Bloque I.A (primera semana)

Contenidos Conceptuales: Estadística descriptiva: distribuciones de frecuencia. Representaciones Graficas: histogramas, diagramas circulares, diagramas de barras, boxplots. Medidas resumen: tendencia central y dispersión.

La estadística provee métodos para organizar y sintetizar datos y para hacer inferencias sobre una población en base a información contenida en una muestra de la misma. Bajo incertidumbre, la estadística provee métodos para tomar decisiones.

En general un estudio se enfoca en un conjunto de objetos o individuos, a los que denominamos población.

Es decir que la población es la totalidad de objetos o individuos de interés en cierto estudio.

Los siguientes son ejemplos de poblaciones de interés:

- Establecimientos industriales del Gran Buenos Aires
- Ruedas de cierto tipo fabricadas por una empresa
- Individuos que padecen cierta enfermedad

Cuando se dispone de la información acerca de todos los individuos u objetos de una población, se está en presencia de un **censo**, mientras que cuando se cuenta con información sólo de un subconjunto de la población se tiene una **muestra**.

En general, estamos interesados en ciertas características de los individuos y objetos de una población. Por ejemplo, para las poblaciones ejemplificadas podriamos estar interesados en:

a. número de empleados de cada establecimiento industrial del Gran Buenos Aires
b. duración en meses de las ruedas de cierto tipo fabricadas por las empresas

c. presencia o ausencia de ciertos síntomas en individuos que padecen la enfermedad
d. distancia entre ciudades de cierta ruta comercial aérea.

Se denominan variables a aquellas características o atributos que cambian de un objeto o individuo a otro.

Una variable puede ser **categórica o numérica**.
Ejemplos de Variables categóricas *son: el* sexo, el estado civil, la profesión del padre, etc.
Ejemplos de Variables numéricas*:* edad, duración de la carrera, número de empleados.

Las variables numéricas o cuantitativas pueden ser de dos tipos:

- *DISCRETAS:* toman un número finito o infinito numerable de valores. (en general asociadas al proceso de 'CONTAR'), tales como número de hijos, materias aprobadas, cantidad de fallas en un lote, etc.
- *CONTINUAS:* toman, o pueden tomar, un número infinito no numerable de valores cualquier valor en un intervalo. (en general asociadas al proceso de 'MEDIR'), tales como estatura de un individuo, duración de una consulta, superficie afectada por un fenómeno, etc-

Si sobre cada objeto o individuo se observa una única variable se obtiene un dato univariado. En cambio, si se observan dos o más características se tiene un dato multivariado.

Ejemplos de datos univariados podrían ser:

- el número de empleados de un establecimiento industrial.
-registro la presión arterial de un paciente después de su almuerzo.
Ejemplos de datos multivariados:
-para cada establecimiento se registran:
(número de empleados, horas trabajadas en un mes, salario medio).

-para cada individuo se le registran:

(edad, sexo, presión arterial, peso).

Observación: un dato multivariado puede tener en sus componentes variables de diferente especie, es decir categórica en una posición y cuantitativa en otra o discreta en una posición y continua en otra, etc.

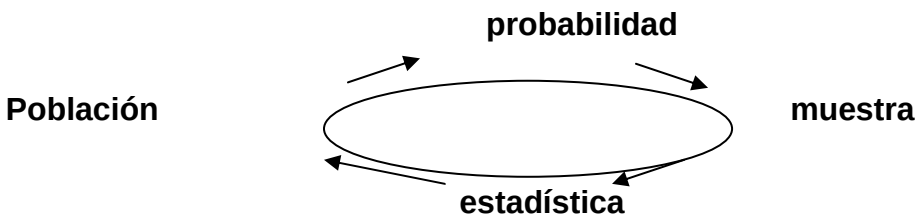
Relación entre las ramas de la estadística

La estadística descriptiva permite organizar y describir conjunto de datos cualitativos o cuantitativos (categóricos o numéricos).
Una vez obtenida una muestra, generalmente el investigador desea usar esa información para llegar a conclusiones, es decir para hacer inferencias acerca de la población.

Para ello se usa métodos de inferencia estadísticas. Los métodos pueden clasificarse en:

-estimación puntual
-pruebas o tests de hipótesis
-estimación por intervalos de confianza

Entre la estadística descriptiva y la inferencial hay un puente, que es la Probabilidad. Tanto la estadística como la probabilidad tratan con poblaciones y muestras, pero lo hacen de manera inversa.



