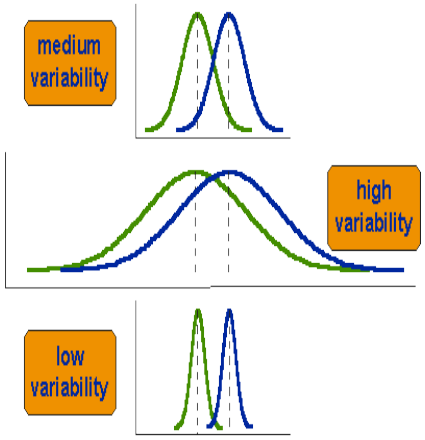
Medidas de variabilidad

Oscar Centeno Mora

Preámbulo.....

- En capítulos anteriores se estudiaron temas como los números relativos y medidas de posición como formas de resumir la información.



- La presente clase cierra el ciclo de las medidas de estadísticas descriptivas para una variable.

- Como se señaló en el tema de medidas de posición, una medida de este tipo SIEMPRE debe estar acompañada de una medida de variabilidad.

- Se expondrán las principales medidas de variabilidad.

Preámbulo.....

- El análisis anterior sobre clientes para una determinada catera del banco, se determinó con el conocimiento y la certeza que los resultados describieron o aproximaron de forma correcta lo que se le solicitó.

- Sin embargo, alguien le pregunta: ¿con que tanta certeza podría dar por válidos los restados? Hoy podrá responder a dicha pregunta…

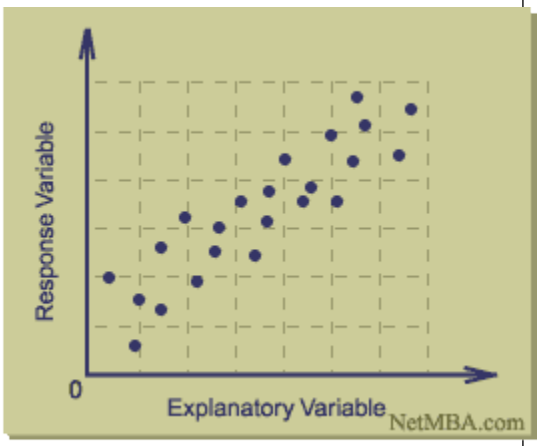
Índice

Índice

Índice

Introducción

- El concepto de variabilidad juega un papel clave dentro de la Estadística como en cualquier ciencia.

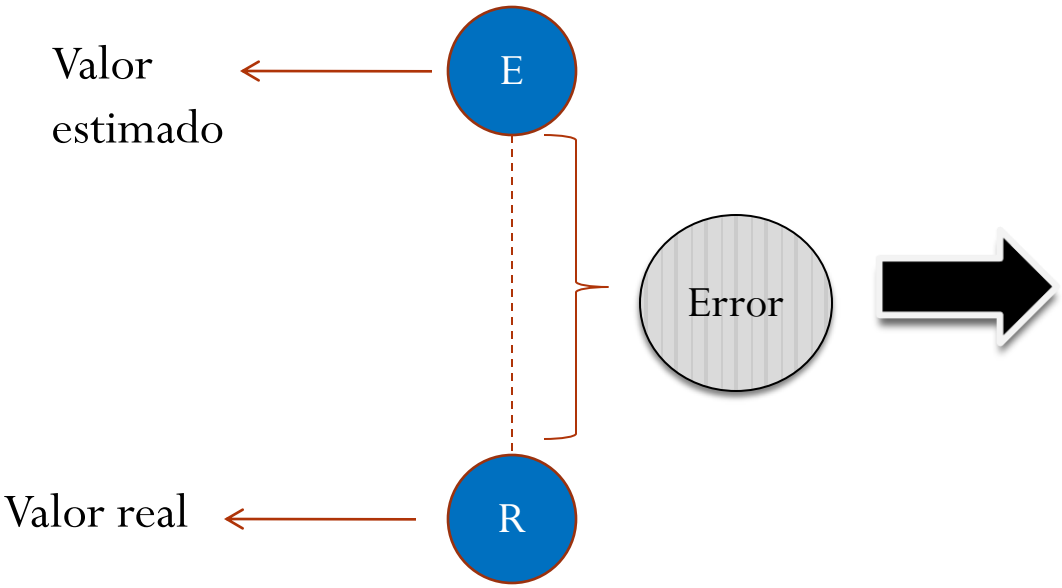


- La razón de ser de la estadística reside en que los datos son variables de mayor a menor intensidad.

- La importancia de la Estadística es suministrar procedimientos válidos y confiables para analizar los hechos variables.

Introducción

- El objetivo está en poder explicar y mediar la variabilidad con tal de realizar inferencias fidedignas.



El objetivo en la estadística es poder explicar con cierto grado de confianza la diferencia entre el valor estimado y el verdadero

Introducción

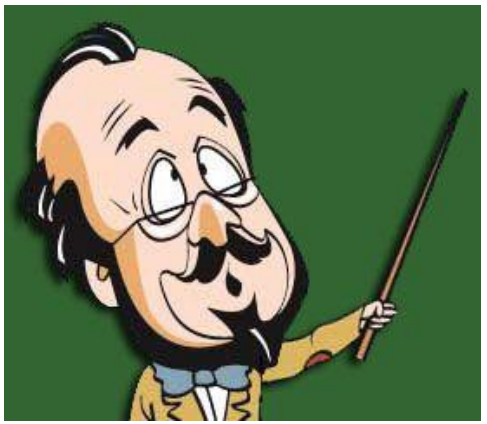
- Para los investigadores, la variabilidad es un fenómeno natural y corriente del cual hay clara conciencia:

-Economista: la predicción del gasto en el gobierno.

-Ingeniero Industrial: control y calibración de un proceso de producción de carros.

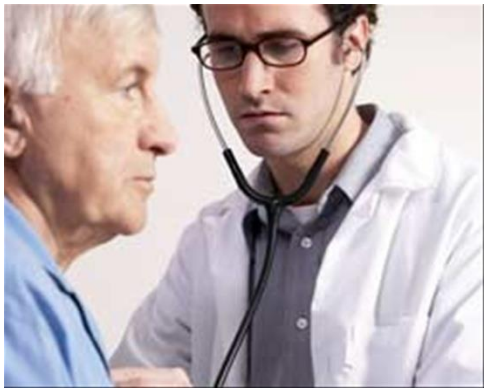
Introducción

-Profesor: rendimiento académico de los alumnos.



-Agrónomo: cuál de dos variedades fertilizante contiene un mayor rendimiento .

-Médico: establece cuál tratamiento es mejor para la rehabilitación del paciente.



Índice

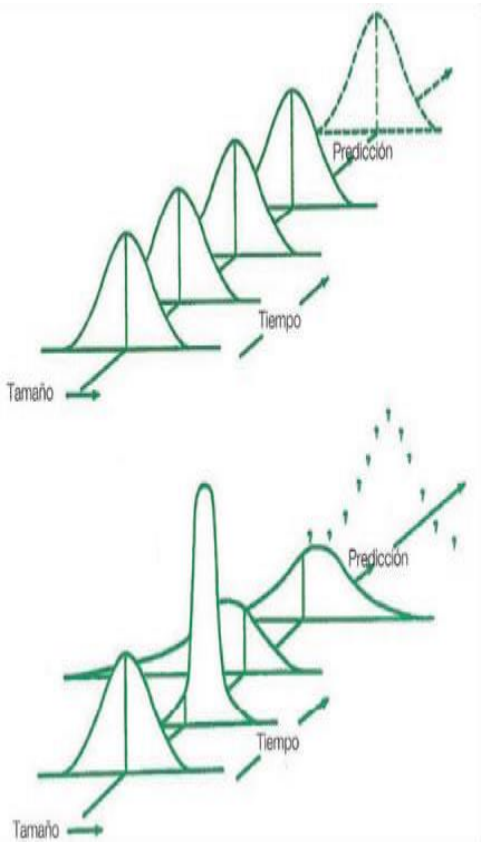
El concepto de variabilidad

- Cuando se analizan los datos, se debe tener en mente 2 objetivos:

  1. Se quiere conocer el fenómeno en cuestión, y la mayoría de veces se utilizan medidas de posición (MDP).

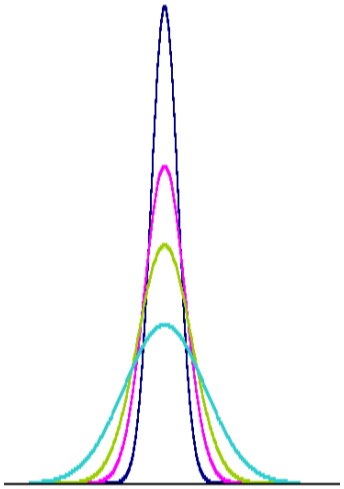
  2. Se utilizan medidas para poder medir la variabilidad en los dato. Por lo tanto se quiere saber tanto la concentración como la dispersión de estos.

- De acuerdo con el segundo objetivo, si existe la presencia de mucha variabilidad, entonces la confianza de los datos estará en juego.



El concepto de variabilidad

- A la hora de analizar un conjunto de datos, es tan importante conocer las MDP (preferiblemente el promedio), como las medidas de variabilidad (la *desviación estándar*).

