ESTADISTICA DESCRIPTIVA UNAB

RESUMEN CLASE N° 2 (19-03-2018)

Clasificación de los Métodos Estadísticos

Los métodos estadísticos pueden clasificarse en dos grandes grupos:

- Descriptivos
- Inferenciales

Definición (Métodos Descriptivos)

Se preocupan de describir el conjuntos de datos. Generalmente están relacionados con el análisis preliminar o exploratorio de los datos.

Definición (Métodos Inferenciales)

Se preocupan de la obtención de conclusiones sobre la población de donde se ha extraído una muestra.

Líneas de razonamiento estadístico

- Recolección de los datos.
- Descripción estadística de los datos.
- Análisis estadístico de los datos.
- Decisión o predicción.

UNIDAD EXPERIMENTAL: Es la unidad de análisis a la cual es el ente que me va a proporcionar la información.

POBLACIÓN: es el universo o conjunto total de unidades experimentales. Sobre este conjunto se obtendrán las conclusiones finales.

MUESTRA: es cualquier subconjunto de la población.

PARÁMETRO: característica relacionada con la población y que es de interés para el investigador.

ESTIMADORES: aproximaciones de los parámetros basadas en la muestra.

Ejemplo*: Medir índice de productividad de las empresas de la V región de X rubro.*

UE: La empresa

Variable: Índice de productividad

Población: La V región

El **parámetro** es un valor determinado de una población, por ejemplo, la media poblacional, la varianza, la proporción. Es información relevante de una población.

Un **estimador** es un valor que puede calcularse a partir de los datos muestrales y que proporciona información sobre el valor del parámetro. Por ejemplo la media muestral es un estimador de la media poblacional, la proporción observada en la muestra es un estimador de la proporción en la población.

Ejemplo: Podemos "estimar" la Media poblacional (Parámetro), a través, de información de la muestra tomada de la población.

¿Cómo se llama la metodología estadística que, a partir de una Muestra, **permite tomar conclusiones con respecto a lo general de la población de interés**? ====> **Inferencia Estadística**

Población está constituida por **Unidades experimentales**. Nosotros vamos a trabajar con un subconjunto llamado **Muestra (Grupo de Unidades Experimentales)**, y esta nos permitirá tomar decisiones con respecto a la población **(Inferencia Estadística)**.

Los Parámetros son **características desconocidas**. ¿Por qué? Porque están asociadas a la población

Las muestras son **características conocidas**.

Las **Variables** es la característica que yo quiero medir de cada Unidad experimental.

VARIABLE

Cualitativas (Cualidad o atributo): Nominal (solo permite la clasificación y no hay un orden) ó Ordinal (denotan un orden jerárquico)

Nominal: Sexo, Nombre, grupo étnico, estado civil, religión, nacionalidad.

Ordinal: Primero, segundo,..... ó Malo, Regular, Bueno, Muy bueno

Ejemplo:

Cuantitativas (Variables numéricas, de cantidad)

Discreta: Valores enteros. Número de hijos, Numero de dientes

Continua: Cuando cualquier valor intermedio es posible (Decimales). Peso 57,6 Kg, Temperatura 36,77 ° C

ORGANIZACIÓN DE DATOS

- Frecuencia Absoluta
- Frecuencia Relativa
- Frecuencia Absoluta Acumulada
- Frecuencia Absoluta Relativa

Variable Continua: Si la variable bajo estudio es cuantitativa continua (o discreta con un alto rango de variabilidad), entonces el esquema de tabla anterior sufre un leve modificación que está relacionada con la creación de los "intervalos de clases". En este caso el esquema de la tabla es el siguiente:

Intervalos (Clases)	Marca de clase	n_i	f_i	N_i	F_i
$[\min; \min + A[$	m_1	n_1	f_1	N_1	F_1
$[\min + A; \min + 2A[$	m_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[\min + (k-1)A; \max]$	m_k	n_k	f_k	N_k	F_k

Donde la marca de clase i-ésima (m_i) corresponde al promedio del intervalo i-ésimo ($i = 1, ..., k$).

Construcción Tablas para Variables Continuas:
Se supone que la amplitud de los intervalos es la misma, se puede seguir los siguientes pasos para la construcción de tablas de frecuencias de variables continuas.

