.UBA económicas
FACULTAD DE CIENCIAS ECONÓMICAS

Estadística I

Estadística Descriptiva

Natalia SALABERRY

Estadística Descriptiva

↓

Nos permite describir un conjunto de datos para luego sacar conclusiones y así obtener información acerca del suceso al cual están asociados los datos.

↓

Dato

↓

Es la menor unidad de representación de un suceso o parte de el.
Surge como consecuencia de una observación o medición del suceso.

↓

Ejemplo:
Objetivo obtener información de desempleo en una fábrica
Datos: cantidad de empleados activos y despedidos de la fábrica
Suceso: despidos en la fábrica
Medición: representación de los despidos.
Por ejemplo tasa de desempleo =
empleados despedidos/ (empleados activos + empelados despedidos)
Información: la tasa de desempleo es x% en la fábrica

Metodología

Conjunto de pasos que me permitirán obtener información a partir de los datos

Diseño → del plan y desarrollo del proceso de investigación

Descripción → de los datos, a través de la exploración

1. Organizar los datos
2. Procesar los datos
3. Obtener características relevantes de los datos
4. Presentar los resultados

Inferencia → de un suceso, a partir de los datos

Conclusión → toma de decisiones entorno al suceso

Descripción

Los Los datos pueden ser de diferente tipo:

- **Cualitativos**: representan una cualidad del objeto o persona medida. Por ejemplo: Color de ojos. A su vez estos pueden ser de dos tipos:
  - **Nominales**: no poseen una jerarquía y permiten identificar a partir de una categoría. Por ejemplo: Cantidad de personas según la actividad que ejercen (deportistas, estudiantes, entre otros)
  - **Ordinales**: poseen un orden dentro de la categoría. Por ejemplo: Cantidad de personas según nivel primerio alcanzado (1-primer grado, 2- segundo grado, y así sucesivamente)
- **Cuantitativos**: son datos numéricos que surgen de una medición o conteo. Por ejemplo cantidad personas, altura de las personas, entre otros. A su vez estos pueden ser de dos tipos:
  - **Discretos**: generalmente surgen de un conteo. Por ejemplo: Cantidad de personas.
  - **Continuos**: generalmente surgen de una medición. Por ejemplo: Altura de personas

Descripción

Según el tipo de datos, se requiere saber su cantidad y porcentaje. Entonces surgen dos conceptos:

Frecuencia Absoluta f_i	→	**Cantidad** de veces que aparece un valor dentro del conjunto de datos
Frecuencia Relativa f_{r_i}	→	**Proporción** de veces que aparece un valor dentro del conjunto de datos

También, es útil poder contar con la frecuencia de valores menores o iguales a un determinado valor. Para ello se definen dos conceptos:

Frecuencia Absoluta acumulada F_i	→	Es la suma de las frecuencias de todos los valores menores a x_i
Frecuencia Relativas acumulada F_{ri}	→	Es la suma de las frecuencias relativas de todos los valores menores a x_i

Descripción

x		f_i	$f_{r_i} = \frac{f_i}{n}$
2			
3	2 -> aparece **2** veces	2	2/7
4			
5	3 -> aparece **2** veces	2	2/7
3			
4	4 -> aparece **2** veces	2	2/7
2			
	5 -> aparece **1** vez	1	1/7

Ejemplo

x_i	f_i	F_i	f_{r_i}	F_{r_i}	F_{r_i} *en %*
2	2	2	2/7	2/7	28,57
3	2	4	2/7	4/7	57,14
4	2	6	2/7	6/7	85,71
5	1	7	1/7	7/7	100

Descripción

A partir de las frecuencia relativa y frecuencia absoluta podemos determinar la **distribución de frecuencia relativa y la distribución de frecuencia absoluta**. Estas nos permitirán conocer características sobre como se distribuyen los datos.

Normalmente surgen de realizar un análisis gráfico. Para ello se utilizan los siguientes gráficos:

- Gráfico de barras
- Gráfico de Polígonos
- Gráfico circulares
- Histograma
- Boxplot