En probabilidad se suponen conocidas las propiedades de la población y se formulan y responden preguntas sobre la muestra.

En Estadística, se dispone de la característica de la muestra y se desea utilizar estos datos para sacar conclusiones sobre la población.

Podríamos preguntarnos: ¿por qué se estudia Probabilidad antes que Estadística?
La respuesta es que para comprender qué nos dice una muestra respecto a la población, debemos conocer la incertidumbre asociada con el hecho de haber tomado una muestra.

Hay poblaciones que en realidad no existen. Por ejemplo, supongamos que para estudiar la precisión de un nuevo instrumento, se realizan 20 mediciones de cierta magnitud conocida. ¿Cuál sería la población de la cual proviene esta muestra? Sería el conjunto de todas las mediciones que podrían hacerse en condiciones experimentales similares. A este tipo de poblaciones se la denomina conceptuales o hipotéticas.

Estadística Descriptiva:

Utiliza dos tipos de métodos:

-gráficos y tabulares

-numéricos

Métodos gráficos y visuales:

Permiten representar un conjunto de datos empleando técnicas visuales que favorezcan el análisis.
Notación: en general el número de observaciones de un conjunto se indica con la letra **n** y las observaciones individuales por $x_1, x_2, ..., x_n$. La i-ésima observación se designa con x_i.(Por supuesto se puede utilizar cualquier otra letra, por ejemplo: $y_1, y_2, ..., y_n$).
Entre los métodos gráficos más usuales se encuentran:

Histograma :

Antes de ver cómo se construye necesitamos poder elaborar una distribución de frecuencias para datos cuantitativos.

Si la variable en estudio es discreta, es decir si el número de valores posibles es finito o infinito numerable, se define la frecuencia o frecuencia absoluta de la observación x como la cantidad de veces que se presenta.

$$\text{Frecuencia relativa} = f_r = \frac{\text{cant. de veces que aparece x}}{\text{numero de observaciones}}$$

El patrón de variación de una variable se denomina distribución. La distribución registra los valores numéricos de la variable y cuan frecuentemente ocurre cada uno.

Una *distribución de frecuencias* se representa mediante una tabla de las frecuencias y/o las frecuencias relativas correspondientes a cada dato registrado.

Pasos para construir el histograma:

-Se marcan los valores de x sobre una escala horizontal.

-Arriba de cada valor se construye un rectángulo cuya altura sea igual a la frecuencia relativa y tenga por base 1.

Observación: De esta forma el área total (suma de las áreas de los rectángulos graficados) será 1.

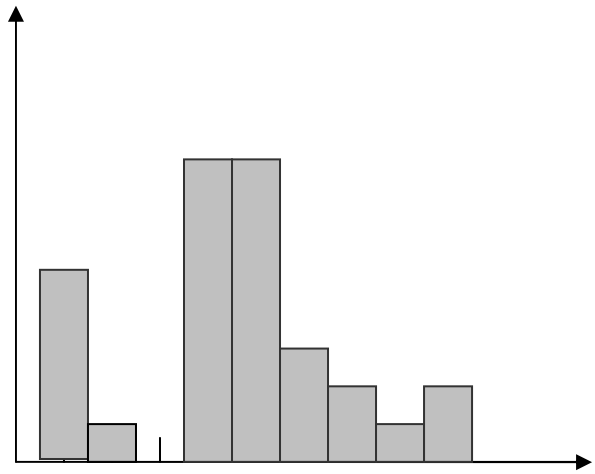
Ejemplo 1:

Dadas las siguientes 30 observaciones, elaborar la tabla de frecuencias y el histograma correspondiente:

1	1	1	1	1	2	4	4	4	4	4	4	4	4	5
5	5	5	5	5	5	5	6	6	6	7	7	8	9	9

Tabla de frecuencias:

valor	*frecuencia*	*Frecuencia relativa*
1	5	5/30 = 0.167
2	1	1/30 = 0.033
3	0	0/30 = 0
4	8	8/30 = 0.0267
5	8	8/30 = 0.0267
6	3	3/30 = 0.1
7	2	2/30 = 0.067
8	1	1/30 = 0.033
9	2	2/30 = 0.067
	N=30	suma de las frec.rel=1



1 2 3 4 5 6 7 8 9

En el caso de una variable continua, se subdivide el eje de las mediciones en una cantidad adecuada de intervalos de clase, o clases, de modo tal que cada observación esté contenida en uno y sólo un intervalo.