Paso 1: Contar el número n de datos.

Paso 2: Calcular el rango (R),

$$R = \max - \min,$$

donde min y max corresponden a los valores mínimos y máximos de los datos, respectivamente.

Paso 3: Escoger el número de clases (intervalos). Se sugiere ,el entero más próximo de la denominada ***fórmula de Sturges***, dada por

$$k = 1 + 3,3 \log(n),$$

donde $\log(\bullet)$ es el logaritmo en base 10.

Paso 4: Calcular la amplitud (A)

$$A = \frac{R}{k}.$$

Paso 5: Para determinar los extremos de la primera clase (intervalo) se debe tomar como límite inferior el valor min y como límite superior el valor $\min + A$.

Paso 6: Para obtener las restantes clases, se suma sucesivamente A al límite inferior, donde el límite inferior de las sucesivas clases corresponderá a límite superior de la clase anterior.

Frecuencia absoluta: Corresponde al número de unidades de análisis que pertenecen a la clase c_i
Frecuencia relativa: Corresponde a la proporción de unidades de análisis que pertenecen a la clase c_i

Considere los siguientes datos

0,36	0,68	0,8	0,87	0,92	1,00
0,48	0,71	0,81	0,87	0,94	1,00
0,60	0,72	0,81	0,88	0,97	1,13
0,61	0,73	0,82	0,92	0,97	1,16
0,68	0,79	0,85	0,92	0,97	1,19

cree una tabla de distribución de frecuencias con la metodología vista anteriormente.

1. El valor del tamaño de muestra es $n = 30$.
2. El rango sería $R = 1,19 - 0,36 = 0,83$.
3. El número de clases a considerar es $k = 1 + 3,3\log(30) = 5,87 \equiv 6$.
4. La amplitud sería $A = \frac{0,83}{6} = 0,138\overline{3}$.
5. La tabla quedaría de la siguiente forma

Intervalos	Marca de clase	n_i	f_i	N_i	F_i
$[0,360; 0,498[$	$0,429$	2	$0,067$	2	$0,0\overline{6}$
$[0,498; 0,637[$	$0,568$	2	$0,067$	4	$0,1\overline{3}$
$[0,637; 0,775[$	$0,706$	5	$0,167$	9	$0,3$
$[0,775; 0,913[$	$0,844$	9	$0,3$	18	$0,6$
$[0,913; 1,052[$	$0,983$	9	$0,3$	27	$0,9$
$[1,052; 1,190]$	$1,121$	3	$0,1$	30	1

n_i = Frecuencia absoluta
f_i = n_i / n = Frecuencia relativa
N_i = Frecuencia absoluta acumulada
F_i = Frecuencia absoluta relativa

Ejemplo: En una muestra de 30 porciones de hierro de 100 gr c/u, se estaba midiendo en c/u de las porciones de hierro, el contenido en gramos de silce.

UE: la barra de hierro

VARIABLE: Cantidad de gramos de silce. Es Cuantitativa - Continua

Interpretaciones:

$n_4 = 9$ → Hay 9 porciones de hierro cuyo contenido en gramos de silce, fluctúan entre 0,775 gr. y 0,913

$f_4 = 0,3$ → Un 30% de las porciones de hierro tienen un contenido en gramos de silce que fluctúan entre 0,775 gr. y 0,913 gr.

$N_4 = 18$ → Hay 18 barras de hierro ***que contienen a lo más*** (a lo más = menor o igual; a lo menos = mayor o igual) 0,913 gramos de silce.

$F_4 = 0,6$ → El 60% de las barras de hierro contiene a lo más 0,913 gr. de silce.