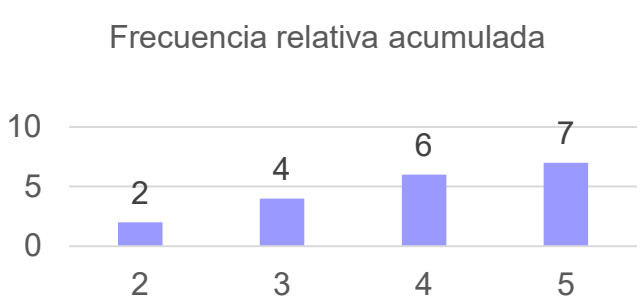


Gráfico de barras

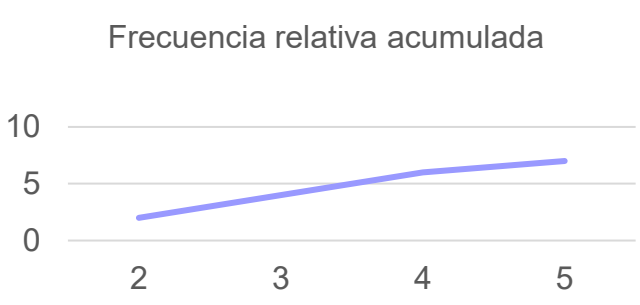


Gráfico de polígono

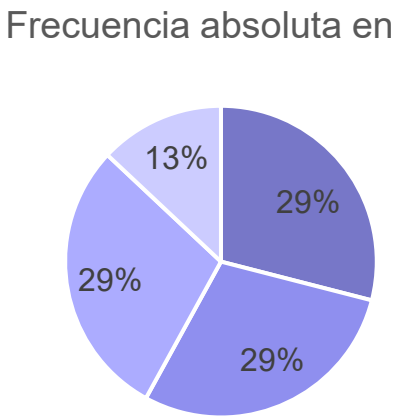


Gráfico circulares

Descripción

La descripción de los datos puede realizarse a través de diferentes medidas. Estas proporcionarán características que permitirán tener un mayor conocimiento de los mismos.

Hay tres tipos de medidas:

- **Medias de posición:** refieren a entorno de que valor se encuentran distribuidos los datos
  - **De tendencia central**: indican el centro de la distribución de frecuencia
  - **De tendencia no central**: brindan información de ciertos valores, no necesariamente centrales

- **Medias de dispersión:** refieren a que tan dispersos se encuentran los datos respecto de su valor central

- **Medias de forma:** refieren a la forma que tiene la distribución de los datos. Permiten saber si hay tendencia de que los datos se agrupen hacia uno de los lados respecto de un valor central y la probabilidad de ocurrencia del valor central

Descripción

Medias de posición – De tendencia central

- **MEDIA MUESTRAL**: Es el promedio simple de los datos. Es muy sensible a valores atípicos siendo aquellos que se alejan del comportamiento "normal" de los datos

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **MEDIANA**: Es el valor central (mitad=50%) cuando los datos se encuentran ordenados de manera creciente.

Para obtener su valor se debe primero calcular:

Distancia=k(n-1) donde k es el porcentaje que en este caso es 0,5*

Si:

Distancia=número entero => es la posición. Buscaré en el ´índice de mis datos dicha posición y me fijare el valor que toma la variable siendo este el valor de la mediana

Distancia=número decimal => debo interpolar . El resultado de esta operación será el valor de la mediana

- **MODA**: Es el valor más frecuente, es decir, el de mayor frecuencia absoluta y relativa. Si dos o más valores son los de mayor frecuencia entonces estamos frente a una distribución multimodal

X_{mo}

Descripción

→ Medidas de Posición

→ De tendencia central

Ejemplo

x
2
3
4
5
2
16

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(2+3+4+5+2)}{5} = 3{,}2$$

X_{mo}=2 siendo el valor que se repite más veces

Posición	x
0	2
1	2
2	3
3	4
4	5
	16

Para calcular la mediana

Distancia=k(n-1) donde k es el porcentaje*

=0,5(5-1)=2 si es entero directamente es la posición =>* X_2=3

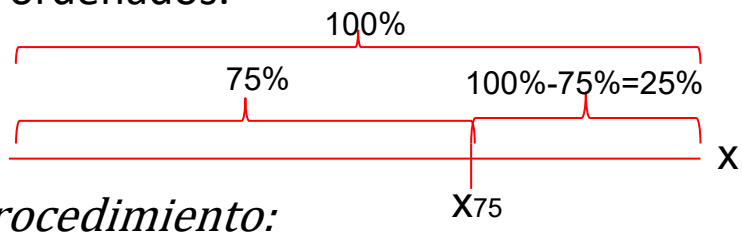
Por lo tanto, la mediana es igual a 3

Descripción

Medias de posición – De tendencia no central

- **PERCENTILES**: Permiten saber el porcentaje de observaciones que se encuentran por debajo de un valor determinado y el (100-k)% por encima (dividen la distribución en 100 partes iguales). Los datos siempre deben estar ordenados.

$$P_k$$



Para hallar su valor, debo seguir el siguiente procedimiento:

Distancia=k(n-1) donde k es el porcentaje que en este caso es 0,5*

Si:

Distancia=número entero => es la posición. Buscaré en el ´índice de mis datos dicha posición y me fijare el valor que toma la variable siendo este el valor de la mediana

Distancia=número decimal => debo interpolar . El resultado de esta operación será el valor de la mediana

- **CUARTILES**: dividen la distribución de frecuencias en cuatro partes iguales.