Posteriormente se calcula la frecuencia relativa para cada clase y se elabora el histograma como antes, se construye sobre cada intervalo de clase un rectángulo, cuya área sea la frecuencia relativa (para que el área total sea 1)

Ejemplo 2:

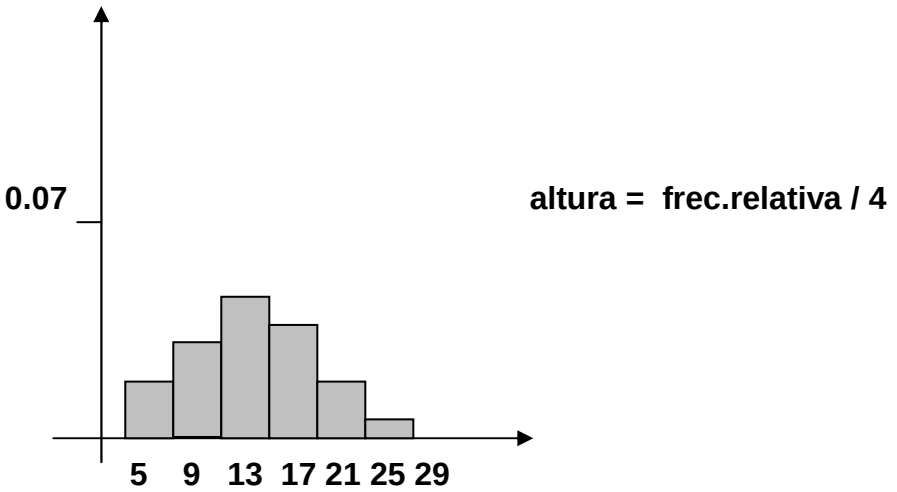
Relación precio-rendimiento de 25 acciones (price-earnings ratios)

20.5	19.5	15.6	24.1	9.9
15.4	12.7	5.4	17.0	28.6
16.9	7.8	23.3	11.8	18.4
13.4	14.3	19.2	9.2	16.8
8.8	22.1	20.8	12.6	15.9

Valor	Frecuencia	Frecuencia relativa
5 – 8.99	3	3/25 = 0.12
9 – 12.99	5	5/25 = 0.20
13 – 16.99	7	7/25 = 0.28
17 – 20.99	6	6/25 = 0.24
21 – 24.99	3	3/25 = 0.12
25 – 28.99	1	1/25 = 0.04
	N = 25	Suma de las frec. Relativas = 1

No hay reglas “óptimas” para seleccionar el número de intervalos. En general, se recomienda usar entre 5 y 20 intervalos. J.L. Devore sugiere:

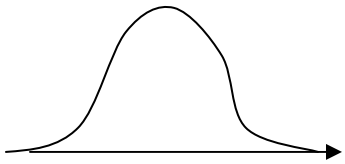
$$cantidad\ de\ clases \approx \sqrt{cantidad\ observaciones}$$



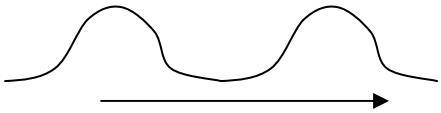
Los intervalos no tienen porqué ser de igual longitud. Si, por ejemplo, hay valores atípicos puede ser conveniente usar intervalos de distinta longitud.

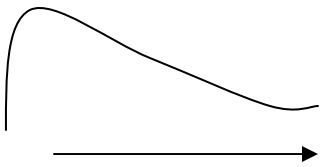
Formas usuales de histogramas:

Unimodal

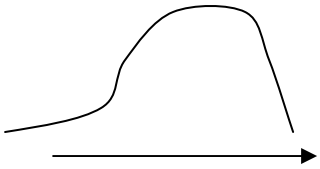


Bimodal simétrico





asimétrico por la derecha



asimétrico por la izquierda

Histogramas para Datos cualitativos:

Si la naturaleza del conjunto de datos es cualitativa, se puede construir un histograma. En estos casos, en general, **no hay un orden** natural para las clases e intervalos deben elegirse de igual longitud.