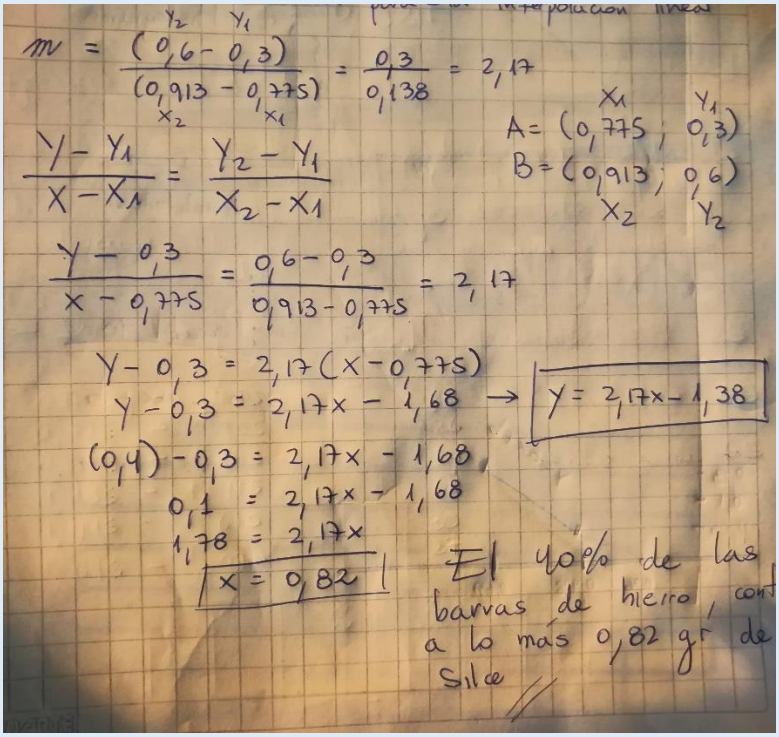
¿Cuál es el valor de la variable que acumula el 60% de la información? = 0,913

¿Cuál es el valor de la variable que acumula el 30% de la información? = 0,775

¿En qué intervalo sé que acumula el 40% de la información? = [0,775 ; 0,913]

Ejercicio en Clase:

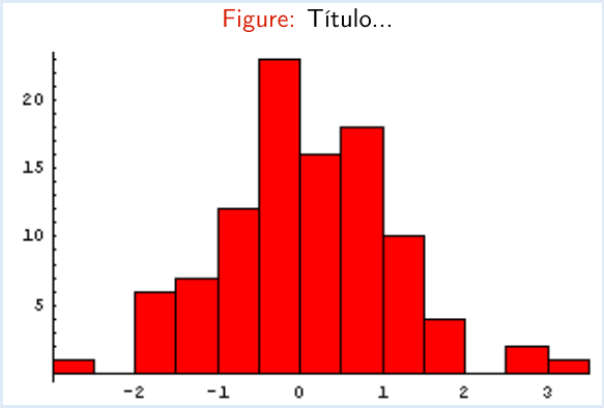
Del ejemplo de los gramos de Silce para datos tabulados. Determine a través de aproximación lineal el valor de la variable que acumulan el 40% de las barras de hierro. Sugerencia: Utilizar la ecuación de la recta para la interpolación lineal.



El 40% de las barras de hierro contienen a lo más 0,82 gramos de Silce

P_{90} = 1,052 (Percentil 90, valor de la variable que acumula el 90% de la información)

El P_{75} se encuentra en el intervalo [0,913 ; 1,052]



Histograma: Se usa para variables continuas. Es un conjunto de rectángulos adyacentes. En el eje horizontal deben ir los intervalos (clases) y en el eje vertical las frecuencias (Absoluta o Relativa).

Patrón de comportamiento: Comportamiento de la variable (simétrico, asimétrico, si es muy dispersa o muy agrupada con respecto al centro)

RESUMEN CLASE N° 3 (26-03-2018)

Medidas Estadísticas de resumen

Estas medidas estadísticas que resumen al conjunto de datos, también se les denomina ***estadísticos***. Se clasifican en medidas de posición y dispersión. Las primeras nos entregan ***la posición relativa que poseen los individuos dentro de la distribución*** y se subdividen en dos:

a) Las de centralidad, que tienden a ubicarse en el centro de la distribución, entre las cuales se encuentran:

- El promedio o media aritmética.
- La mediana o valor del centro.
- La moda, modo o valor más frecuente.
- La media geométrica.
- La media armónica.
- Entre otras.

b) Los cuantiles, que tienden a ubicarse en distintas partes de la distribución de la variable, entre las que se encuentran:

- Los cuartiles (dividen al conjunto en cuatro partes iguales).
- Los percentiles (dividen al conjunto en cien partes iguales).
- Entre otras.