Para hallar su valor, opero igual que con los percentiles

Descripción

Medias de posición – De tendencia central

- **DECILES**: dividen la distribución de frecuencias en diez partes iguales.

$$D_k$$

Para hallar su valor, opero igual que con los percentiles

- Equivalencias

$P_{50} = Q_2 = D_5 = X_{me}$

$P_{25} = Q_1$

$P_{75} = Q_3$

$P_{10} = D_1$

$P_{20} = D_2$ y así sucesivamente

Descripción → Medidas de Posición → De tendencia no central → Ejemplo

Orden	X
0	2
1	2
2	3
3	4
4	5

Decil 5: Distancia=k(n-1)=0,5*(5-1)=2 si es entero directamente es la posición =>* X_2=3 Entonces, el decil 5 es igual a 3

Cuartil 2: Distancia=k(n-1)=0,5*(5-1)=2 si es entero directamente es la posición =>* X_2=3 Entonces, el cuartil 2 es igual a 3

Percentil 50: Distancia=k(n-1)=0,5*(5-1)=2 si es entero directamente es la posición =>* X_2=3 Entonces, el percentil 50 es igual a 3

Y además coinciden con la mediana, lo cual es correcto ya que en todos los casos estoy buscando el valor que se corresponde con la mitad de la distribución de los datos

Descripción → Medidas de Posición → De tendencia no central

Posición	X
0	2
1	2
2	3
3	4
4	5

Por ejemplo, si queremos calcular P_{70}, se calcula la distancia

Distancia=k(n-1)=0,70*(5-1)=2,8*

=> como no es entero, se debe interpolar

Interpolación: Xparteentera+ Partedecimal(Xposterior – Xparteentera)*

Xparteenterea= X_2=3 *Xposterior=* $X_{parteentera+1}$ $=X_3$=4

ParteDecimal 0,8

Luego *3 + 0,8*(4– 3) =3+[0,8* (1)]=3,8 =>* P_{70} =3,8 siendo el valor del pecentil 70

Descripción

Medias de dispersión

- **RANGO**: Diferencia entre el máximo y mínimo valor de los datos (longitud del intervalo)

$$Rango = x_{max} - x_{min}$$

- **RANGO INTERCUARTILICO**: Longitud del intervalo donde se encuentran el 50% de los datos.

$$RIC = Q_3 - Q_1 = P_{75} - P_{25}$$

- **VARIANZA MUESTRAL**: indica cuanto se alejan los datos de su media en términos cuadráticos

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

- **DESVÍO ESTANDAR:** indica cuanto se alejan los datos de su media en la misma unidad de medida que los datos. Se calcula como la raíz cuadrada de la varianza muestral.

- **COEFICIENTE DE VARIACIÓN:** Mide la relación entre la media y el desvío de la muestra. En general es útil para comparar dos o más conjuntos de datos. Cuanto mayor es el CV, mayor es la dispersión y por lo tanto menos representativa es la media

$$CV = \frac{S}{\bar{X}}$$

Descripción → Medidas de Dispersión

Ejemplo

$RIC = P_{75} - P_{25}$= $X_3 - X_1$=4-2 = 2

Orden	X
0	2
1	2
2	3
3	4
4	5

P_{75} : Distancia=k(n-1)=0,75*(5-1)=3 como es entero entonces directamente es la posición =>*$P_{75} = X_3 = 4$

P_{25} : Distancia=k(n-1)=0,25*(5-1)=1 como es entero entonces directamente es la posición =>*$P_{25} = X_1 = 2$

Descripción → Medidas de Dispersión

Ejemplo

Orden	X	$(x_i - \bar{X})^2$
0	2	1,44
1	2	1,44
2	3	0,04
3	4	0,64
4	5	3,24
	16	6,8

$$Rango = x_{max} - x_{min} = 5 - 2 = 3$$

$$S^2 = \frac{6,8}{5-1} = 1,7 \qquad S = \sqrt{1,7} = 1,3$$

$$CV = \frac{S}{\bar{X}} = \frac{1,3}{3,2} = 0,41$$ En general si es menor de 0,5 entonces se considera chico

Descripción

Ejemplo

Dadas las ventas de supermercados entre enero y noviembre 2022

Período	VENTAS en Millones de pesos
ene-22	$ 23,33
feb-22	$ 24,21
mar-22	$ 29,38
abr-22	$ 30,03
may-22	$ 31,50
jun-22	$ 33,92
jul-22	$ 39,73
ago-22	$ 37,06
sep-22	$ 38,92
oct-22	$ 41,67
nov-22	$ 46,61

se buscar generar información para responder las siguientes preguntas:

A) ¿Cuál es el monto promedio vendido por los supermercados?

