Bloque 4 ESTADISTICA DESCRIPTIVA

Marco histórico:

Según los historiadores, en el año 3000 A.C. los babilonios recopilaban datos en sus tabillas acerca de la producción agrícola y los intercambios mediante trueque. También se sabe que los egipcios y griegos realizaban registros similares. Y en la Biblia se comenta sobre la realización de censos de personas y bienes. En esta época, surge la estadística como una herramienta que le permite a los gobernantes administrar los recursos materiales. Por lo general la gente común tiene este concepto al pensar en estadística. Sin embargo, actualmente se considera a la ciencia estadística como algo mucho más amplio, siendo la estadística descriptiva una herramienta que su utiliza para analizar en forma cuantitativa un conjunto de datos, una rama muy importante de la misma.

Podemos estructurar nuestro estudio en función de tres preguntas:

1)¿Cómo puede reemplazarse un conjunto de datos por un solo valor que sea "lo más representativo" de todos ellos?

2)¿Cuán representativa es esta medida de un conjunto de datos?

3)¿Qué otros valores se necesita para describir de manera adecuada nuestro conjunto de datos?

Como respuesta a estas tres preguntas surgen las medidas de tendencia central, las medidas de variabilidad y las medidas de posición y de forma.

Pero antes de centrarnos en el estudio de estas medidas, veamos diversas formas de organizar un conjunto de datos. La idea de agrupar los datos según algún criterio, constituye de alguna manera un resumen de los datos, y como todo resumen, implica una pérdida de información, en pos de simplificar su lectura e interpretación.

Actividad 5 Cuestionario:

¿Quienes fueron los primeros en registrar datos de la producción agrícola?

¿Qué referencias estadísticas hay en La Biblia?

¿Para qué les sirve a los gobernantes?

¿Para qué se utiliza la Estadística Descriptiva?

¿Cuales son las preguntas que me permiten estructurar el estudio de la Estadística

Tablas de frecuencias:

Una tabla de frecuencias es un arreglo de doble entrada, en donde los datos se agrupan por clase. Por lo general una clase es un dato (en el caso de variables que toman pocos valores), o un intervalo (en el caso en que tome muchos valores).

Primera columna: Se representan las clases. Se considera por lo general que las clases deben ser homogéneas. (En el caso de intervalos, significa que conserven todos de la misma amplitud).
Segunda columna: Se indica la frecuencia de cada clase: El número de observaciones pertenecientes a cada una de ellas.
Tercer columna (opcional): Se indica la frecuencia acumulada: El número de observaciones menores o iguales al límite superior de la clase.

Ejemplo: En la siguiente tabla se representaron los tiempos (en minutos) que necesitan 20 operarios para realizar determinada tarea.

Clase	Frecuencia f_i	Frecuencia acumulada F_i	frecuencias Porcentuales $f_{\%} = \frac{f_i}{N} \cdot 100$
0-1 minutos	8	8	8/30*100=26.7%
1-2 minutos	12	20	40%
2-3 minutos	10	30	33.3%

Por lo general el preferible, no representar frecuencias absolutas, sino relativas (la frecuencia dividida entre el total de datos), o porcentuales (las frecuencias representadas como porcentaje del total de los datos). En este caso, se denominará tablas de frecuencias relativas o tablas de frecuencias porcentuales). En nuestro ejemplo, esto está representado en la tercera columna. Observar que los valores corresponden a realizar el cálculo $f_r = \frac{f_i}{N}$ y en el caso e frecuencias porcentuales, $f_{\%} = \frac{f_i}{N} \cdot 100$

El histograma:

Una manera gráfica de volcar esta información es mediante un diagrama de barras o histograma: es un gráfico en el cuál se dibujan en el eje de las variables independientes las clases, y en el eje de las dependientes las frecuencias absolutas.