Las segundas medidas estadísticas de resumen, ***las de dispersión***, nos entregan el grado de dispersión, variabilidad u homogeneidad que poseen los datos dentro del conjunto, generalmente respecto de una medida de tendencia central, entre las que se encuentran:

- El rango o desviación máxima
- El rango intercuartil.
- La varianza.
- La desviación estándar o típica.
- El coeficiente de variación.
- Entre otras.

PROMEDIO:

⇒ El uso de esta medida es exclusivamente para variables cuantitativas.

⇒ Su cálculo puede ser afectado de manera desproporcionada por la existencia de datos atípicos (fuera de lo común).

MEDIANA:

⇒ El uso de esta medida es para variables cualitativas que poseen orden jerárquico o cuantitativas.

⇒ Su cálculo no es afectado por la existencia de datos atípicos.

*[Cuando una medida no se ve afectada por datos atípicos, se denomina medida **RESISTENTE** ó **ROBUSTA**]*

MODA:

⇒ El uso de esta medida es para cualquier tipo de variable.

⇒ En el caso de variables cuantitativas, los datos pueden ser agrupados en clases y la moda se define como la marca de clase que tiene la mayor frecuencia.

⇒ Si existe un único valor que se repite más hablamos de una *distribución unimodal.*

RELACIÓN ENTRE PROMEDIO, MEDIANA Y MODA

Caso 1: *Distribución Simétrica (No Sesgada)*

$$\overline{x} = M_e = M_o.$$

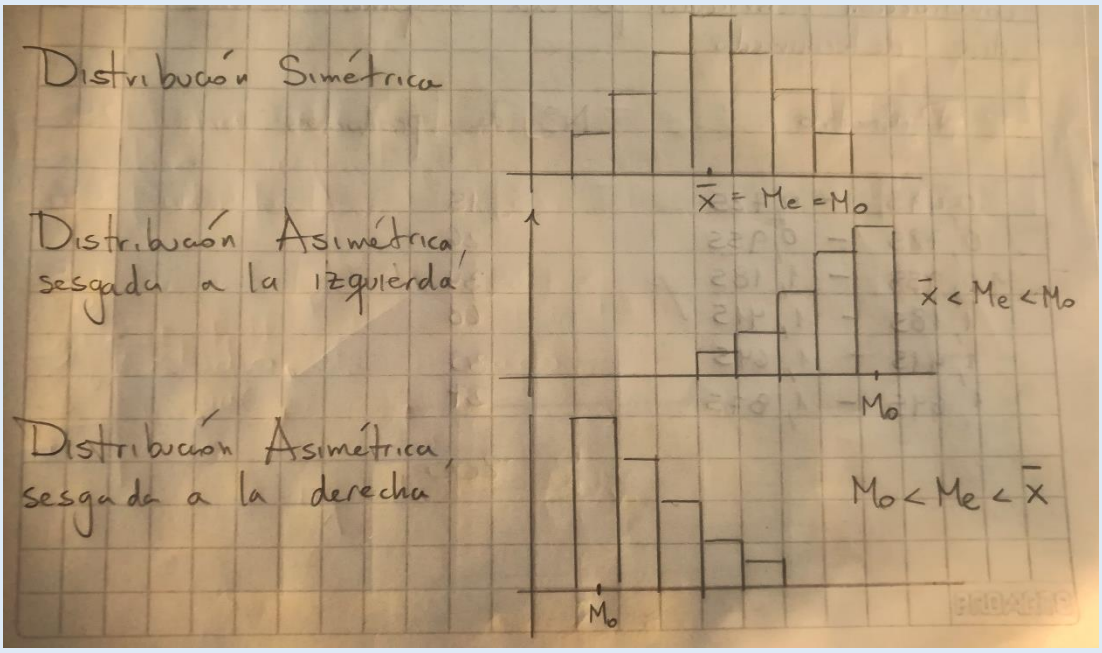
Caso 2: *Distribución Asimétrica, sesgada a la derecha.*

$$M_o < M_e < \overline{x}.$$

Caso 3: *Distribución Asimétrica, sesgada a la izquierda.*

$$\overline{x} < M_e < M_o.$$

Nota: La mediana (M_e) siempre se hallará entre el promedio $\overline{x}$ y la moda (M_o) ya que no es afectada por datos extremos, es decir, es **robusta.**



Ejemplo: En una base de datos el promedio, la mediana y la moda son muy similares. Existe un fuerte indicio de que los datos estén distribuidos de manera simétrica en forma de campana, es decir, quizás estemos en presencia de un modelo teórico Gaussiano.