B) ¿En qué mes el monto vendido alcanza el 50% de las ventas?

C) ¿En qué mes el monto vendido alcanza el 75% de las ventas?

D) ¿En qué mes el monto vendido es mínimo para el 25% superior de las ventas?

E) ¿En qué mes el monto vendido es máximo para el 25% de las ventas?

F) ¿Entre qué valores se encuentran las ventas de los supermercados?

G) ¿Entre qué meses se encuentra el 50% de las ventas y cuál es el monto acumulado?

H) ¿Cuál es la variabilidad de montos vendidos en el período dado?

I) ¿Cuál es la variabilidad de montos vendidos en la misma unidad de medida que los montos en el período dado?

J) ¿Recomendaría utilizar la media de los montos vendidos como medida a ser informada?

Lo primero es observar si los datos están ordenados (siempre miro la variable que en este caso es las Ventas)

Descripción

Ejemplo

Como en agosto baja el valor (es más chico que el anterior) entonces los ordeno y establezco posición

Posición	Período	VENTAS en Millones de pesos
0	ene-22	$ 23,33
1	feb-22	$ 24,21
2	mar-22	$ 29,38
3	abr-22	$ 30,03
4	may-22	$ 31,50
5	jun-22	$ 33,92
6	ago-22	$ 37,06
7	sep-22	$ 38,92
8	jul-22	$ 39,73
9	oct-22	$ 41,67
10	nov-22	$ 46,61
	Total	$ 376,36

A) ¿Cuál es el monto promedio vendido por los supermercados?

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{376{,}36}{11} = 34{,}21$$

El monto promedio vendido en el período por los supermercados es $34,21 (en millones)

B) ¿En qué **mes** el monto vendido alcanza el 50% de las ventas?

Al preguntar "alcanza el 50%" => Mediana (o Percentil 50)

Entonces, calculo distancia

Distancia=k(n-1)=0,5*(11-1)=5 => como es entero, es directamente la posición =>* $X_5 = X_{me} = \$33{,}92$ *que se corresponde a Junio 2022*

El mes donde se alcanza el 50% de las ventas es Junio 2022

Descripción

Ejemplo

Como en agosto baja el valor (es más chico que el anterior) entonces los ordeno y establezco posición

Posición	Período	VENTAS en Millones de pesos
0	ene-22	$ 23,33
1	feb-22	$ 24,21
2	mar-22	$ 29,38
3	abr-22	$ 30,03
4	may-22	$ 31,50
5	jun-22	$ 33,92
6	ago-22	$ 37,06
7	sep-22	$ 38,92
8	jul-22	$ 39,73
9	oct-22	$ 41,67
10	nov-22	$ 46,61
	Total	$ 376,36

C) ¿En qué **mes** el monto vendido alcanza el 75% de las ventas?
Al preguntar "alcanza el 75%" => Percentil 75
Entonces, calculo distancia
Distancia=k(n-1)=0,75*(11-1)=7.5 => como no es entero, interpolo*

Interpolación: Xparteentera+ Partedecimal(Xposterior – Xparteentera)*

Xparteenterea= X_7=38,92 *Xposterior=* $X_{parteentera+1}$ = X_8=39,73
ParteDecimal 0,5

Luego 38,92 + ***0,5*****(39,73 – 38,92) =* 38,92 *+[0,5* (0,81)]=39,325*
=> P_{75} = ***$39,325 => busco el más cercano***

El mes con el monto vendido que alcanza el 75% de las ventas es Julio 2022

Descripción

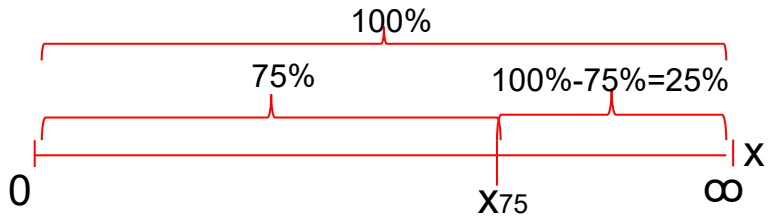
Ejemplo

Como en agosto baja el valor (es más chico que el anterior) entonces los ordeno y establezco posición