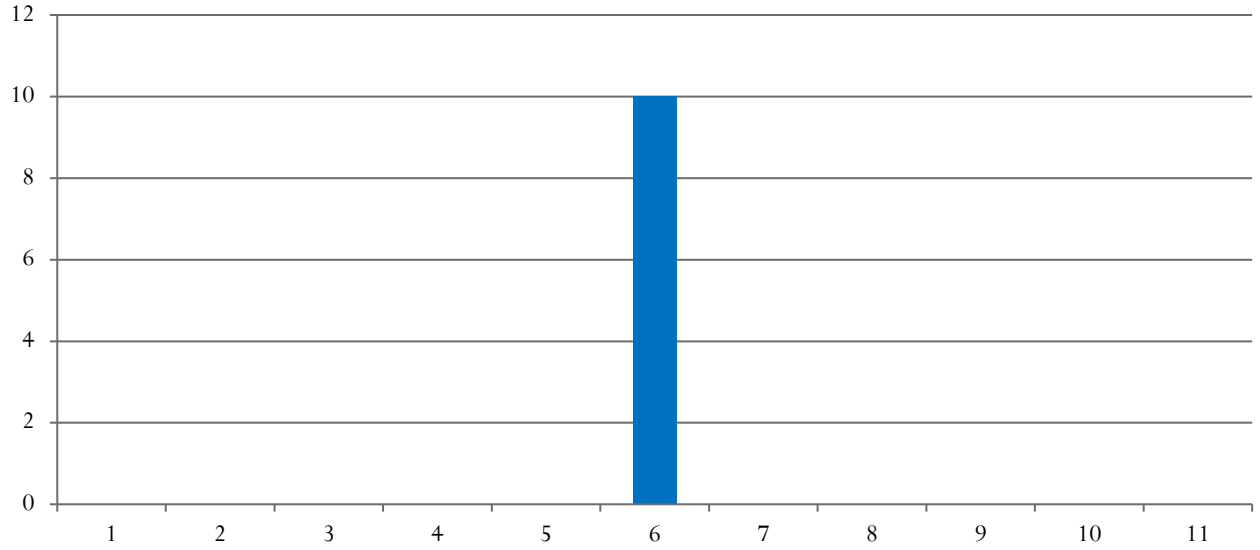


- Obligatoriamente se deben conocer siempre estas dos medidas, ya que la confianza de todo análisis está en función de la medida de tendencia y la medida de variabilidad.

- A continuación se detalla lo que es la variabilidad de una forma más práctica. Supóngase que el ejemplo es la cantidad de dinero que las personas reciben por semana (1=100 000 ¢).

El concepto de variabilidad

- Variabilidad “nula”.

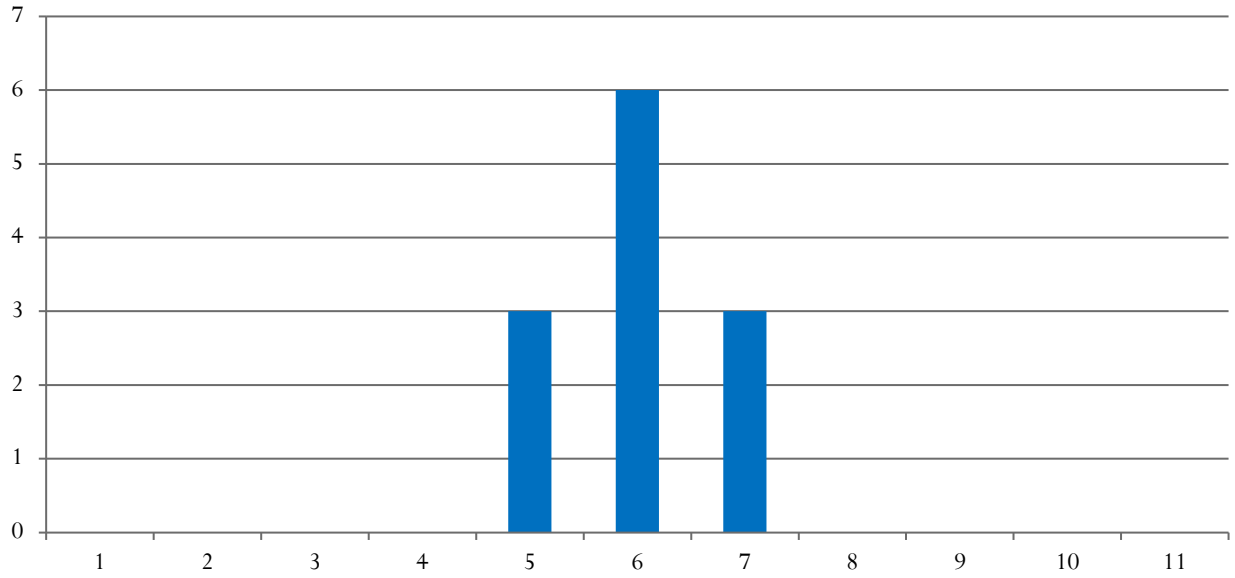


Promedio = 6

- Cuanto más concentrados estén los datos alrededor del promedio, se tendrá más confianza para caracterizar o representar el conjunto de datos.

El concepto de variabilidad

- Pequeña variabilidad.

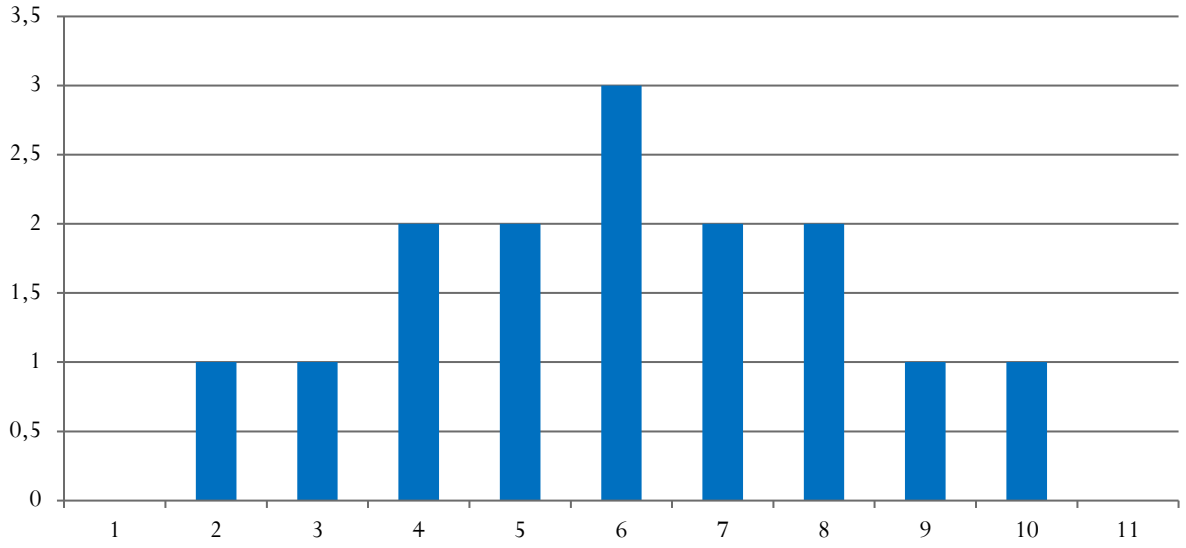


Promedio = 6

- Cuando los datos estén concentrados alrededor del promedio, y estos no se alejen mucho, entonces se tendrá un nivel de confianza aceptable para caracterizar a los datos, con un pequeño grado de error asociado.

El concepto de variabilidad

- Mayor variabilidad

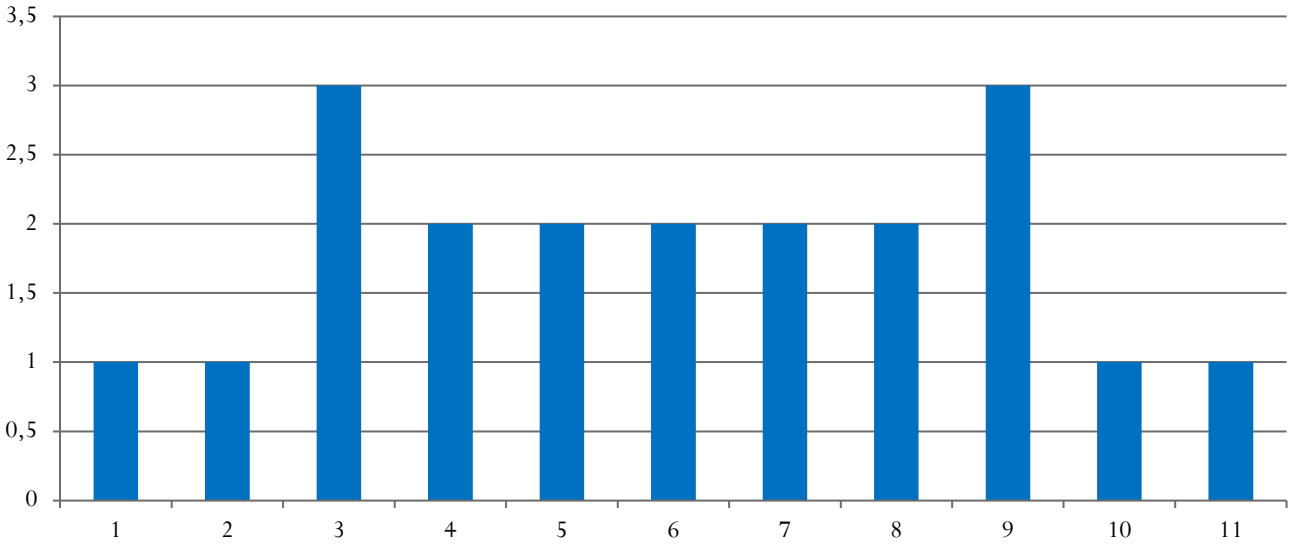


Promedio = 6

- Cuando los datos estén concentrados alrededor del promedio, pero estos se alejan relativamente bastante, entonces se tendrá un nivel de confianza no tan bueno para caracterizar a los datos, y un mayor grado de error asociado.