CUARTILES:

Los cuartiles dividen a un conjunto ordenado de datos en 4 grupos de igual tamaño:

- ⇒ El cuartil 1 (Q_1) marca la parte alta del primer cuarto de los datos.
- ⇒ El cuartil 3 (Q_3) marca la parte baja del último cuarto de los datos.
- ⇒ El cuartil 2 (Q_2) corresponde a la M_e.

Metodología para el cálculo de Q_1 y Q_3

Paso 1: Ordene los datos de menor a mayor y encuentre la M_e.

Paso 2: Divida los datos en 2 mitades, por encima y por debajo de la M_e. Si n es impar incluya la mediana en ambas mitades.

Paso 3: Encuentre la mediana en ambas mitades, estas son Q_1 y Q_3.

PERCENTIL:

Los percentiles dividen a un conjunto ordenado de datos en 100 grupos de igual tamaño.

P_α, el percentil de orden α, corresponde al valor de la variable que es mayor o igual al α% de los datos y es menor o igual que el $(100-\alpha)$% de los datos (ordenados de menor a mayor).

Al Q_1 se le denomina también percentil 25, al Q_2 percentil 50 y al Q_3 percentil 75.

La mediana acumula el 50% y por sobre este valor se acumula el otro 50%. Corresponde al Q_2, P_{50}, D_5.

El cálculo exacto del percentil es complicado realizarlo a mano, claro está que los ordenadores pueden hacerlo. Una forma de obtener una aproximación (lineal) del percentil es mediante la fórmula:

$$P_\alpha = LI + \left(\frac{\frac{n\alpha}{100} - N_{i-1}}{n_i} A \right),$$

donde $\alpha \in [0, 100]$.

$LI =$ Límite inferior del intervalo donde se encuentra P_α,

$N_{i-1} =$ Frecuencia absoluta acumulada del intervalo anterior donde se encuentra P_α,

$n_i =$ Frecuencia absoluta del intervalo donde se encuentra P_α,

$A =$ Amplitud del intervalo donde se encuentra P_α,

$n =$ Tamaño de la muestra.

Calcular del ejercicio la clase anterior el P_{40} : Trabajamos en el 4to intervalo [0,775 ; 0,913], porque aquí se encuentra el Percentil 40.

$$P_{40} = 0{,}775 + \left(\frac{\frac{(30 \times 40)}{100} - 9}{9} \times 0{,}1383 \right) = 0{,}8211$$

Donde el 40% de las barras de hierro contienen a lo más 0,82 gramos de Silce.

RESUMEN CLASE N° 4 (02-04-2018)

MEDIDAS DE DISPERSIÓN

RANGO:

Rango (R): Corresponde a la diferencia entre el mayor y menor de los datos.

$$R = \text{Máx} - \text{Mín}$$

$\Rightarrow$ Su cálculo es afectado por la existencia de datos atípicos.

Rango Intercuartil (RI): Esta medida de variabilidad es resistente a valores atípicos y se concentra en el 50% central de los datos.

$$RI = Q_3 - Q_1$$

Bajo el Q_1 tengo el 25% de los datos, Sobre el Q_1 tengo el 75% de los datos
Bajo el Q_3 tengo el 75% de los datos, Sobre el Q_1 tengo el 25% de los datos
Entonces entre el Q1 y Q3 tengo el 50% de los datos

VARIANZA MUESTRAL ESTIMADA (s^2):