Posición	Período	VENTAS en Millones de pesos
0	ene-22	$ 23,33
1	feb-22	$ 24,21
2	mar-22	$ 29,38
3	abr-22	$ 30,03
4	may-22	$ 31,50
5	jun-22	$ 33,92
6	ago-22	$ 37,06
7	sep-22	$ 38,92
8	jul-22	$ 39,73
9	oct-22	$ 41,67
10	nov-22	$ 46,61
	Total	$ 376,36

D) ¿En qué **mes** el monto vendido es mínimo para el 25% superior de las ventas?

Si los datos que ya están ordenados fuesen representados en una recta numérica



El X75 es valor mínimo del 25% superior de las ventas de mayor monto

Entonces,

$\Rightarrow P_{75} = \$39{,}325 \Rightarrow$ *siendo el mismo que en el punto C.*

El mes en el cual es monto vendido es mínimo para el 25% superior de las ventas es Julio 2022

Descripción

Ejemplo

Como en agosto baja el valor (es más chico que el anterior) entonces los ordeno y establezco posición

Posición	Período	VENTAS en Millones de pesos
0	ene-22	$ 23,33
1	feb-22	$ 24,21
2	mar-22	$ 29,38
3	abr-22	$ 30,03
4	may-22	$ 31,50
5	jun-22	$ 33,92
6	ago-22	$ 37,06
7	sep-22	$ 38,92
8	jul-22	$ 39,73
9	oct-22	$ 41,67
10	nov-22	$ 46,61
	Total	$ 376,36

E) ¿En qué **mes** el monto vendido es máximo para el 25% de las ventas?

Al preguntar "máximo para el 25%" (alcanza) => Percentil 25 ya que es valor máximo para el 25%
Entonces, calculo distancia
Distancia=k(n-1)=0,25*(11-1)=2,5 => como no es entero, interpolo*

Interpolación: Xparteentera+ Partedecimal(Xposterior – Xparteentera)*

Xparteenterea= X_2=29,38 *Xposterior=* $X_{parteentera+1}$ $=X_3$=30,03
ParteDecimal 0,5

Luego 29,38 *+ **0,5***(*30,03 – 29,38*) =* 29,38 *+[0,5* (0,65)]=29,705*
$=> P_{25}$ *= **$29,705 => busco el más cercano => 30,03***

El mes en el que monto de las ventas alcanza el 25% es Abril 2022

Descripción

Ejemplo

Como en agosto baja el valor (es más chico que el anterior) entonces los ordeno y establezco posición

Posición	Período	VENTAS en Millones de pesos
0	ene-22	$ 23,33
1	feb-22	$ 24,21
2	mar-22	$ 29,38
3	abr-22	$ 30,03
4	may-22	$ 31,50
5	jun-22	$ 33,92
6	ago-22	$ 37,06
7	sep-22	$ 38,92
8	jul-22	$ 39,73
9	oct-22	$ 41,67
10	nov-22	$ 46,61
	Total	$ 376,36

F) ¿**Entre qué valores** se encuentran las ventas de los supermercados?
Si me preguntan "Entre que valores" entonces se trata de un **rango** (intervalo) que tiene un valor mínimo y un valor máximo
El valor mínimo es 23,33 y el valor máximo 46,61
Las ventas para el período se encuentran entre $23,33 y $46,61 millones

G) ¿**Entre qué meses** se encuentra el 50% de las ventas y cuál es el monto acumulado?
Si me preguntan "Entre que meses" entonces se trata de un **intervalo** donde acumula el 50%
RIC= P75-P25 => tengo acumulado el 50% (75%-25%=50%)
Utilizando los valores ya calculados en C y E, entonces:
Monto acumulado = P75-P25= $\$39{,}325 - \$29{,}705 = \$9{,}62$

Entre los meses de abril y julio se encuentra el 50% de las ventas que acumula un monto vendido de $\$9{,}62$ millones

Descripción

Ejemplo

Posición	Período	VENTAS en Millones de pesos	$(x_i - \bar{X})^2$
0	ene-22	$ 23,33	$ 118,47
1	feb-22	$ 24,21	$ 100,09
2	mar-22	$ 29,38	$ 23,37
3	abr-22	$ 30,03	$ 17,51
4	may-22	$ 31,50	$ 7,37
5	jun-22	$ 33,92	$ 0,09
6	ago-22	$ 37,06	$ 8,10
7	sep-22	$ 38,92	$ 22,14
8	jul-22	$ 39,73	$ 30,42
9	oct-22	$ 41,67	$ 55,58
10	nov-22	$ 46,61	$ 153,65
	Total	$ 376,36	$ 536,79