Ejemplo 3:

A cada miembro de una muestra de 120 dueños de motocicletas, se les pregunta la marca de su máquina. Se obtiene la siguiente tabla de frecuencias :

Marcas	Frecuencia	Frec. Relativa
Honda	41	0.34
Yamaha	27	0.23
Kawasaki	20	0.17
Suzuki	18	0.15

Harley Davidson	3	0.03
Otras	11	0.09
	N = 120	1 (por redondeo: 1.01)

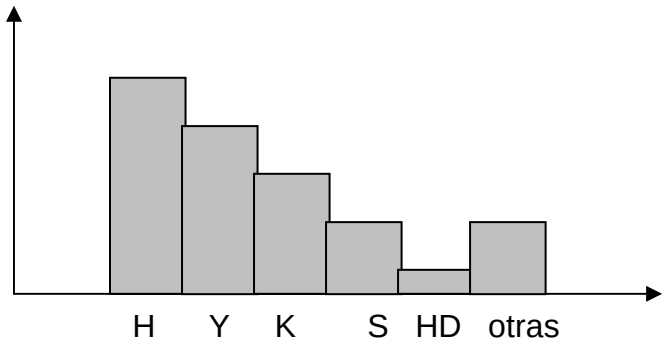
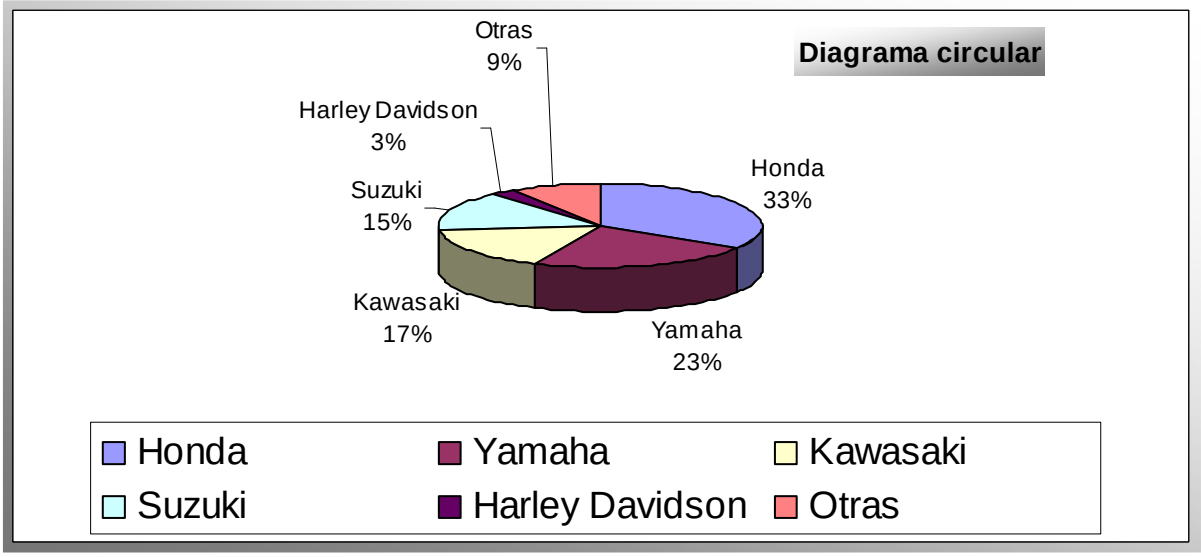


Diagrama circular

Esta distribución de frecuencias también puede representarse mediante un **diagrama circular**. Esta representación es muy frecuente en los medios masivos de comunicación dada su sencillez de interpretación.

Se representa un sector circular de amplitud proporcional a la frecuencia relativa:

Marcas	Frecuencia	Frec. Relativa	Sector circular
Honda	41	0.34	0.34*360=122°24’
Yamaha	27	0.23	0.23*360=82°48’
Kawasaki	20	0.17	0.17*360=61°12’
Suzuki	18	0.15	0.15*360=54°
Harley Davidson	3	0.03	0.03*360=10°48’
Otras	11	0.09	0.09*360=32° 24’
	N = 120	1 (por redondeo: 1.01)	



propiedades sobresalientes.

Por ahora nos referiremos a datos numéricos. Supongamos que nuestro conjunto de datos es $x_1, x_2, \ldots, x_n$.

Una de las características de interés es localizar el centro de estos datos.

Este centro se llama Media o Promedio muestral.