El concepto de variabilidad

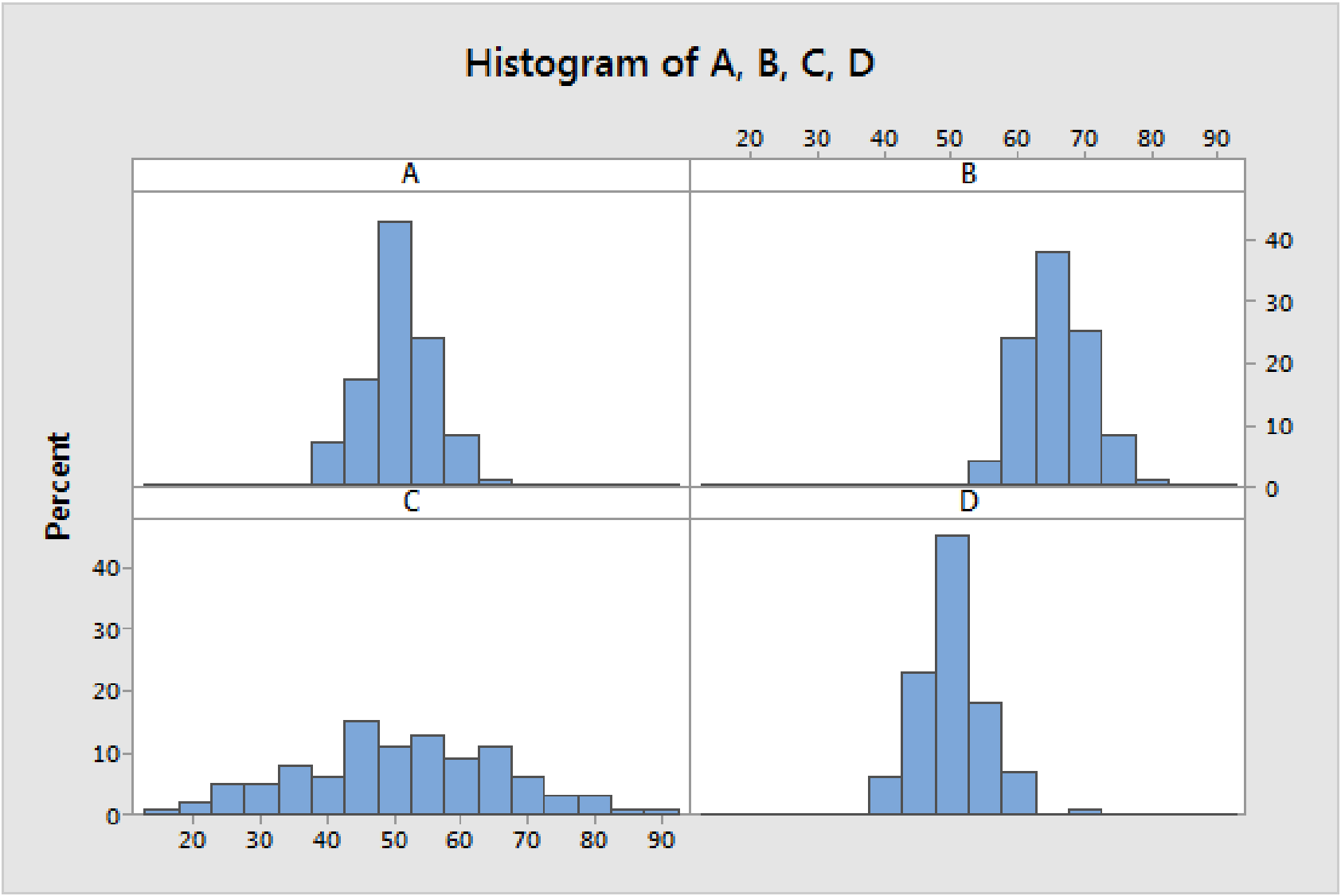
- Variabilidad extrema



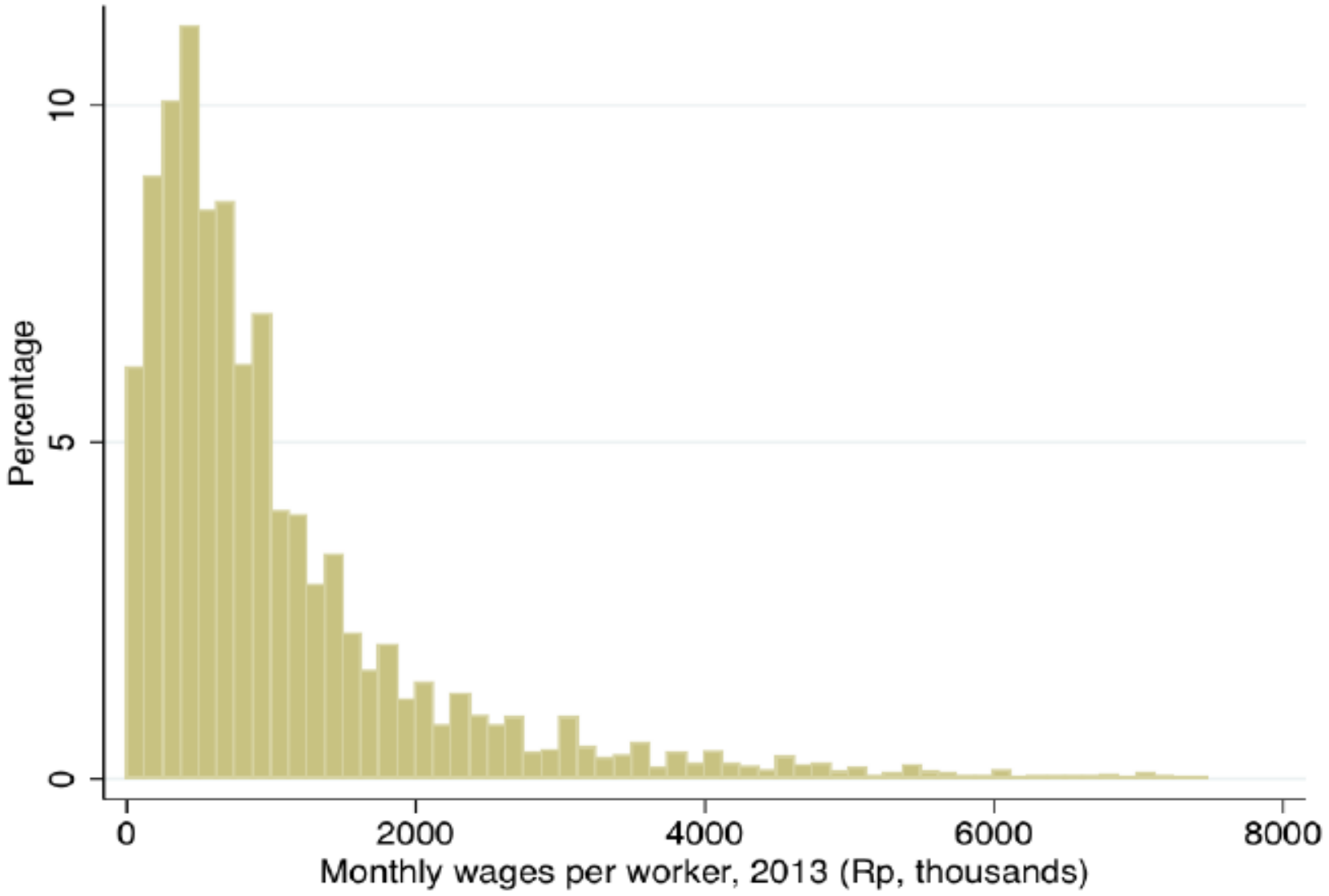
Promedio = 6

- Cuando los datos no están concentrados alrededor del promedio, y estos se alejan bastante del promedio, el nivel confianza será muy malo, y habrá un grado de error asociado bastante grande.