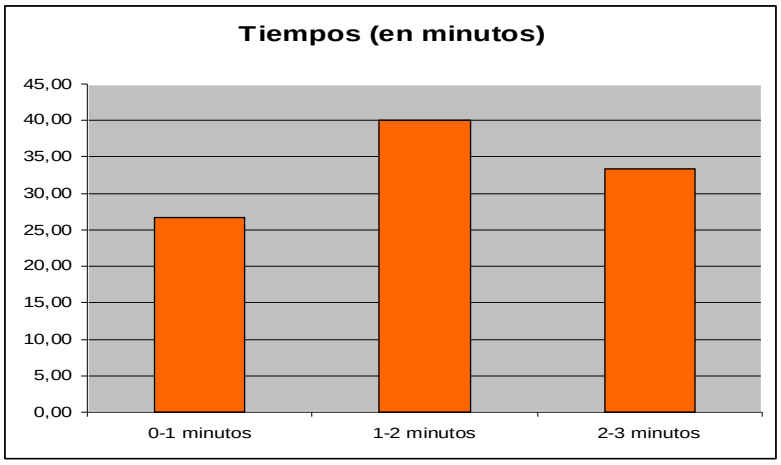


Diagrama de torta:

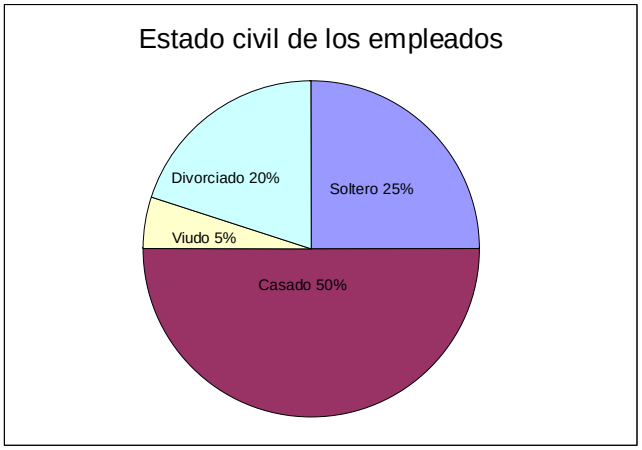
Otra forma de representar los datos, es mediante los conocidos diagramas de torta, o diagramas circulares. Un diagrama de torta, es preferido cuando las variables son nominales. (Sexo, estado civil, por ejemplo). En el diagrama de torta, se representa un círculo en el cuál cada área representa una categoría de la variable. El tamaño de área es proporcional a la cantidad de datos. Por ejemplo, considerar el siguiente conjunto de datos representada por la tabla de frecuencias que se adjunta, donde se detallan el estado civil de los empleados de una empresa:

Estado Civil	Porcentaje
Soltero	25%
Casado	50%
Viudo	5%
Divorciado	20%

Para construir un diagrama de torta, se debe calcular el área correspondiente a cada categoría, que será proporcional a la cantidad de los datos en esta. Para ello, tomamos en cuenta que el área del sector circular será proporcional al ángulo que lo determina.

*Por ejemplo, para calcular el ángulo que le corresponde a la categoría **Soltero**, podemos calcular: $\acute{a}ngulo = \frac{25}{100} \cdot 360^{\circ} = 90^{\circ}$

El aspecto del gráfico será:



Actividad 6:

1. Un grupo de alumnos necesita de los siguientes tiempos para realizar una tarea: (Pablo, 1hora); (José, 2 horas); (Estela, 1 hora); (Julio, 3 horas); (Eugenia, 2 horas); (Susana, 4 horas).

Realiza una tabla de frecuencias indicando clase, frecuencia y frecuencia acumulada.

2. Confecciona un histograma con los datos del Ítem 1.
3. Confecciona un diagrama de torta con los datos del Ítem 1.

Medidas de tendencia Central:

Media: La media aritmética o promedio, es probablemente la medida descriptiva más ampliamente difundida. Su fórmula es:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_N}{N}$$

A pesar de ser una medida tan utilizada, presenta serios inconvenientes si se presentan datos muy atípicos (Datos con un comportamiento sensiblemente distinto del conjunto de datos). Considerar por ejemplo.