H) ¿Cuál es la **variabilidad** de montos vendidos en el período dado?
Si me preguntan por la "variabilidad" y no se aclara nada más, entonces es el concepto es varianza

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{11-1} = \frac{536,79}{10} = 53,679$$

La variabilidad de montos vendidos es de $53,679 millones

I) ¿Cuál es la **variabilidad** de montos vendidos **en la misma unidad de medida** que los montos en el período dado?
A diferencia del punto anterior ahora me aclaran "en la misma unidad", entonces es el desvío

$$S = \sqrt{S^2} = \sqrt{53,679} = 7,326595389 \approx 7,33$$

La variabilidad de montos vendidos en la misma unidad de medida que las ventas es $7,33 millones

J) ¿Recomendaría utilizar la media de los montos vendidos como medida a ser informada?
Lo que me permite saber si la media es representativa (y por tanto determinar si es recomendable informarla) es el

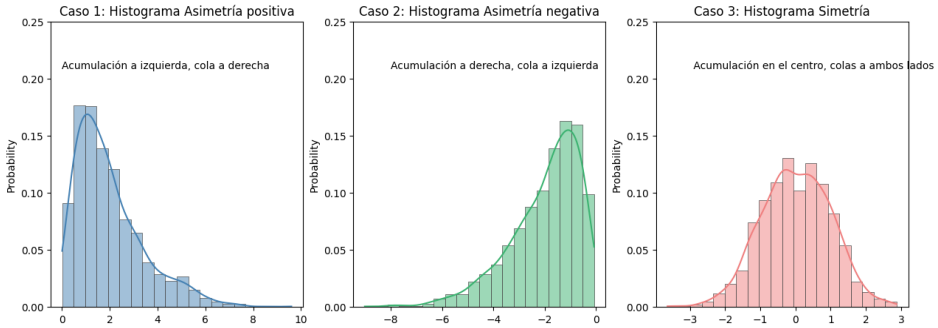
$$CV = \frac{S}{\bar{X}} = \frac{7,33}{34,21} = 0,214264834 \approx 0,21$$ => Al ser un valor bajo, es recomendable informar la media

De hecho, si se observa $\bar{X} = 34,21$ y $X_{me} = \$33,9$ son valores muy similares.

Descripción

Histograma

Un histograma es una representación visual de los datos agrupados en intervalos.
Los intervalos se denominan intervalos de clase y su forma gráfica es un rectángulo.
Al ser un intervalo tendrá un límite inferior y un límite superior por lo que los datos serán asignados entre estos límites, conformándose así varios intervalos de clase que hacen a la distribución total de los datos. De este modo permitirá observar gráficamente la forma que toma la distribución y su comportamiento.
Además, cada intervalo de clase tendrá una marca de clase que es el punto medio del intervalo (es decir, Límite inferior + Límite superior dividido 2).
La amplitud del intervalo de clase vendrá dada por la diferencia entre el límite superior y el límite inferior.
Luego, el área correspondiente a cada intervalo de clase representa la frecuencia absoluta (o relativa) de la clase. Observaremos que las sumas de las áreas de los rectángulos es 1 (o 100 %).