El concepto de variabilidad



El concepto de variabilidad

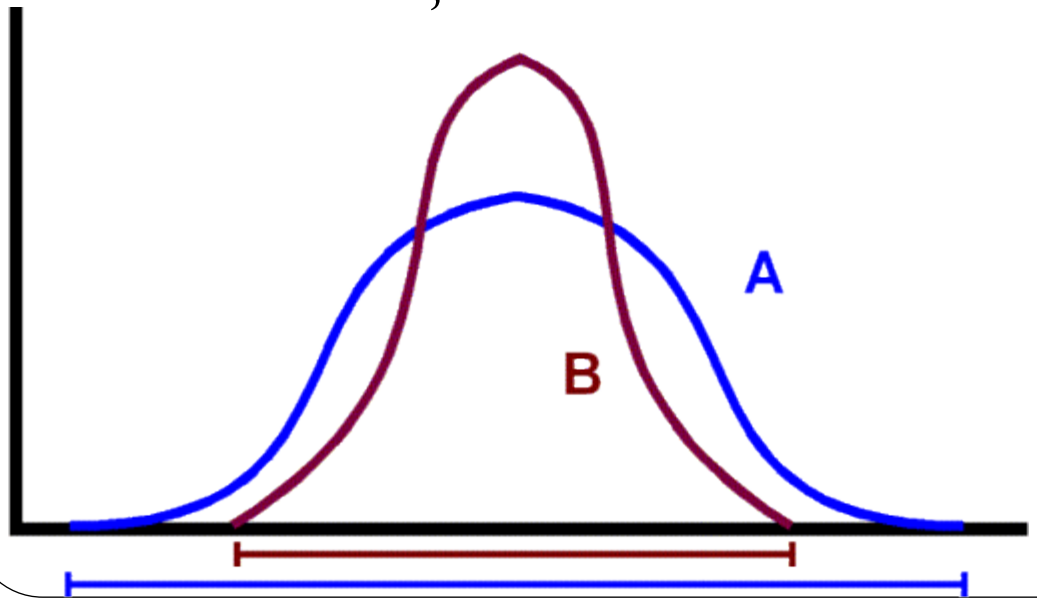


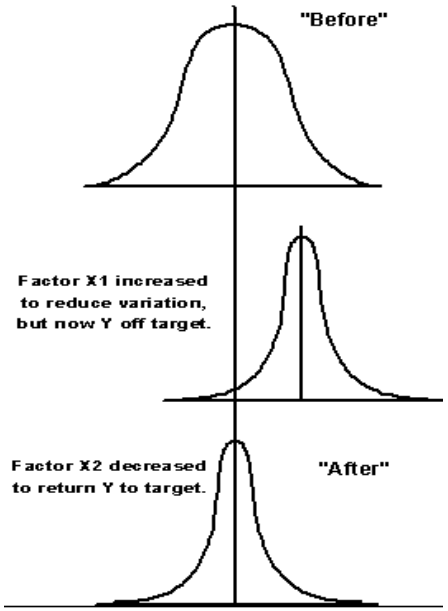
El concepto de variabilidad

El concepto de variabilidad

- Por lo tanto, se evidencia que las MDP no son suficientes para describir un conjunto de datos, y la necesidad de aplicar también una medida de variabilidad.

- Dentro de cualquier aplicación y área que requiera de la Estadística, es necesario estudiar la variabilidad.





Medidas de variabilidad

- Existen diferentes formas de medir la dispersión o variabilidad de los datos.

- La elección de cierta medida depende de la situación y de las posibles ventajas en relación a las otras medidas.

- Las medidas de variabilidad más corrientes son:

  -El recorrido o amplitud.

  -La desviación media (DM).

  -La variancia y desviación estándar (DE).

Índice

Recorrido o amplitud

- La forma más simple de apreciar la variabilidad es considerando los valores extremos del conjunto de datos (el valor menor y el valor mayor).

- La diferencia ente el valor mayor y el menor del conjunto de datos da origen al recorrido o amplitud.

$$\text{Recorrido} = X_{máximo} - X_{mínimo}$$

$$\text{Recorrido} = \text{Valor mayor} - \text{valor menor}$$

Recorrido o amplitud

- El cálculo es muy simple. Considérese las siguientes edades:

19, 20, 21, 20, 21, 19, 18, 22, 20, 19, 23.

Recorrido = 23 – 18 =5

- En este caso se tiene que las edades tienen una dispersión o amplitud de 5 años. Dependiendo del contexto, puede ser mucho o poco.

Recorrido o amplitud

- Supongamos que en el conjunto de datos se incluye el número “32”, lo cual daría el siguiente conjunto de datos:

19, 20, 21, 20, 21, 19, 18, 22, 20, 19, 23, 32.

Recorrido =32– 18 =14

- La edad tiene una amplitud de 14 años. De nuevo, dependiendo del contexto se puede considerar mucho o poco.

Recorrido o amplitud

- El recorrido es una medida que se ve influenciada 100% por los valores extremos, y únicamente describe la lejanía absoluta que hay entre los datos.

- El recorrido no es muy usado debido a ciertas limitaciones:

  -No toma en cuenta todas las observaciones del grupo, sino únicamente el valor mayor y el menor.

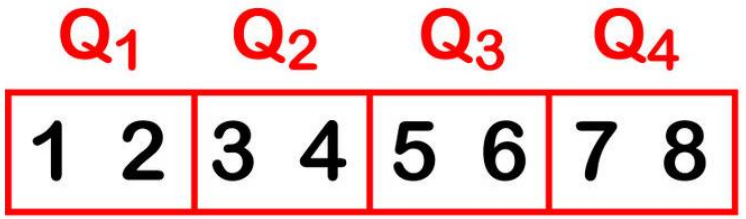
  -El aumentar el número de observaciones casi nunca produce información útil en la variabilidad.

- Es por lo anterior que se debe recurrir a otras medidas.

Índice

Intervalo de cuartiles

- Conocido también como Amplitud del intercuartil, constituye otra medida de variabilidad.



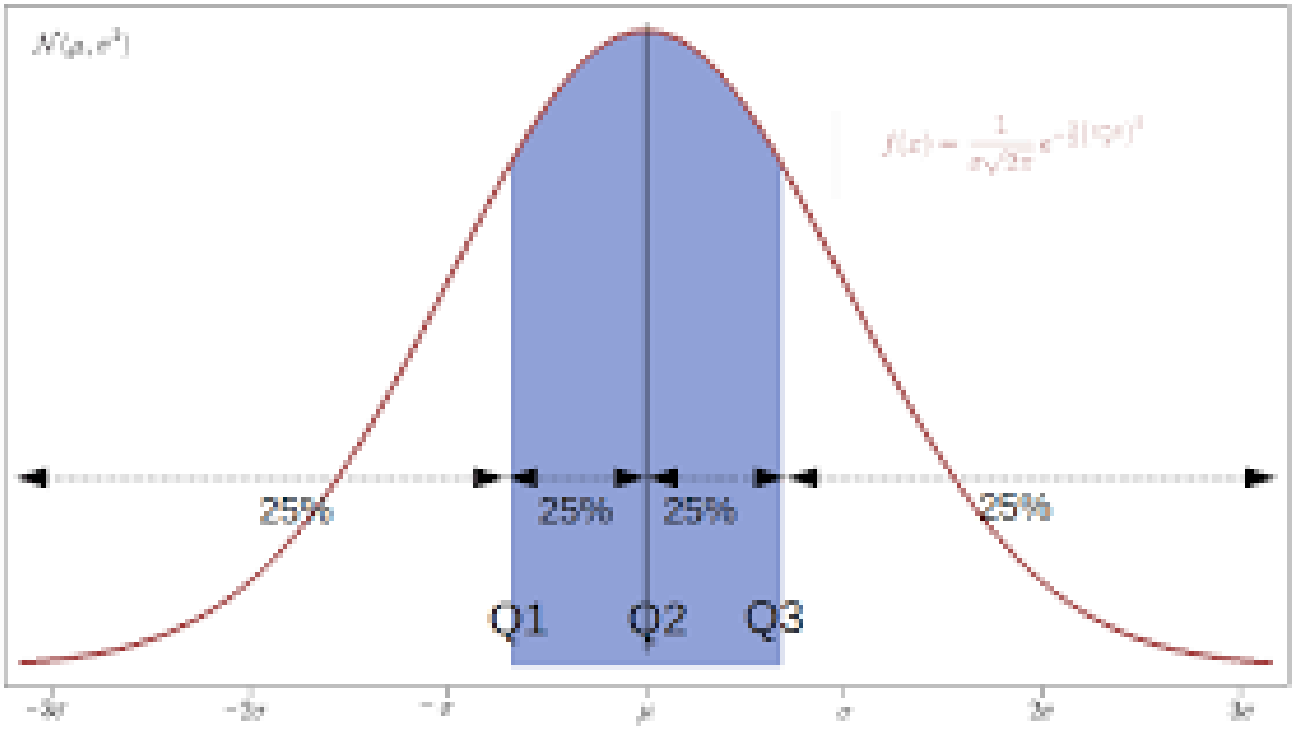
- Se expresa mediante la fórmula:

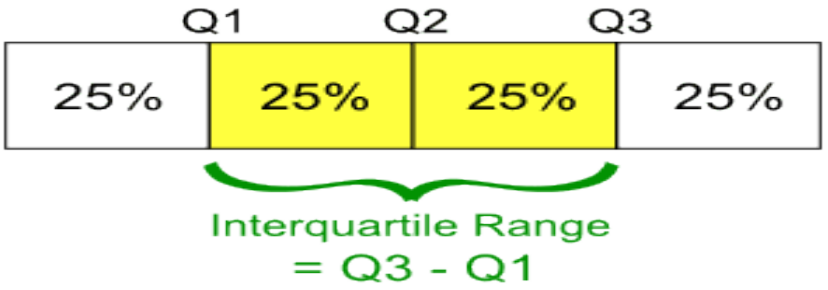
$$\mathrm{IQR} = \mathrm{Q}_3 - Q_1$$

- La medida indica donde se aglomera el 50% de los datos respecto a la mediana.

- Permite saber tanto donde está el 50% aglomeración central (posición), como el rango de tal aglomeración (variabilidad).

Intervalo de cuartiles





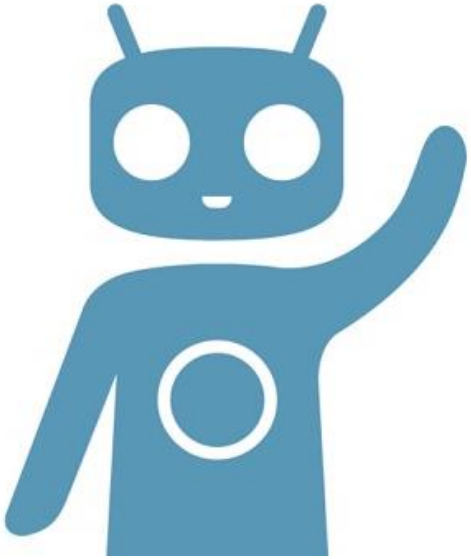
Índice

La desviación media

- Dadas las limitaciones del recorrido, es necesario definir una medida de dispersión que tome en cuenta en el cálculo todos los datos.