4, 2, 1, 3, 100

$$\mu = \frac{110}{5} = 22{,}5$$

Luego, si calculamos el promedio sin incluir el 100:

$$\mu = 2$$

Otro inconveniente del promedio, es que no puede utilizarse cuando se tienen magnitudes ordinales.

Altura (en metros)		frecuencia observada	frecuencia acumulada	frecuencias relativas	frecuencias relativas acumuladas
1.4	1.45	6	6	0.02	0.02
1.45	1.5	15	21	0.05	0.07
1.5	1.55	21	42	0.07	0.14
1.55	1.6	83	125	0.27	0.42
1.6	1.65	67	192	0.23	0.64
1.65	1.7	55	247	0.18	0.83
1.75	1.8	22	269	0.07	0.9
1.8	1.85	15	284	0.05	0.95
1.9	1.95	10	294	0.03	0.98
1.95	2	6	300	0.02	1

Mediana: Es el valor tal que si se ordenaran los datos en forma ascendente ocuparía la posición central, es decir, tendría igual número de datos menores que mayores.

Indiquemos por $x_{(i)}$ el dato que ocupa la posición i-esima cuando se los ordena en forma ascendente. Por ejemplo, en el conjunto de datos ordenados: 1,1,2,3,6,7. $x_{(2)} = 1 \quad x_{(5)} = 6$.

Luego: $\tilde{x} = \begin{cases} x_{(\frac{N+1}{2})} & \text{Si N es impar} \\ \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} & \text{Si N es par} \end{cases}$.

La mediana tiene la propiedad de ser mucho más "robusta", es decir, mucho menos sensible a la influencia de datos extremos.

Observación: La mediana no tiene porqué tratarse de un dato, sino que es un valor que puede haber o no sido observado.

Moda: Es el dato con mayor frecuencia, es decir, el dato que más se repite.

Forma de cálculo para datos agrupados: Por lo general, los datos no vienen ordenados en forma suelta: Por cuestiones relativas a la facilidad en su recolección o la dificultad en imputar un dato concreto (por ejemplo en las encuestas relativas a gastos, el valor del gasto familiar no puede ser determinado mas allá de un rango de valores, o tomando como ejemplo las mediciones con aparatos calibrados, muchas veces se informa un rang de valores donde el valor se encuentra determinado). Tomar como ejemplo el conjunto de datos correspondientes a las alturas de un conjunto de varones adultos jóvenes de entre 19 y 30 años.

Medias de tendencia central con datos agrupados:

Media Aritmética:

Cuando los datos están agrupados, el cálculo de las medidas de resumen cambia. Debemos elegir un representante de cada clase, que denotamos por x_i. Por lo general, se elige el punto medio de cada clase. $x_i = \frac{L_s + L_i}{2}$, en el caso de tratarse de intervalos. La fórmula de la media queda: $\mu = \frac{\sum_{i=1}^{n} x_i f_i}{n}$

Mediana:

En caso de datos agrupados se recomienda la siguiente formula por interpolación lineal:

$$\tilde{x} = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} C$$

Donde L_i es el límite inferior del intervalo de la mediana; F_{i-1} es la frecuencia acumulada del intervalo anterior al cual se encuentra la mediana; f_i es la frecuencia del intervalo en el que se encuentra la mediana y C es el ancho de la clase que la contiene.

Moda:

Al igual que la media se calcula por interpolación lineal como:

$$\hat{x} = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} C$$

Donde L_i es el limite inferior del intervalo modal Δ_1 es la diferencia de frecuencia con el intervalo inmediatamente anterior Δ_2 es la diferencia de frecuencia con el intervalo inmediatamente posterior y C es el ancho de clase que la contiene.

Actividad 7:

Calcula media, mediana y moda con los datos de la Actividad 6 Item 1.

Medidas de variabilidad

Un interrogante que surge naturalmente, al calcular el promedio, es ¿Cuán representativo es este promedio del conjunto de datos? Ya que si los datos se encuentran muy dispersos, tendrá escasa representatividad.