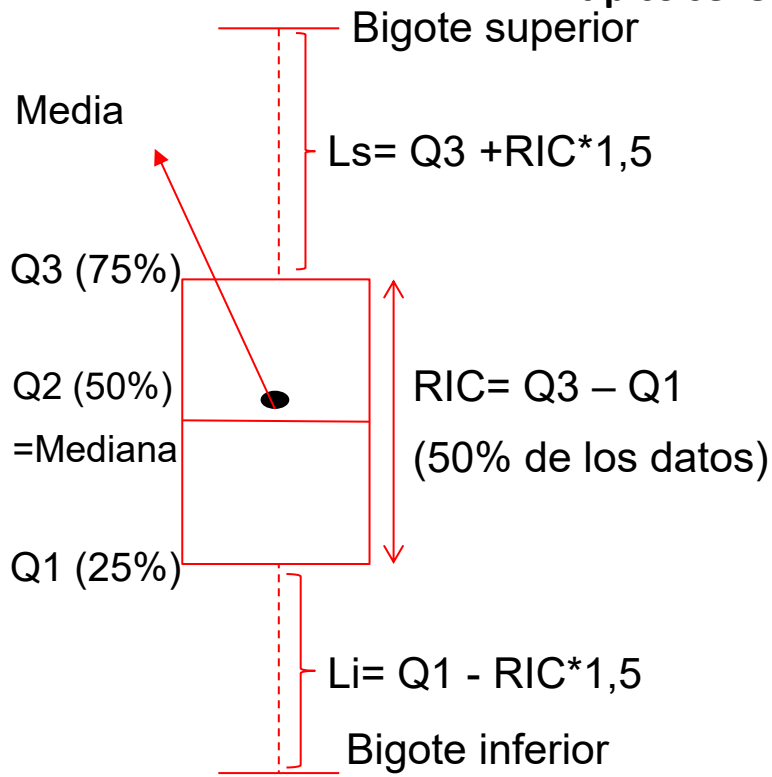


Descripción

Boxplot

Valor atípico ○

- **Atípico moderado** si se encuentra a una distancia entre el bigote superior y un umbral=Q3 +RIC*3
- **Atípico severo** si se encuentra a una distancia por encima de Q3 +RIC*3



El Box plot o diagrama de caja y bigotes es un método para representar gráficamente una serie de datos numéricos a través de sus cuartiles. El diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos.

Este tipo de gráficos nos permite identificar valores atípicos y comparar distribuciones además de conocer de una forma cómoda y rápida como se distribuyen el 50% de los valores centrales.

A partir de un box-plot podemos apreciar los siguientes aspectos de la distribución de datos: posición, dispersión, asimetría, longitud de las colas, Valores atípicos (outliers)

Valor atípico ○

- **Atípico moderado** si se encuentra a una distancia entre el bigote inferior y un umbral=Q1 -RIC*3
- **Atípico severo** si se encuentra a una distancia por debajo de Q1 - RIC*3

Descripción

Medias de forma

- **COEFICIENTE DE ASIMETRÍA**: Permite saber el tipo de asimetría de una distribución, es decir, cuan desviados se encuentran los datos respecto de la media

$$A_s = \frac{n}{(n-1)*(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{S} \right)^3$$

Si A_s=0 distribución simétrica
Si A_s >0 asimétrica positiva
SI A_s <0 asimétrica negativa

$Si\ x_{mo} < x_{me} < \bar{X}$ => asimétrica positiva (sesgo a derecha)

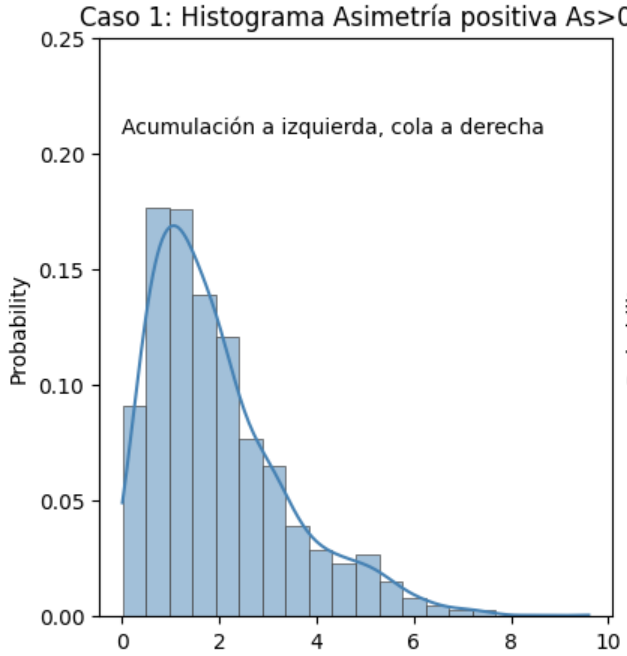
$Si x_{mo} = x_{me} = \bar{X}$ => simétrica

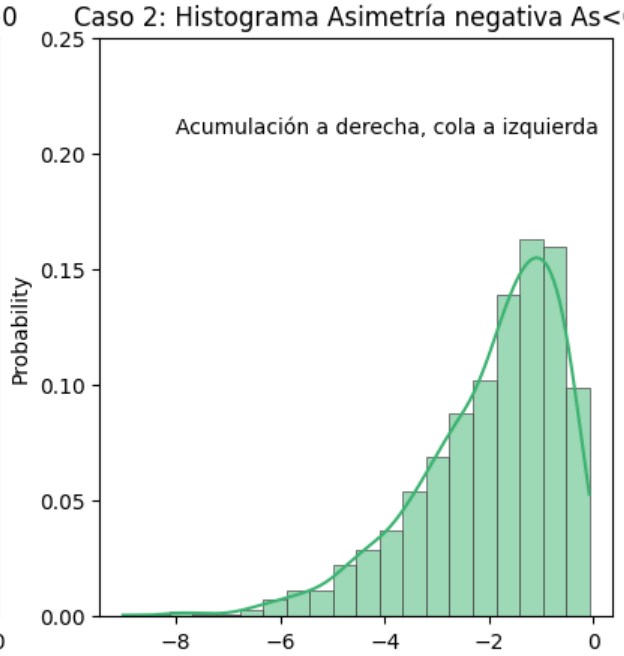
$Si\ x_{mo} > x_{me} > \bar{X}$ => asimétrica negativa (sesgo a izquierda)