- La medida tiene que estar basada en las desviaciones o diferencias de los datos individuales respecto a un valor central o típico (como el promedio).

- La mejor opción es considerar la suma de las desviaciones de los datos con respecto al promedio.



La desviación media

- Sin embargo, si esto se hiciera de buenas a primeras, esta opción siempre arrojaría valores iguales a "0".

- Para obviar este problema, se suele usar los valores absolutos de las diferencias, y dividirlas por el número de datos para obtener una medida de dispersión promedio o por observación.

- Así se origina la desviación media:

$$DM = \frac{\sum|X_i - \bar{X}|}{n} = \frac{desviación\ absoluta}{número\ de\ datos}$$

La desviación media

- Sea el siguiente conjunto de datos:

3, 10, 2, 8 y 7.

El promedio: $\bar{x} = 6$

- La DM se expresa:

xi	xi -x	|xi-x|
3	-3	3
10	4	4
2	-4	4
8	2	2
7	1	1
30	0	14

La desviación media

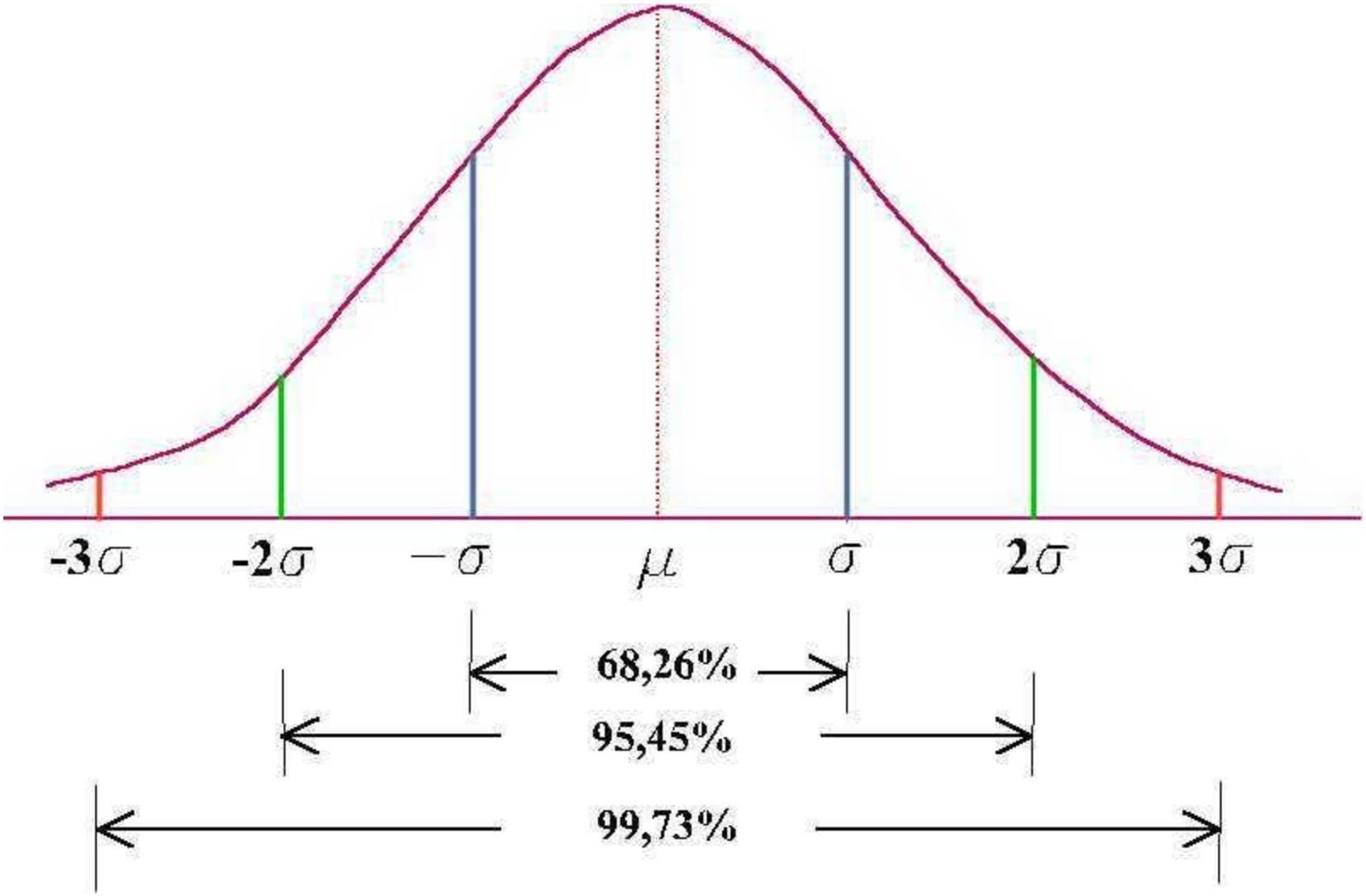
- $$DM = \frac{\sum|X_i - \bar{X}|}{n} = \frac{14}{5}$$

$$= 2.8$$

Por lo tanto, la desviación media de los datos es de 2,8.

- Sin embargo, ahora es importante saber si 2.8 es un valor "grande" o "pequeño"; o más importante es si los datos poseen mucha o poca variabilidad.

La desviación media

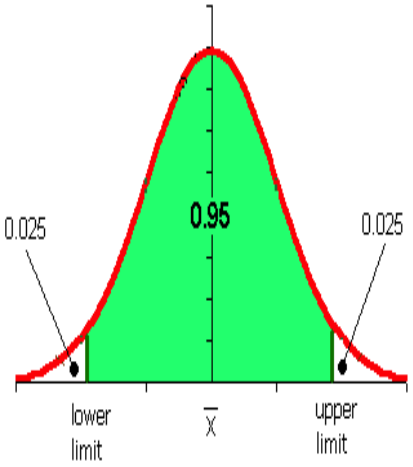


La desviación media

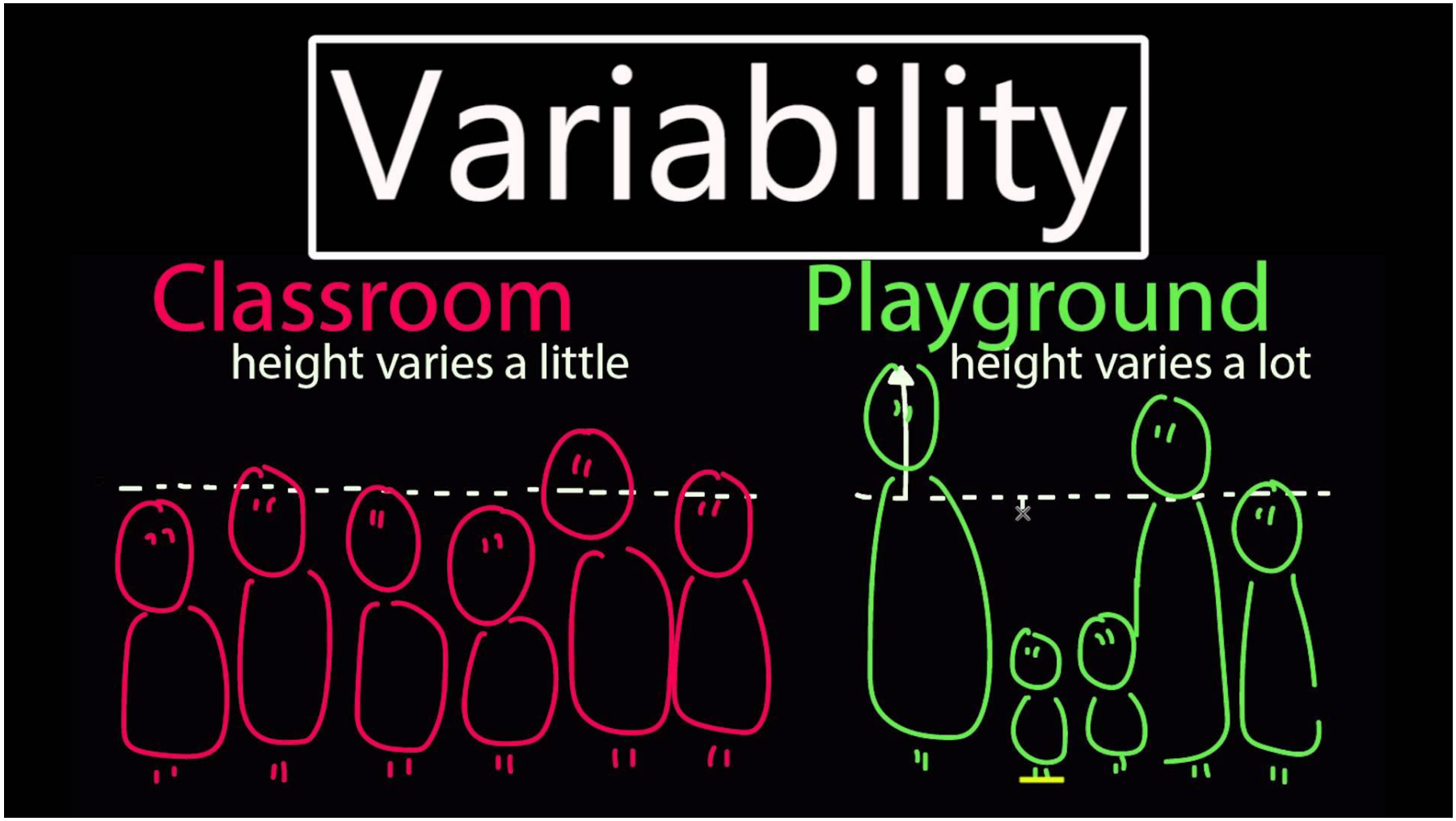
- El tamaño de la desviación media va a depender mucho de la medida que se esté utilizando. Sin embargo el resultado por si solo no dice nada.

