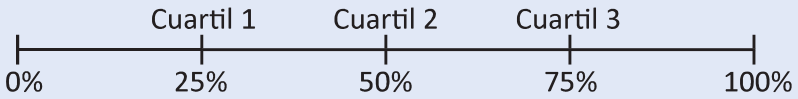
La mediana es $\frac{12 + 12}{2} = 12$.

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

2.2 Diagrama de caja y bigotes

Problema inicial

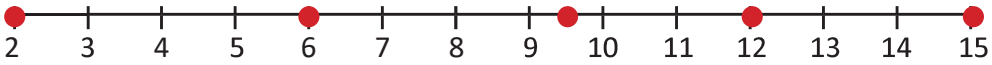
Considerando los datos de las inasistencias de los estudiantes que recolectó el profesor:

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

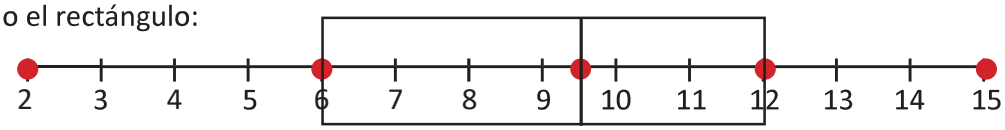
a) Ubica en una recta numérica el valor mínimo, el cuartil 1, el cuartil 2, el cuartil 3 y el valor máximo.
b) Dibuja un rectángulo que cubra desde el cuartil 1 hasta el cuartil 3.

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:

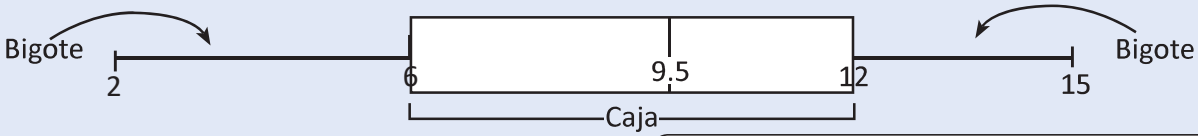


b) Dibujando el rectángulo:



Definición

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

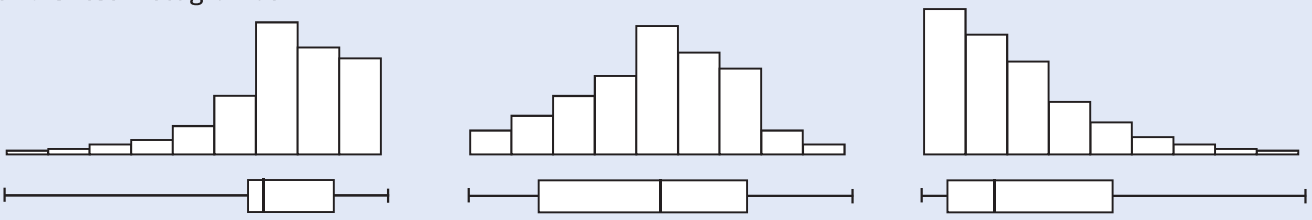


Si el bigote de la izquierda es más corto que el de la derecha significa que el 25% de los datos menores tienen menor rango que el 25% de los datos mayores, así mismo si una caja es angosta significa que el 50% central de los datos tienen rango más estrecho.

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.

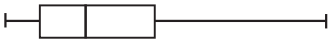


Problemas

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

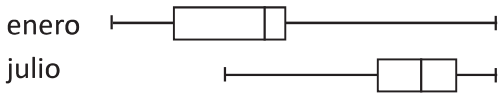
2. Observa los siguientes diagramas de caja y estima la forma en que se distribuyen los datos.

a) 
b) 
c) 

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:

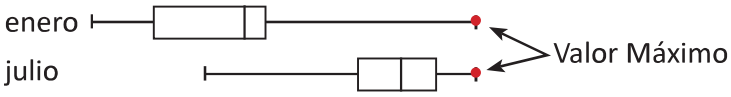


Nota que la cantidad de días es igual para ambos meses.

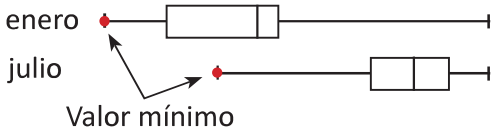
a) ¿En qué mes se dio la mejor venta?
b) ¿En qué mes sucedió la peor venta?
c) Determina cómo fue la variabilidad entre los cuartiles de ambos meses.
d) ¿En qué mes hubo mayor cantidad de ventas?

Solución

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

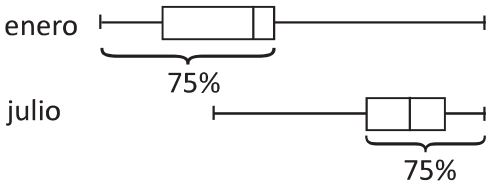


b) La peor venta sucedió en enero, puesto que el valor mínimo del diagrama 1 es menor que el valor mínimo del diagrama 2.