La varianza de las observaciones $x_1, x_2, ..., x_n$ es

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

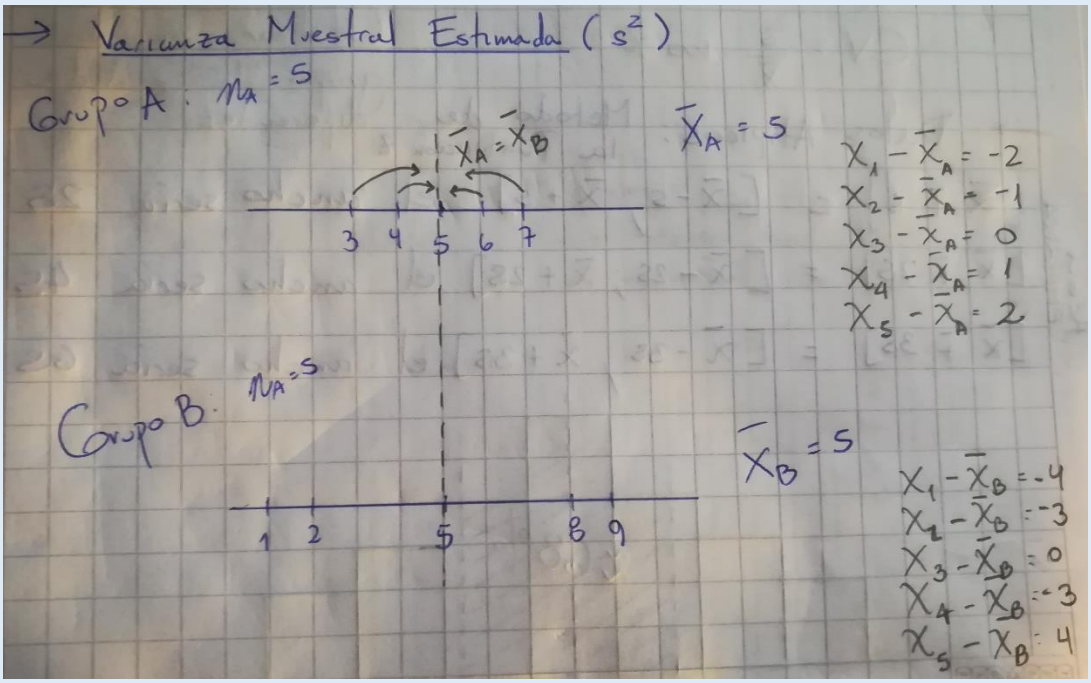
Por otro lado, si los datos se encuentran tabulados, la varianza se obtiene de la siguiente forma:

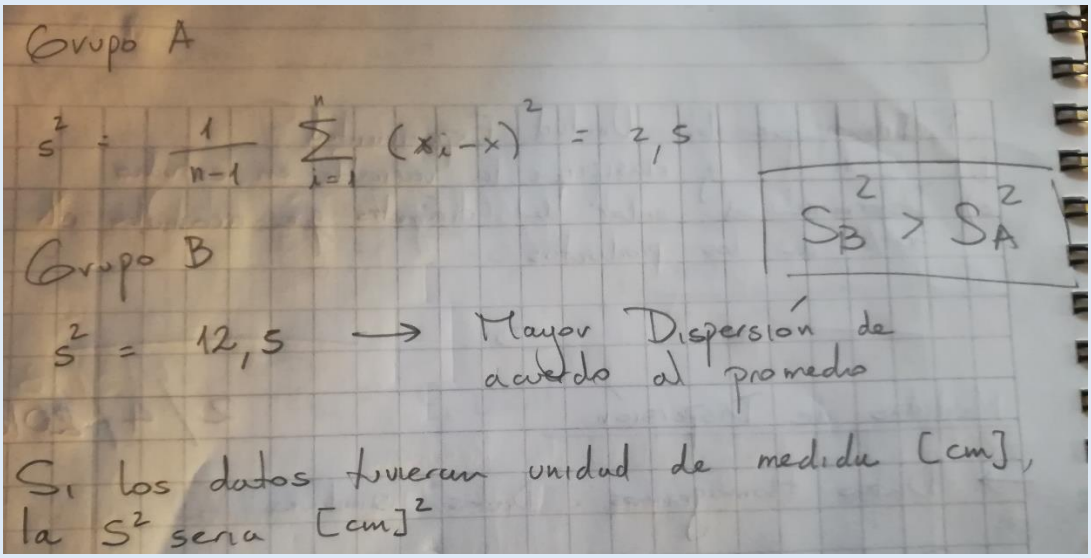
$$s^2 \approx \frac{n}{n-1}\sum_{i=1}^{n} f_i(m_i - \overline{x})^2.$$