1.-Media o Promedio muestral:

Se nota $\overline{x}$ y se lo define y calcula como: $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Interpretación física de la media:

Si sobre un eje horizontal representamos con un punto los valores de las observaciones y colocamos sobre cada uno de ellos, dependiendo del punto con un hilo, un peso de una libra; observaremos que el eje horizontal se

mantiene en equilibrio si se lo sostiene en el punto correspondiente a $\bar{x}$. Dicho en otras palabras, $\bar{x}$ es el centro de gravedad de nuestra variable.

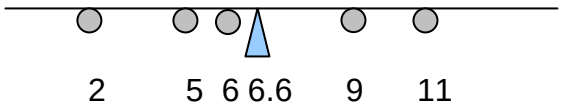
Ejemplo 4:

Consideremos las siguientes observaciones: 2 – 9 – 11 – 5 – 6

$$\bar{x} = \frac{\sum_{i=1}^{5} X_i}{5} = \frac{2+9+11+5+6}{5} = 6.6$$

Así como $\bar{x}$ representa el promedio de las observaciones de la muestra, podría pensarse en el promedio de todos los valores de la población.

Este promedio se denomina media poblacional y se lo denota con la letra griega μ.



Cuando se trata de una población finita, de tamaño N, entonces la media poblacional es: $\mu = \frac{\sum_{i=1}^{n} x_i}{N}$

La media muestral posee una propiedad que la hace no satisfactoria como medida de posición en algunos casos. El valor de $\bar{x}$ puede verse afectado por la presencia de una o unas pocas observación /es alejadas de los datos restantes.

Dicho de otra forma es muy sensible a la presencia de valores extremos, salvajes o outliers. Los especialistas la caracterizan como un resumen no robusto.

Ejemplo 5:

Sean: $x_1 = 47, x_2 = 46, x_3 = 40, x_4 = 57, x_5 = 50.$ Entonces $\bar{x} = 48$

En cambio si : $x_1 = 47, x_2 = 46, x_3 = 40, x_4 = 57, x_5 = 200$. Entonces $\bar{x} = 78$

Propiedades de la media:

Sean $x_1, x_2, \ldots, x_n$ un conjunto de observaciones y c una constante no nula:

a.- Si $y_1 = x_1 + c;\ y_2 = x_2 + c; \ldots ;\ y_n = x_n + c \Rightarrow \bar{y} = \bar{x} + c$

b.- Si $y_1 = c.x_1 ;\ y_2 = c.\ x_2 ;\ldots;\ y_n = c.\ x_n \Rightarrow \bar{y} = c.\bar{x}$

En muchos conjuntos de datos pueden presentarse valores atípicos o inusuales, por ejemplo si tomamos distribuciones de ingresos por sueldos. Existe una medida de localización que es menos sensible a la presencia de valores atípicos, es decir más robusta, se denomina mediana.

2.- Mediana:

Sean $x_1, x_2, \ldots, x_n$ las observaciones. Se ordenan de menor a mayor y se define la mediana muestral y se nota $\tilde{x}$ como:

$$\tilde{x} = \begin{cases} \text{valor medio o central} & \text{si n es impar} \\ \text{promedio de las dos observaciones centrales} & \text{si n es par} \end{cases}$$

Así, si con $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ notamos la muestra ordenada, calculamos la mediana

como: $\tilde{x} = x_{(k+1)}$ si n = 2k + 1 ó $\tilde{x} = \dfrac{x_{(k)} + x_{(k+1)}}{2}$ si n = 2k

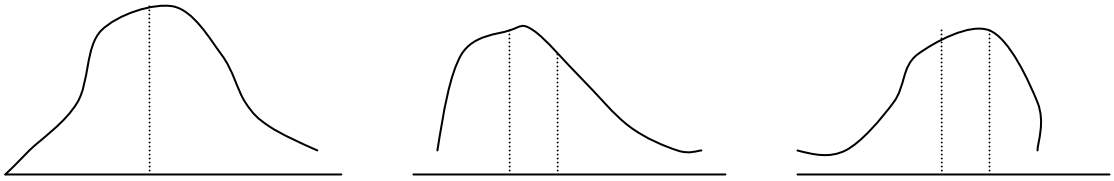
Ejemplo 6:

Sean : $x_1 = 47, x_2 = 46, x_3 = 40, x_4 = 57, x_5 = 50.$ Entonces $\bar{x} = 47$

Si : $x_1 = 47, x_2 = 46, x_3 = 40, x_4 = 57, x_5 = 200$. Entonces $\bar{x} = 47$

Al igual que en el caso de la media, podemos pensar en una mediana poblacional a la que notaremos $\tilde{\mu}$.