Varianza:

Surge naturalmente la idea de calcular el promedio de las desviaciones de los datos respecto del promedio, esto es:

$$\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}$$

Sin embargo, definido de esta forma, tendría escasa utilidad ya que las desviaciones positivas se compensarían con las negativas de forma tan que la suma de siempre cero (¿Por qué?). Por lo tanto, debemos definir alguna transformación de los datos que salve este inconveniente. Por ejemplo: Se podrían elevar las diferencias al cuadrado, y por otro lado, el peso de los datos más alejados sería mayor. De esta forma surge la definición de varianza como:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}.$$

En el caso de datos agrupados, debemos utilizar la fórmula: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{n}$

Donde x_i es la marca de clase

Desvío estándar:

Un problema de la varianza es que se pierden las unidades naturales de la variable. Esto es, si se tratan de alturas la varianza se expresa en m2, si se trata de pesos, en kg^2, lo cual dificulta su interpretación. Por lo tanto, si definimos el desvío estándar de los datos como: $\sigma = \sqrt{\sigma^2}$, podemos interpretarlo como el promedio de las desviaciones respecto de la media. Una variable con mayor desvío es menos homogénea

Desvío Medio:

Si en vez de aplicar la transformación cuadrática a los datos aplicamos la función módulo, obtenemos:

$$D.M = \frac{\sum_{i=1}^{N}|x_i - \mu|}{N}$$

El desvío medio puede calcularse también respecto de la mediana, y de la moda.

Tiene la misma interpretación que el desvió estándar, solo que se utilizan desvíos absolutos para salvar el hecho de que su suma es cero. Es esperable que su resultado sea similar al desvío

estándar, salvo que existan valores demasiado alejados de la media (en este caso la varianza será sensiblemente mayor y por lo tanto el desvío también).

MAD: Surge de calcular la mediana de los desvíos absolutos respecto de la media. Es aceptado que este valor es mucho más robusto que el desvío, y debe preferirse cuando se encuentran valores atípicos.

Coeficiente de variación:

Se define como la proporción de desvío respecto de la media. Es decir, el cociente entre el desvío estándar y la media aritmética. Es una medida "a dimensional" por lo tanto es de fácil interpretación. $c.v = \frac{\sigma}{\mu}$

Medidas de posición: Las medidas de posición nos dan información entre otras cosas acerca de la forma de la distribución de los datos.

Percentil i-esimo:

Se define como el valor tal que si se orden los datos de forma ascendente, acumula el i% de los datos. Si se calcula por interpolación lineal:

$p_i = L_i + \frac{\frac{n}{100} i - F_{i-1}}{f_i} C$ Donde i es el numero de percentil ; L_i es el limite inferior del intervalo que contiene al percentil i-esimo; F_{i-1} es la frecuencia acumulada del intervalo anterior que contiene al percentil; f_i es la frecuencia del intervalo en el que se encuentra el percentil y C es el ancho de clase que lo contiene.

Cuartiles:

Definimos el cuartil i-esimo, de forma análoga, como el dato que acumula la i-cuarta parte del total de datos ordenados.

$Q_1 = L_i + \frac{\frac{n}{4} - F_{i-1}}{f_i} C$ Donde L_i es el límite inferior del intervalo que contiene al cuartil i-esimo; F_{i-1} es la frecuencia acumulada del intervalo anterior que contiene al cuartil; f_i es la frecuencia del intervalo en el que se encuentra el cuartil y C es el ancho de clase que lo contiene.

A partir de las medidas de posición pueden definirse los intervalos percentil 10-90 y rango semiintercuartílico que dan idea acerca de la homogeneidad de los datos.