- $\Rightarrow$ Su cálculo es afectado por la existencia de datos atípicos.
- $\Rightarrow$ El uso de esta medida es exclusivamente para variables cuantitativas.
- $\Rightarrow$ Valores grandes de s^2 $\Rightarrow$ una alta variabilidad.
- $\Rightarrow$ Si los datos corresponden a los de una población $\Rightarrow$ para calcular la varianza poblacional (σ^2) se reemplaza el factor $\frac{1}{n-1}$ por $\frac{1}{n}$.
- $\Rightarrow$ Se define la **desviación estándar muestral** (típica) como $s = \sqrt{s^2}$.

Lo atractivo de la ***desviación estándar*** es que traemos de regreso a la unidad de medida original, que se encuentra al cuadrado.

No porque dos bases de datos tengan el mismo promedio, significa que tengan el mismo comportamiento en cuanto a la dispersión (Lejanía de la información con respecto al centro). Nos centraremos en $(x_i - \overline{x})$ de la fórmula.





COEFICIENTE DE VARIACIÓN (CV):

Corresponde a una medida de dispersión relativa a la media. Está dada por

$$CV = \frac{s}{\overline{x}} 100\%$$

- ⇒ No depende de la unidad de medida de los datos.
- ⇒ Útil para comparar variabilidad de grupos que poseen unidades de medidas distintas.
- ⇒ Mientras más pequeño es el valor del CV más homogéneos son los datos.

Es una medida adimensional (no posee unidad de medida). Se hace atractivo utilizarlo cuando tengo dos bases de datos con unidades de medidas diferentes (cm con Kg).

IDENTIFICACIÓN DE DATOS ATÍPICOS: MÉTODO DE LA PUNTUACIÓN Z

El uso que tiene el promedio y la desviación estándar para detectar o identificar datos atípicos.

Regla empírica Para un conjunto de valores que tienen un histograma en forma de ***campana***, el intervalo:

$\overline{x} \mp s \quad \rightarrow$ contiene aprox. al 68% de los valores
$\overline{x} \mp 2s \quad \rightarrow$ contiene aprox. al 95% de los valores
$\overline{x} \mp 3s \quad \rightarrow$ contiene aprox. al 100% de los valores

Método de la puntuación z

Si consideramos la regla empírica, sabemos que aproximadamente el 100% de los datos está en el intervalo $[\overline{x} - 3s; \overline{x} + 3s]$. Es muy improbable que un dato esté fuera de este intervalo, y en caso que fuese, éste se llamaría un dato atípico. Es decir, un dato es no atípico si

$$x_i \in [\overline{x} - 3s; \overline{x} + 3s] \Leftrightarrow \frac{x_i - \overline{x}}{s} \in [-3; 3] \Leftrightarrow \left| \frac{x_i - \overline{x}}{s} \right| \leq 3$$

$\therefore$ Si consideramos la transformación $z_i = \frac{x_i - \overline{x}}{s}$, entonces un dato x_i es atípico si $|z_i| > 3$.

¿En qué intervalo yo tengo que prestar atención para ver si un punto se escapa? En el 3er intervalo.

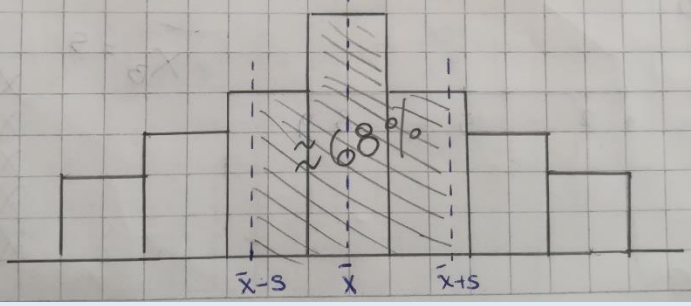
→ Datos Atípicos: Método de la Puntuación z