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.



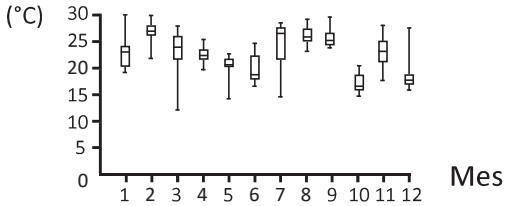
Conclusión

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?



2.4 Deciles y percentiles

Problema inicial

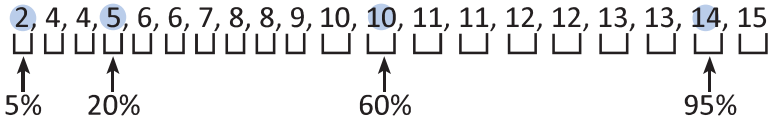
Al finalizar el año escolar el profesor cuenta las inasistencias de sus estudiantes y obtiene los siguientes datos:

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%).

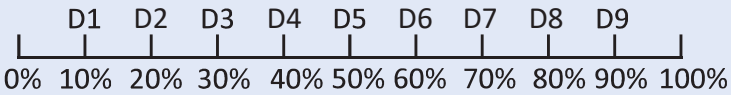


a) El 20% de estudiantes con menos inasistencias tiene a lo sumo 5 inasistencias.
b) El 60% de estudiantes con menos inasistencias tiene a lo sumo 10 inasistencias.
c) El 5% de estudiantes con menos inasistencias tiene a lo sumo 2 inasistencias.
c) El 95% de estudiantes con menos inasistencias tiene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



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.

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.

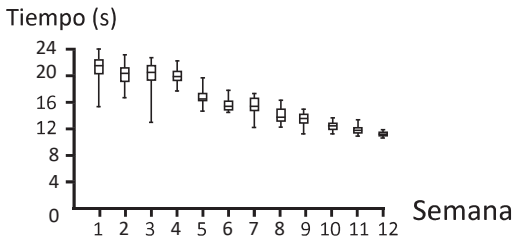
   a) ¿En qué semana se obtuvo el mejor rendimiento?

   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.

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.

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?

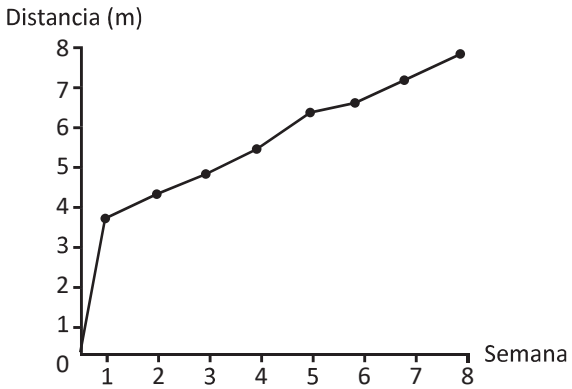
g) ¿Cómo fue el desempeño en la semana 7?

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.



3.1 Práctica en GeoGebra: análisis estadístico

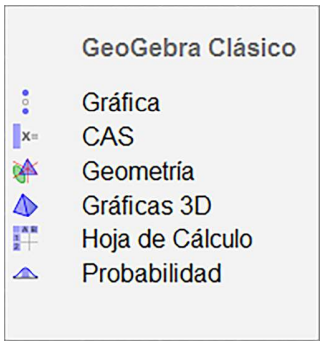
Para esta práctica se utilizarán los recursos de GeoGebra para realizar análisis estadístico de una variable, y construir diagramas de caja y bigotes sobre las situaciones planteadas en la unidad. Para ello sigue los pasos indicados en la parte de Práctica y construye los diagramas y el análisis necesario.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

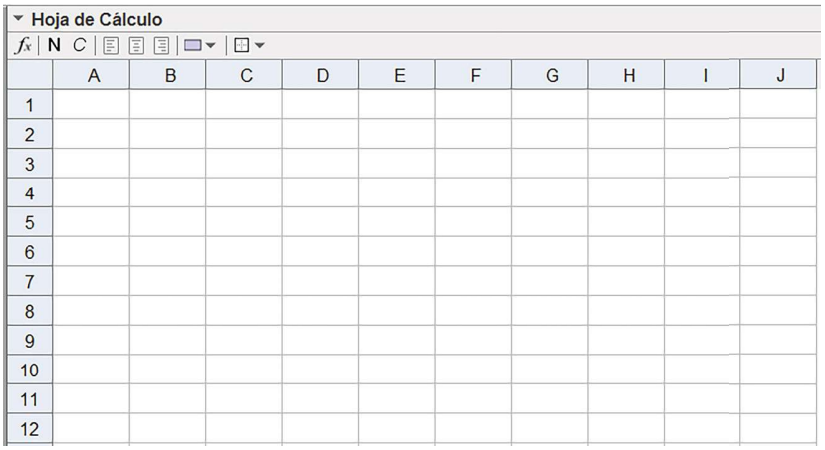
Práctica

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.