- Si la desviación media se utiliza como la desviación estándar, para saber si los datos son muy o pocos dispersos podemos utilizar el criterio de que si el 95% de los datos están "± 2" desviaciones, entonces no hay tanta variabilidad.

- El criterio anterior aplica a conjunto mayores a 30.

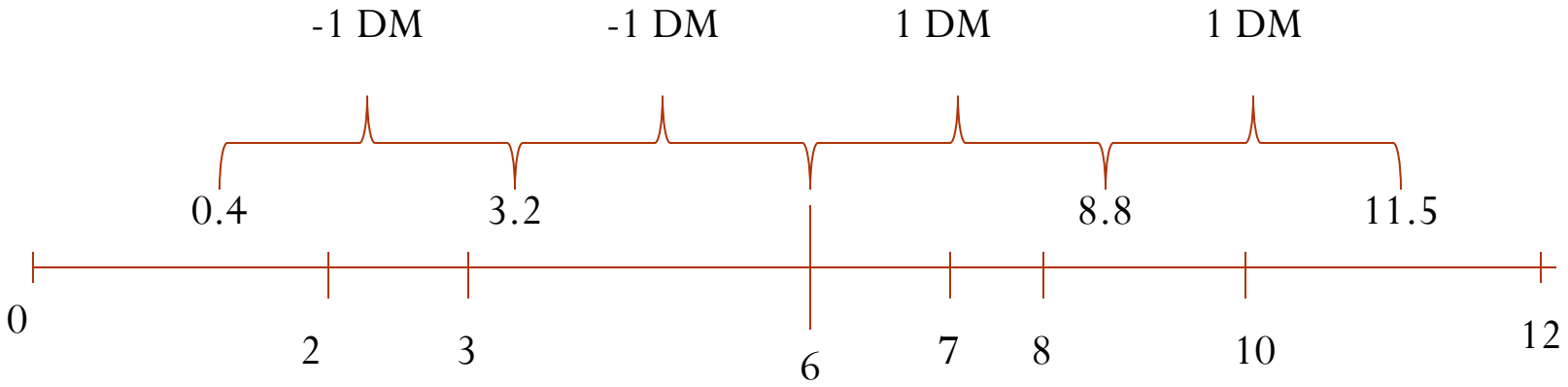


La desviación media

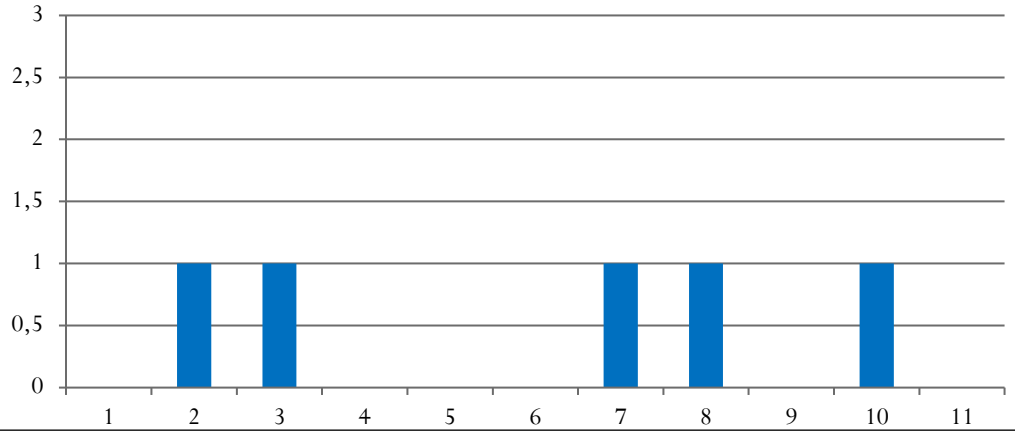


La desviación media

- Para el presente caso, como tan sólo hay 5 datos el criterio anterio no aplica.



- Sin embargo, si observamos la distribución de datos, se observa una gran variabilidad del conjunto.



La desviación media

- Debido principalmente a la falta de requerimientos teóricos y muchas veces prácticos, la desviación media no es la medida de dispersión más utilizada.

- Por eso se debe recurrir la gran mayoría de veces a la variancia y desviación estándar.

$$\sigma^2 \; and \; \sigma$$

Índice

Variancia y desviación estándar

- La variancia y la desviación estándar son las medidas de dispersión más cómodas y útiles que reúnen numerosas ventajas prácticas y teóricas.

- Como para la DM, se toma en cuenta las desviaciones o diferencias de todos los datos con respecto al promedio.

- Sin embargo, en esta ocasión en vez de considerar las desviaciones absolutas de cada dato con respecto al promedio, se van a considerar las desviaciones cuadráticas de los datos con respecto al promedio.

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

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

Variancia y desviación estándar

- Las desviaciones cuadráticas permiten obtener la medida de dispersión denominada "variancia". Si a la variancia se le aplica una raíz cuadrada, entonces se puede obtener la desviación estándar o desviación típica.

- La variancia se formula de la siguiente forma:

$$Variancia = s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{desviación\ cuadrática}{número\ de\ datos}$$

Variancia y desviación estándar

- La desviación estándar (DE) se describe de la siguiente forma:

$$DE = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{desviación\ cuadrática}{número\ de\ datos}}$$

- La desviación estándar nos indica cuánto se alejan, en promedio, las observaciones de la media aritmética del conjunto.

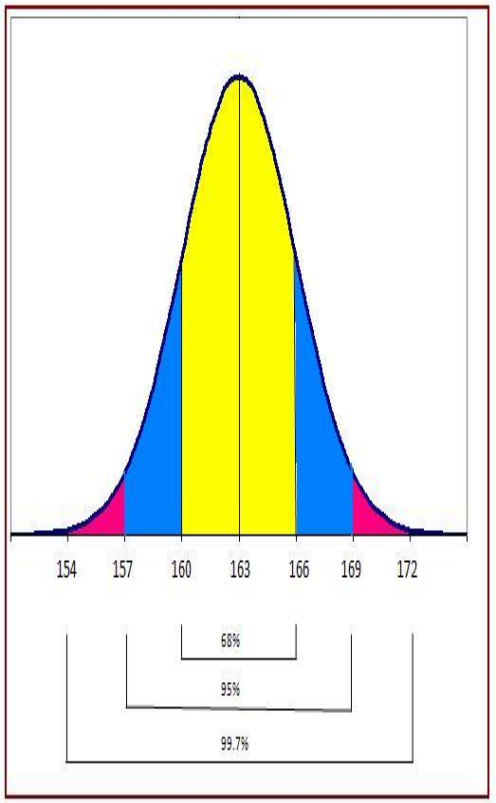
- Es la medida de dispersión más usada en estadística, tanto en aspectos descriptivos como analíticos.

Variancia y desviación estándar

- Para fines prácticos, la medida de dispersión que se utiliza en informes, reportes, descripciones y se interpreta es la desviación estándar.

- La variancia se utiliza para facilitar los cálculos, pero no se interpreta ni nada por el estilo, ya que las medidas están modificadas por el cuadro, e interesa la medida original de la variable.

- De ahí que si interpretamos algo será la DE.



Variancia y desviación estándar

- Para el cálculo de la variancia se puede proceder de dos formas. La primera es la forma "normal" y la otra es la "simplificada".

- El cálculo de la variancia "normal", cuando se tiene una muestra de "n" datos, se expresa como "S^2" , se denota:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Variancia y desviación estándar

- Sea el conjunto de datos anterior,

$$3, 10, 2, 8, 7.$$

- Para obtener la variancia:

xi	xi -x	(xi-x)^2
3	-3	9
10	4	16
2	-4	16
8	2	4
7	1	1
30	0	46

$$\bar{x} = 6$$

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{(n-1)} = \frac{46}{4} = 11.5$$

Variancia y desviación estándar

- Por lo tanto la variancia es igual a 11.5.

- Para obtener la desviación estándar lo que tenemos que hacer es sacar raíz cuadrada al resultado anterior.

$$\sqrt{s^2} = s = \sqrt{11.5} = 3.39$$

- Nótese que este es el resultado que se interpreta, pero esto lo haremos más adelante.

Variancia y desviación estándar

- La otra forma para obtener la variancia "simplificada" se denota de la siguiente forma:

$$s^2 = \frac{\left(\sum {x_i}^2\right) - \frac{(\sum x_i)^2}{n}}{n-1}$$

- Para obtener la variancia mediante este método, se utiliza las observaciones individuales en todo momento.