REGLA EMPÍRICA

$[\bar{x} \mp s] = [\bar{x} - s ; \bar{x} + s]$ el ancho sería 2S

$[\bar{x} \mp 2s] = [\bar{x} - 2s ; \bar{x} + 2s]$ el ancho sería 4S

$[\bar{x} \mp 3s] = [\bar{x} - 3s ; \bar{x} + 3s]$ el ancho sería 6S



PROARTE

NO ATIPICO

$x_i \in [\bar{x} - 3s ; \bar{x} + 3s]$

$\Leftrightarrow \bar{x} - 3s \leq x_i \leq \bar{x} + 3s \;/ -\bar{x}$

$-3s \leq x_i - \bar{x} \leq 3s \;/ \cdot \frac{1}{s}$

$-3 \leq \frac{x_i - \bar{x}}{s} \leq 3$ → z_i → Puntuación z

$\Leftrightarrow -3 \leq z_i \leq 3$

$\Leftrightarrow \boxed{|z_i| \leq 3}$

ATIPICO

$|z_i| > 3$

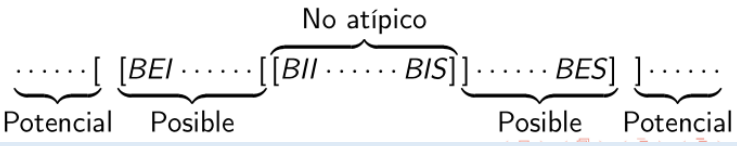
IDENTIFICACIÓN DE DATOS ATÍPICOS: MÉTODO DE TUKEY

Método de Tukey: Considere las siguientes barreras (bisagras),

- ✓ **Barrera Interior Inferior:** $BII = Q_1 - 1,5RI$
- ✓ **Barrera Interior Superior:** $BIS = Q_3 + 1,5RI$
- ✓ **Barrera Exterior Inferior:** $BEI = Q_1 - 3RI$
- ✓ **Barrera Exterior Superior:** $BES = Q_3 + 3RI$

Entonces,

- ⇛ Cualquier valor fuera de las barreras interiores es considerado como un posible valor atípico.
- ⇛ Cualquier valor fuera de las barreras exteriores es considerado como un potencia ico.



IDENTIFICACIÓN DE DATOS ATÍPICOS: DIAGRAMA DE CAJA

Definición (Diagrama de Caja o Cajón con Bigote)

El diagrama de caja, entrega información sobre

- la tendencia central y dispersión de los datos,
- la asimetría de los datos,
- identifica valores atípicos y
- es útil para comparar dos o más distribuciones.

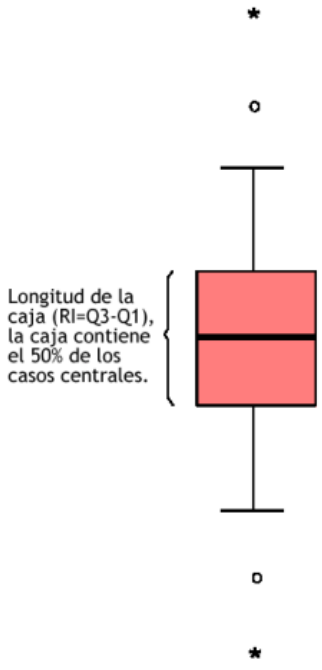
Procedimiento para realizar esta gráfica

Paso 1: Los bordes de la caja se representan por Q_1 y Q_3, se debe trazar una linea vertical que atraviese la caja en la M_e.

Paso 2: Trazar líneas (bigotes) desde los bordes de la caja hasta los valores adyacentes (el menor y mayor de los datos no atípicos).

Paso 3: marque los posibles valores atípicos con o y los potenciales con $*$.

Detalles de un Diagrama de Caja



BES=Q_3+3*RI: Potenciales valores atípicos, valores alejados más de 3 longitudes de caja del Q_3.

BIS=Q_3+1,5*RI: Posibles valores atípicos, valores alejados más de 1,5 longitudes de caja del Q_3.

Valor más grande que no llega a ser atípico.

Q_3: Cuartil 3, valor bajo el cual se acumula el 75% de los casos.

Me: Mediana

Q_1: Cuartil 1, valor bajo el cual se acumula el 25% de los casos.

Valor má pequeño que no llega a ser atípico

BII=Q_1-1,5*RI: Posibles valores atípicos, valores alejados más de 1,5 longitudes de caja del Q_1.

BEI=Q_1-3*RI: Posibles valores atípicos, valores alejados más de 1,5 longitudes de caja del Q_1.