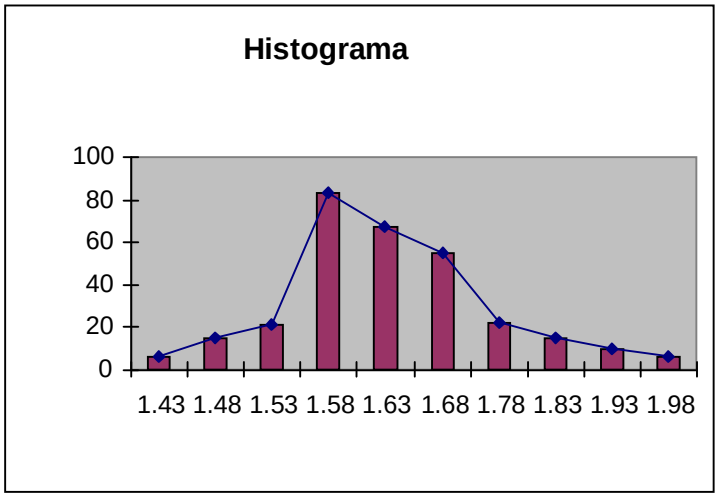
El rango percentil 10-90, es la diferencia entre el 90 y 10 percentil, es un intervalo que contiene el 80% de los datos centrales.

Análogamente, podemos definir el rango semiintercuartílico como el rango de datos que acumula el 50% de los datos.

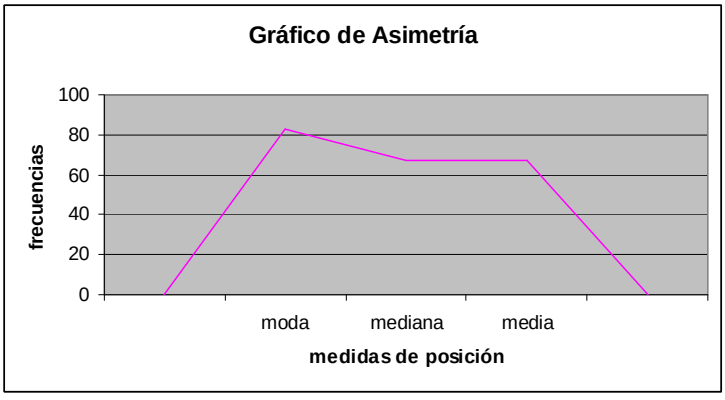
Medidas de resumen gráficas

Al analizar el histograma para los datos de alturas presentados en la tabla de alturas, podemos obtener información relevante sobre la forma en que los datos se distribuyen:

Análisis de la simetría:



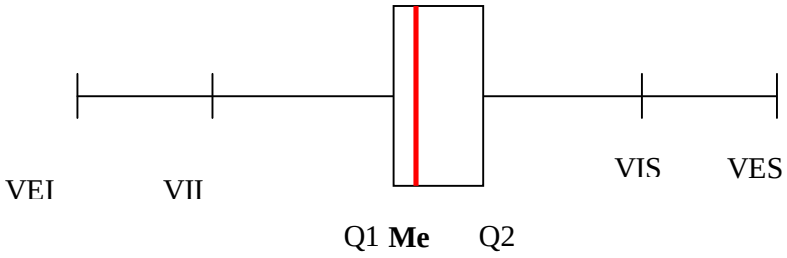
El análisis de las tres medidas de tendencia: Media, mediana y Moda, pueden dar idea de la asimetría de la distribución. Diremos que si las tres medidas son similares entonces, la distribución es simétrica, en otro caso es asimétrica.



Otra información relevante que se puede extraer del gráfico es la correspondiente a "apuntalamiento" de la distribución: Si la distribución de los datos es acampanada, se dice que los datos presentan una distribución mesocúrtica. En el caso en que los datos presenten una distribución de frecuencias aplastada y con colas pesadas, se dirá que la distribución es platicúrtica, y en caso de una distribución apuntalada de colas finas, diremos que es leptocúrtica.