Variancia y desviación estándar

- 1. Se obtiene el cuadro de todas las Xi observaciones:

xi	xi ^2
3	9
10	100
2	4
8	64
7	49
30	226

- 2. Se obtiene la suma de todas las observaciones, se suma el cuadrado y finalmente se divide por el número de observaciones.

$$\frac{(\sum x_i)^2}{n} = \frac{(30)^2}{5} = \frac{900}{5} = 180$$

Variancia y desviación estándar

- 3. Se obtiene los términos cuadráticos:

$$\left(\sum {x_i}^2\right) - \frac{(\sum x_i)^2}{n} = 226 - 180 = 46$$

- Finalmente se obtiene la variancia:

$$s^2 = \frac{\left(\sum {x_i}^2\right) - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{46}{4} = 11.5$$

Variancia y desviación estándar

- Finalmente, la desviación estándar es la raíz cuadrado de la varianza.

$$\sqrt{s^2} = s = \sqrt{11.5} = 3.39$$

- La desviación estándar es de 3.39. Se interpretaría como la desviación con respecto al promedio.

- Al igual que para la DM, interesa saber si el valor de la DE es grande o pequeño; y finalmente valorar la magnitud de la dispersión de los datos. De nuevo, el valor por si mismo no dice mucho.

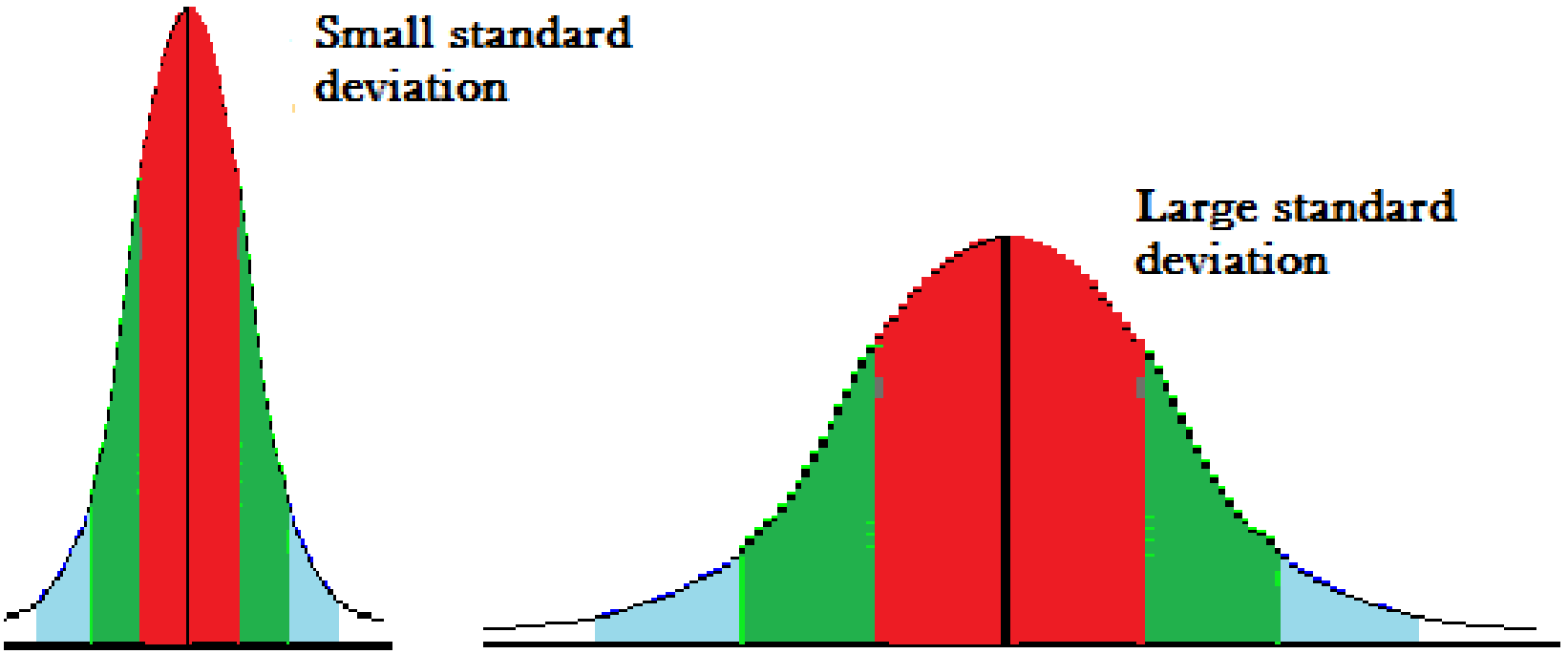
Variancia y desviación estándar

- Si los datos son muy o pocos dispersos podemos utilizar el criterio de que si el 95% de los datos están "± 2" desviaciones, entonces no hay tanta variabilidad.

- Sin embargo, este regla aplica para tamaños de muestra superior a 30.

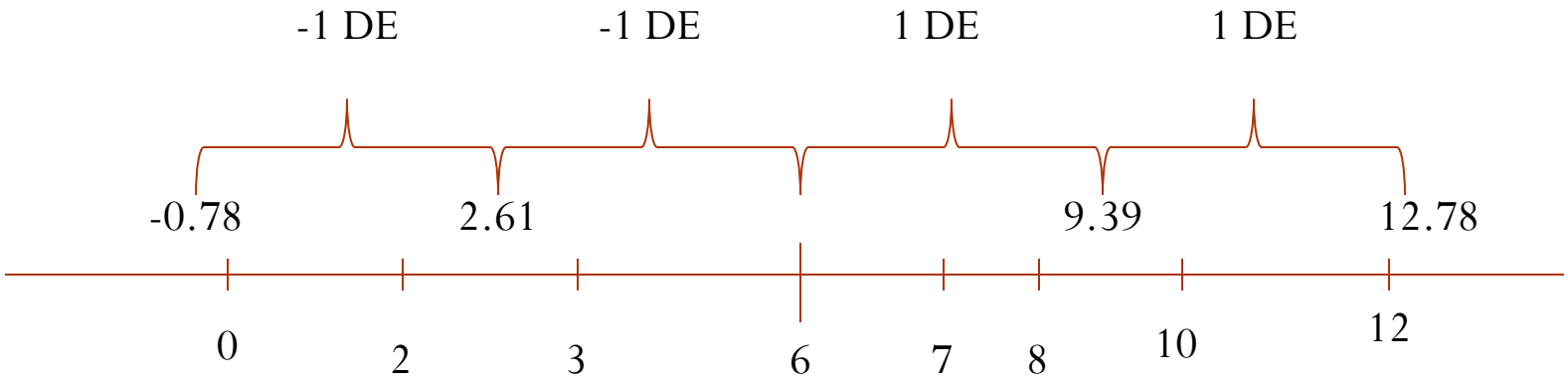
- Como se observó anteriormente, la distribución de los es muy dispersa, y además, realmente una dispersión de 3.4 alrededor de la media es muy grande.

Variancia y desviación estándar



Variancia y desviación estándar

- Como tan sólo hay 5 datos el criterio anterior no aplica. Pero su representación se muestra a continuación:



- Antes se había observado que los datos están muy dispersos alrededor del promedio.

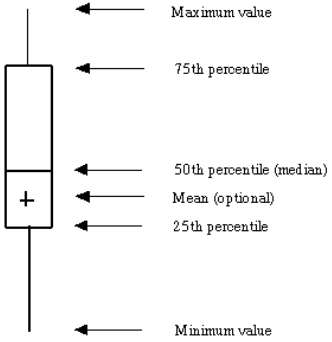
Índice

Gráfico de cajas (box plot)

- ¿Es posible resumir en un solo gráfico los aspectos de las medidas de posición como los de variabilidad?

- Un **Diagrama de caja**, también conocido como diagrama de caja y bigotes, es un gráfico que está basado en cuartiles y mediante el cual se visualiza la distribución de un conjunto de datos. Está compuesto por un rectángulo, la "caja", y dos brazos, los "bigotes"…

- Es un gráfico que suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y la simetría de la distribución. Primero es necesario encontrar la mediana para luego encontrar los 2 cuartiles restantes



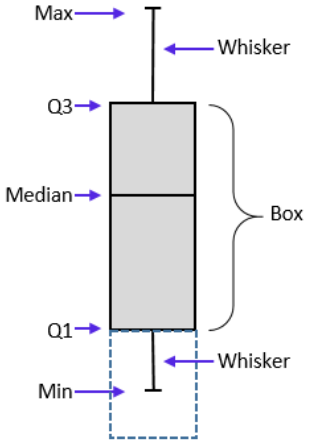


Gráfico de cajas (box plot)