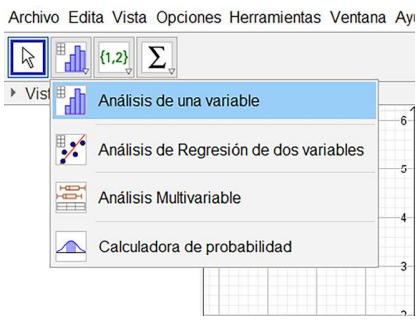


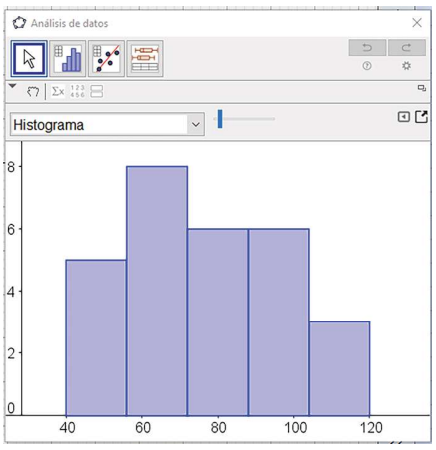


2. Ingresa los datos de la tabla de velocidades uno por uno en la columna A de la vista de Hoja de Cálculo.

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.

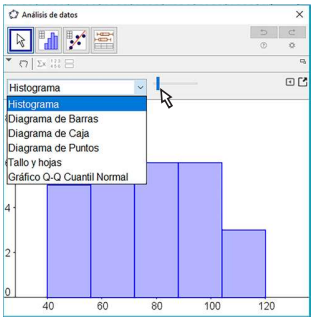




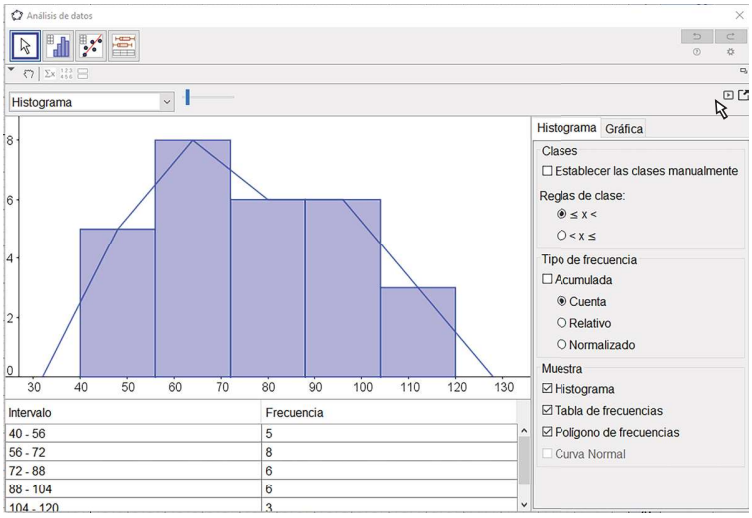
	A
1	60
2	65
3	40
4	80
5	80
6	90
7	45
8	70
9	100
10	70
11	50
12	80
13	55
14	120
15	75
16	65
17	90
18	85
19	70
20	100
21	55
22	110
23	70
24	95
25	70
26	80
27	115
28	100

	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.

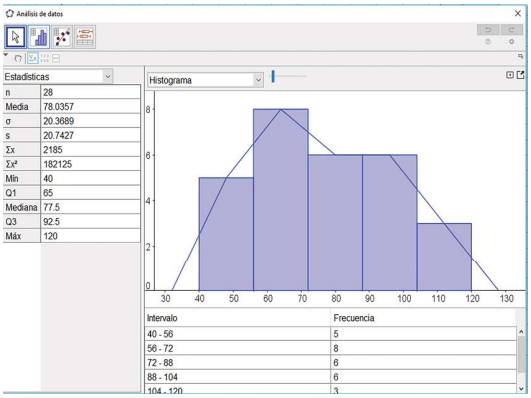


6. En la parte superior derecha puedes extender las opciones de la vista, en ella marca las opciones de tabla de frecuencias (la cual puedes construir manualmente) que creará una tabla de distribución de frecuencias de forma automática; y marca la opción polígono de frecuencias, y se obtendrá el siguiente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