En general μ y $\tilde{\mu}$ no son iguales. Si esto ocurre se debe a la simetría de la distribución.



$\mu = \tilde{\mu}$ $\quad$ $\tilde{\mu} < \mu$ $\quad$ $\mu < \tilde{\mu}$

Simétrica $\quad$ asimetría por la derecha $\quad$ asimetría por la izquierda

3.-Media α - podada ó α - recortada:

Esta medida es un resumen intermedio entre la media y la mediana. Para calcularla, se ordena la muestra y se poda el 100α% de las observaciones en cada extremo. Luego se promedian los valores restantes.

$\overline{x}$ equivale a la media podada con $\alpha = 0$

$\tilde{x}$ equivale a la media podada con el valor máximo posible de α

Ejemplo 7:

Sean las siguientes observaciones ordenadas:

590; 612; 623; 666; 777; 883; 898; 964; 970; 983.

Si queremos obtener la media podada al 10 % (α=0.10) $\Rightarrow$ como n = 10 resulta 100α% (10%) de 10 es 1, de donde se debe eliminar una observación en cada extremo y se promedian las restantes. Así :

$$\bar{x}_{0.10} = \bar{x} = \frac{612+623+666+777+883+898+964+970}{8} = 799.125$$

Nota : Si n.α no es entero, por ejemplo, n = 24 y α = 0.10, resulta n.α = 2.4, se calculan dos medias podadas, una podando dos valores en cada extremo y otra podando tres valores y se interpola linealmente entre ambas.

Datos categóricos y porciones muestrales:

En el caso de datos categóricos la distribución de frecuencias ya proporciona un adecuado resumen de la información, pues nos indica el número de casos en cada clase y la proporción en cada clase (frecuencia relativa x 100).

En el caso de una población dicotómica (sólo dos categorías), dada una muestra de tamaño n, podemos denotar con x al número de casos en la primera de las categorías, por lo tanto n-x será el número de casos en la segunda.

La frecuencia relativa de la categoría 1 será x/n y la de la categoría 1 será entonces $\frac{n-x}{n} = 1 - \frac{x}{n}$.

Si asignáramos un 1 a cada observación que pertenece a la categoría 1 y un cero a la que pertenece a la categoría 2 entonces $\bar{x} = \frac{x_1, x_2, \ldots, x_n}{n} = \frac{x}{n}$. O sea que la media muestral es la proporción muestral.

4.- Moda o Modo

Es el valor de varibale correspondiente a la mayor frecuencia observada. En el caso de distribuciones simétricas coincide con la media y con la mediana. Su

obtención es en general muy sencilla y es aplicable a datos de cualquier tipo, incluso categóricos, sin embargo; el valor modal es informativo cuando es único. Pero existen distribuciones bimodales y hasta multimodales para las cuales el modo no provee uan buena información.

.

Ejemplo 8:

a) datos categóricos: Mo=Honda

Marcas	**Frecuencia**	**Frec. Relativa**
Honda	41	0.34
Yamaha	27	0.23
Kawasaki	20	0.17
Suzuki	18	0.15
Harley Davidson	3	0.03
Otras	11	0.09

b) variable cuantitativa discreta unimodal

valor	*frecuencia*
1	5
2	1
3	1
4	9
5	8
6	3
7	2
8	1
9	2

En esta distribución la Mo=4

c) Variable cuantitativa discreta bimodal

valor	*frecuencia*
1	5
2	1

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

Esta distribución es bimodal $Mo_1=4$ y $Mo_2=5$

Medidas de variabilidad:

Ninguna medida de localización puede dar un resumen completo de los datos. Distintas muestras o poblaciones pueden tener la misma posición (medida adecuadamente) y sin embargo, diferir en otros aspectos esenciales.

Ejemplo 9:

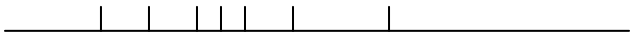
Consideremos los siguientes dos conjuntos de datos:

x_i	40	120	20	80	90
y_i	80	40	100	80	50

En ambos: $\bar{x} = \bar{y} = 70$; $\tilde{x} = \tilde{y} = 80$

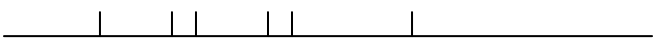
Sin embargo se observa fácilmente, que un conjunto tiene mayor dispersión que el otro.

x_i :



0 20 40 50 60 80 120

y_i :



0 20 30 50 60 80

Presentemos la primera medida de "dispersión" o variabilidad es el rango.