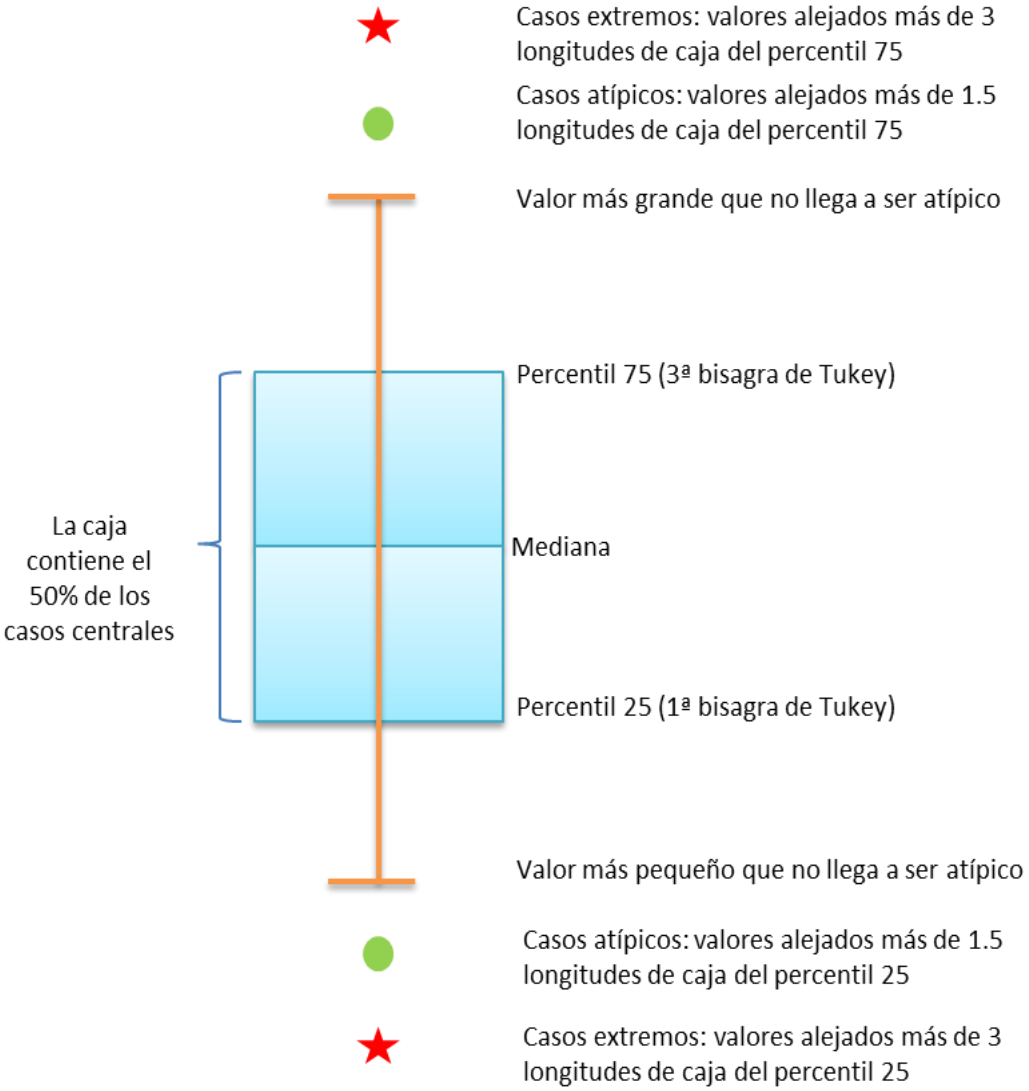


Gráfico de cajas (box plot)

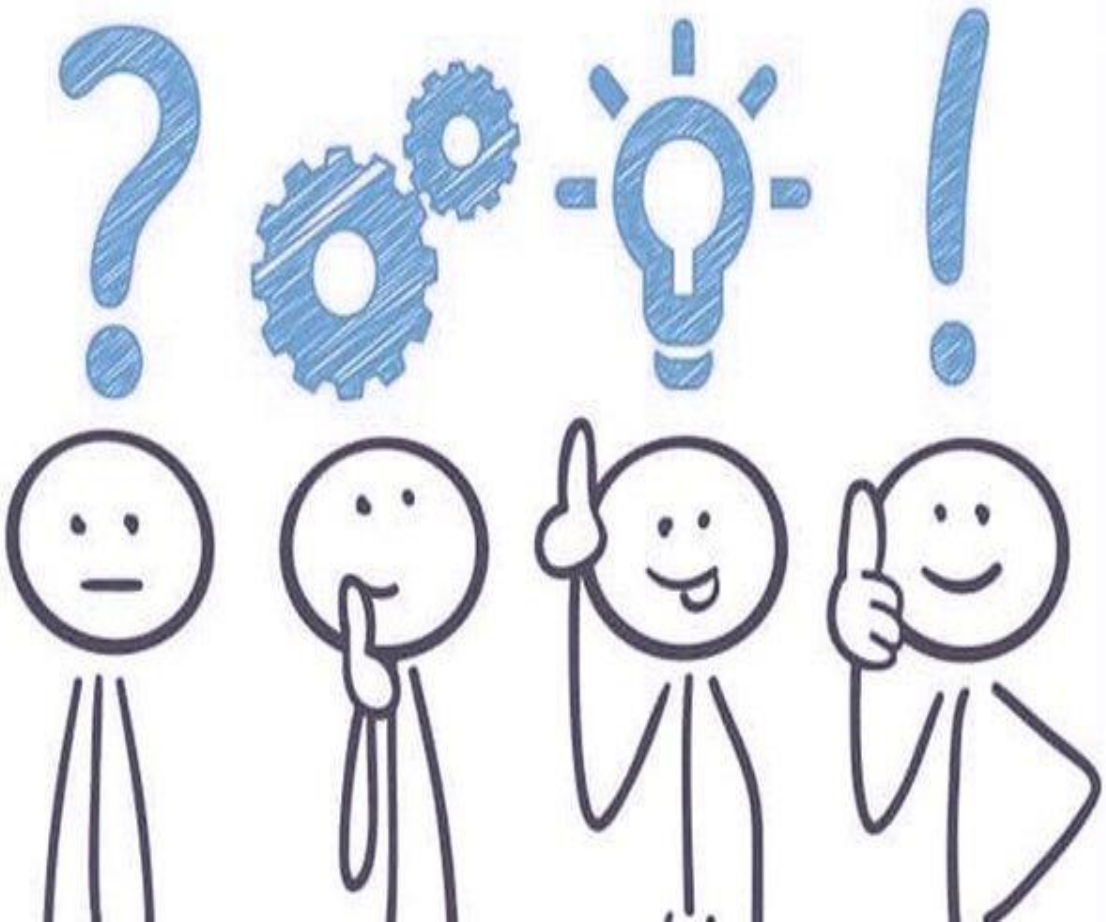
- La utilidad del gráfico de cajas son las siguientes:

1. Proporcionan una visión general de la simetría de la distribución de los datos; si la mediana no está en el centro del rectángulo, la distribución no es simétrica.
2. Son útiles para ver la presencia de valor atípicos también llamados ***outliers***.
3. Pertenece a las herramientas de las estadística descriptiva. Permite ver como es la dispersión de los puntos con la mediana, los percentiles 25 y 75 y los valores máximos y mínimos.
4. Ponen en una sola dimensión los datos de un histograma, facilitando así el análisis de la información al detectar que el 50% de la población está en los límites de la caja.

Índice

Coeficiente de variación

¿Si vamos a comprar el salario medio entre hombres y mujeres, qué tendríamos que hacer para comprar los salarios de estas dos poblaciones?



Coeficiente de variación

¿Si vamos a comprar la variabilidad del salario entre hombres y mujeres, qué tendríamos que hacer para comprar cuál salario es más o menos variable entre las poblaciones?

Coeficiente de variación

- La desviación estándar es útil como medida de variación en un determinado conjunto de datos.

- Cuando se quiere comparar la dispersión de dos conjuntos de datos, la comparación de las dos desviaciones estándar puede dar un resultado equivocado.

$$CV = \frac{S}{|\bar{x}|}$$

- Esto puede ocurrir si las dos variables involucradas tienen medidas en diferentes unidades o provienen de dos tipos de población.

- Ejemplo: los pesos de los niños de primer grado de primaria son comparadas contra la desviación estándar de los pesos de los estudiantes de primer año de universidad, esta última es numéricamente mayor que la anterior, debido a que los pesos mismos son mayores y no porque la dispersión sea mayor

Coeficiente de variación

- Se necesita una medida de variancia relativa en lugar de una de variancia absoluta.

- Esta medida se conoce como *coeficiente de variación*, el cual expresa la desviación estándar como un porcentaje de la media. Se expresa como:

$$\text{C.V.} = \frac{s}{\bar{x}} * 100$$

- Como la media y desviación estándar se expresan en la misma unidad de medición, la unidad de medición se cancela al calcular el coeficiente de variación. Esto hace que se obtenga una medida independiente de la unidad de medición

Coeficiente de variación

- En la comparación de dos tipos de población, personas con edades promedio de 11 años, y personas con edades promedio de 25 años, se desea saber cuál posee mayor variabilidad, los pesos de los de 25 años o los de los 11 años.

- Los datos son los siguientes

	Muestra 1	Muestra 2
Edad	25 años	11 años
Peso medio	145 libras	80 libras
Desviación estándar	10 libras	10 libras

Coeficiente de variación

- Una comparación de las desviaciones estándar puede conducir a la conclusión de que las dos muestras poseen igual variabilidad. Sin embargo, si se calcula el coeficiente de variación se obtiene por grupo:

$$\text{Personas 25 años:} \quad C.V. = \frac{10}{145} * 100 = 6.9$$

$$\text{Personas 11 años:} \quad C.V. = \frac{10}{80} * 100 = 12.5$$

….la impresión recibida es diferente

- Como el CV es independiente de la escala de medición, constituye una estadística útil para compara la variabilidad de dos o más variables en medidas en escalas diferentes.

Reseñas finales

- Para tener idea de la dispersión, se puede utilizar tanto el criterio numérico como gráfico.

- El conocer las medidas de posición como variabilidad no sólo es una buena herramienta para la descripción; todos los métodos inferenciales utilizan medidas de variabilidad.

- Obtener conjuntamente el promedio y la desviación estándar es útil tanto en la estandarización (prueba de hipótesis), como en la estimación (intervalos de confianza).

- Esto último es parte de la Estadística Inferencial, y se estudiar en próximos capítulos.


The End