El Box Plot, o diagrama de tallo y Bigotes:

El Box Plot es una forma alternativa de graficar ciertos aspectos salientes de un conjunto de datos. La información más relevante que puede brindarnos el Box- Plot es la homogeneidad/heterogeneidad de los datos, la presencia de valores influyentes u outliers, y la asimetría de la distribución. Para construir el diagrama de caja y bigotes, se parte de graficar una caja con los cuartiles primer y tercero y la mediana:



Para construir los brazos, se calcula la distancia intercuartil: dQ=Q2- Q1
Se construyen las vallas:
La valla interna inferior: VII=Q1-1.5dQ
La valla interna superior: VIS=Q2+1.5dQ
La valla externa inferior: VEI=Q1-3dQ
La valla externa superior: VES=Q2+3dQ

El Box-Plot nos permite clasificar las observaciones de una manera clara: Los valores que se encuentra entre las vallas internas y externas, son outliers, y los que se encuentra más allá de las vallas externas, son outliers severos.
La forma de la caja, nos permite clasificar la distribución en simétrica o asimétrica en función de la forma ubicación relativa respecto a los datos.

Actividad 8

1) Dada la siguiente serie de datos

20, 28, 27, 0, 0, 9, 27, 6, 5

a) Calcula medidas de tendencia central y medidas de variabilidad.

b) Construye un diagrama de tallo-hoja.

2) Completa la serie de frecuencias, sabiendo que corresponde a una serie de datos de 180 elementos:

Xi	F	fr	f%
2	45		25
2,8		23/180	
3			
3,3			10
3,4	5		
4		54/180	
5			5
5,01		15/180	

3) Basándose en el ejercicio anterior, calcula:
a) La media, la mediana y la moda.
b) El percentil 12, 16 y 76.

4) En el banco XXX $.A, las tasas de interés pasiva mensual fue:

Mes	Tasa
Enero	1,5%
Febrero	1%
Marzo	1,2%
Abril	0,9%
Mayo	3,3%
Junio	1,5%
Julio	1,4%
Agosto	1,45%
Septiembre	1,7%
Octubre	1,8%
Noviembre	1,6%
Diciembre	1,9%

Luis depositó durante todo el año anterior cierta cantidad de dinero en el banco XXX $A. ¿Cuál es la tasa media a la que estuvo depositado el dinero?
a) Si se capitalizó como interés simple.
b) Si se capitalizó como interés compuesto
c) Construye un Box- Plot para los datos. ¿Qué conclusiones sacas?

5) Una consultora de consumo masivo realizó una encuesta a 700 adolescentes para conocer sus hábitos de consumo. Frente a la pregunta ¿Cuántos litros de gaseosa tomas por semana?, se recogieron los siguientes datos:

Litros semanales	Cantidades de adolescentes
Nada	30
Menos de 3	45
Entre 3 y 6	150
Entre 6 y 9	300
Entre 9 y 12	130
Entre 12 y 15	18
Entre 15 y 18	15
Más de 18	12

a) Realiza un histograma sobre los datos.
b) ¿Cuál es la cantidad máxima de gaseosa que toman el 15% de los adolescentes que menos toman?
c) ¿Cuál es la cantidad mínima de gaseosa que toman el 70% de los que mas toman?
d) ¿Qué porcentaje de adolescentes toman entre 1,5 y 17 litros de gaseosa?

6) En una heladería, el 15% de los clientes toman un helado por semana, el 50% dos helados por semana, y del resto, el 15% toma cuatro mientras que los demás toman 5.
a) Calcula la media de helados que toman los clientes por semana.
b) Calcula el desvío de la cantidad de helados que toman los clientes de la heladería.
c) ¿Es homogénea la variable?

7) De la siguiente serie de datos, calcula media aritmética, media armónica, geométrica, y media recortada.

16, 17, 17, 200, 15, 14, 13, 23, 23, 21, 30, 23, 18, 17, 16

8) Para el caso del Ítem 7, da un intervalo para la media de los datos.

9) Dibuja una distribución leptocúrtica, una platicúrtica y una mesocúrtica. Encuentra un ejemplo de la vida real para cada una de ellas.

10) Para el conjunto de datos del Ítem 8, calcula la mediana y el MAD. ¿Qué conclusiones puedes sacar?