1.- Rango muestral:

Se define como la diferencia entre el valor máximo muestral y el valor mínimo. En símbolos, si $x_1, x_2, \ldots, x_n$ son las observaciones de la muestra: r =máx(x_i) - mín (x_i).

Con nuestra notación: r = $x_{(n)} - x_{(i)}$

En nuestro ejemplo: r_x = 100 y r_y = 60.

Esta medida de variabilidad parecería adecuada, pero resulta insuficiente.

Ejemplo 10 :

x_i	10	15	15	15	20
y_i	10	11	15	19	20

Ahora:

$\bar{x} = \bar{y} = 15$; $\tilde{x} = \tilde{y} = 15$; $r_x = r_y = 10$

Sin embargo hay mayor variabilidad en el segundo conjunto que en el primero.

Necesitamos una medida que tenga en cuenta a más valores y no sólo los valores extremos. Podría ocurrir que el rango de dos variable fuera el mismo pero que mientras la mayoría de los valores de una se encuentra próximo a su centro, la mayoría de los valores de la otra se encuentran alejados de su centro, incluso si tienen el mismo centro.

2.- Desviaciones respecto de la media:

Se llama i-ésima desviación respecto de la media a la diferencia $d_i = x_i - \bar{x}$.

Si $x_i > \bar{x} \Rightarrow x_i - \bar{x} > 0$. Si $x_i < \bar{x} \Rightarrow x_i - \bar{x} < 0$.

Si la magnitud de todas las desviaciones es pequeña $\Rightarrow$ todas las observaciones x_i se encuentran cerca de la media y hay poca variabilidad.

Una forma sencilla de combinar todas las desviaciones en una única cantidad sería promediarlas, pero: $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})=0$.

Necesitamos, pues, convertir los desvíos en valores positivos, o mejor dicho, no negativos, antes de promediar.

Una forma es usar $\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{x}\right|$ (desviación absoluta promedio), pero por dificultad de cálculo, es preferible usar $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ ó bien $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$.

3.- Varianza muestral y Desvío standard muestral:

Sean $x_1, x_2, \ldots, x_n$ un conjunto de observaciones y $\overline{x}$ su promedio muestral $\Rightarrow$

1. –la varianza muestral se define como: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$
2. –el desvío standard muestral, notado con s, como la raíz cuadrada positiva de la varianza $\therefore s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$
3. –Designaremos con σ^2 a la correspondiente varianza poblacional. Si la población es finita $\Rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2$. Observar que no se corrige el denominador como se hace en el desvío muestral.

Nota :

Es necesario corregir el denominador en el desvío muestral pues como μ es en general desconocida, se la debe estimar con el valor de $\overline{x}$. Los valores de las observaciones muestrales están más próximos a $\overline{x}$ que a μ. Por lo tanto si dividiéramos por n estaríamos “subestimando” la varianza.

Al denominador de s^2 se lo suele llamar “grados de libertad”. Esta terminología resulta del hecho que si bien s^2 está basada en n diferencias $x_i-\overline{x}$,como la

suma de ellas es igual a cero, resulta que especificando los primeros n-1 valores el último queda determinado.

Ejemplo 11 :

n=4 Si tomamos a: $x_1 - \bar{x}$ =8; $x_2 - \bar{x}$ =-6; $x_3 - \bar{x}$ = -4 $\Rightarrow x_4 - \bar{x} = 2$

Sólo 3 de las desviaciones pueden elegirse libremente, por eso tenemos es este caso, 3 grados de libertad.

Fórmula para calcular s ² :

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Propiedades de la varianza y del desvío estándar muestral:

Sean $x_1, x_2, \ldots, x_n$ un conjunto de observaciones y c una constante no nula:

a. –Si $y_1 = x_1 + c\,; \quad y_2 = x_2 + c\,; \ldots; y_n = x_n + c \Rightarrow s_y^2 = s_x^2$ y $\Rightarrow \quad s_y = s_x$

b. - Si $y_1 = x_1 c\,; \quad y_2 = x_2 c\,; \ldots; y_n = x_n c \Rightarrow s_y^2 = s_x^2$ y $\Rightarrow \quad s_y = |c|\, s_x$