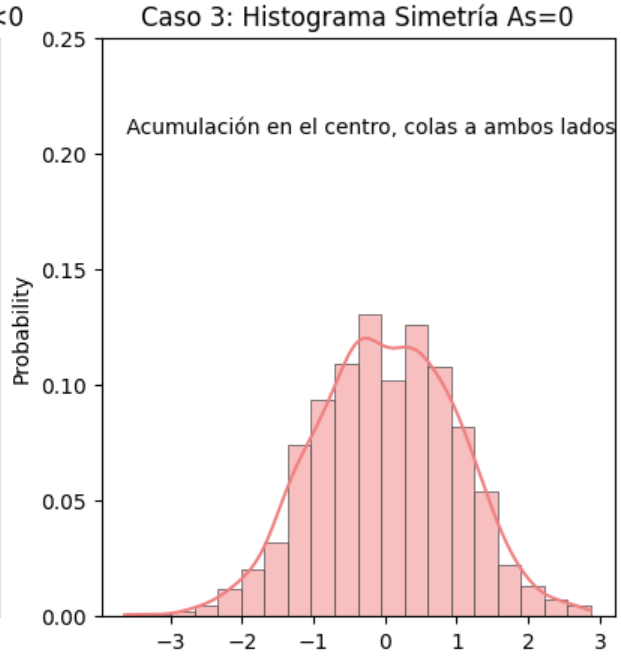
Descripción

Histograma

La simetría o asimetría es fácilmente observable en un histograma







Sesgo a derecha:

- Cola de la distribución a derecha.
- Mayor concentración de datos a izquierda

Sesgo a izquierda:

- Cola de la distribución a izquierda.
- Mayor concentración de datos a derecha

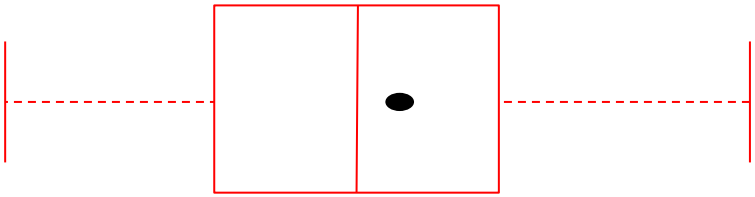
Sin sesgo:

- Colas a ambos lados.
- Mayor concentración de datos en el centro

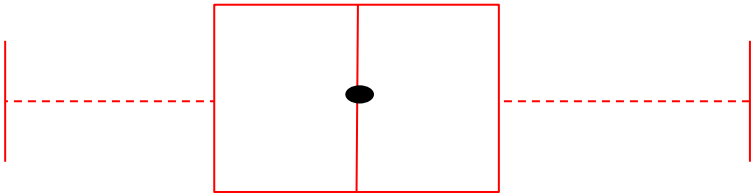
Descripción

También podemos ver la simetría o asimetría en un boxplot

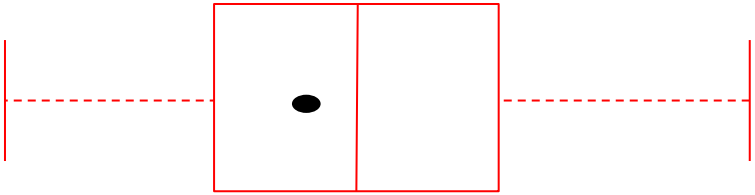
(siempre girar un boxplot vertical siempre hacia la derecha)



$Como\ x_{me} < \bar{X}$ => asimétrica positiva (sesgo a derecha)



$Como\ x_{me} = \bar{X}$ => simétrica



$Como\ x_{me} > \bar{X}$ => asimétrica negativa (sesgo a izquierda)

Descripción

Medias de forma

- **COEFICIENTE DE CURTOSIS**: Permite saber cómo es la concentración de datos (cuan puntiaguda es la forma de la distribución)

$$k = \left[\frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{S}\right)^4\right] - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$

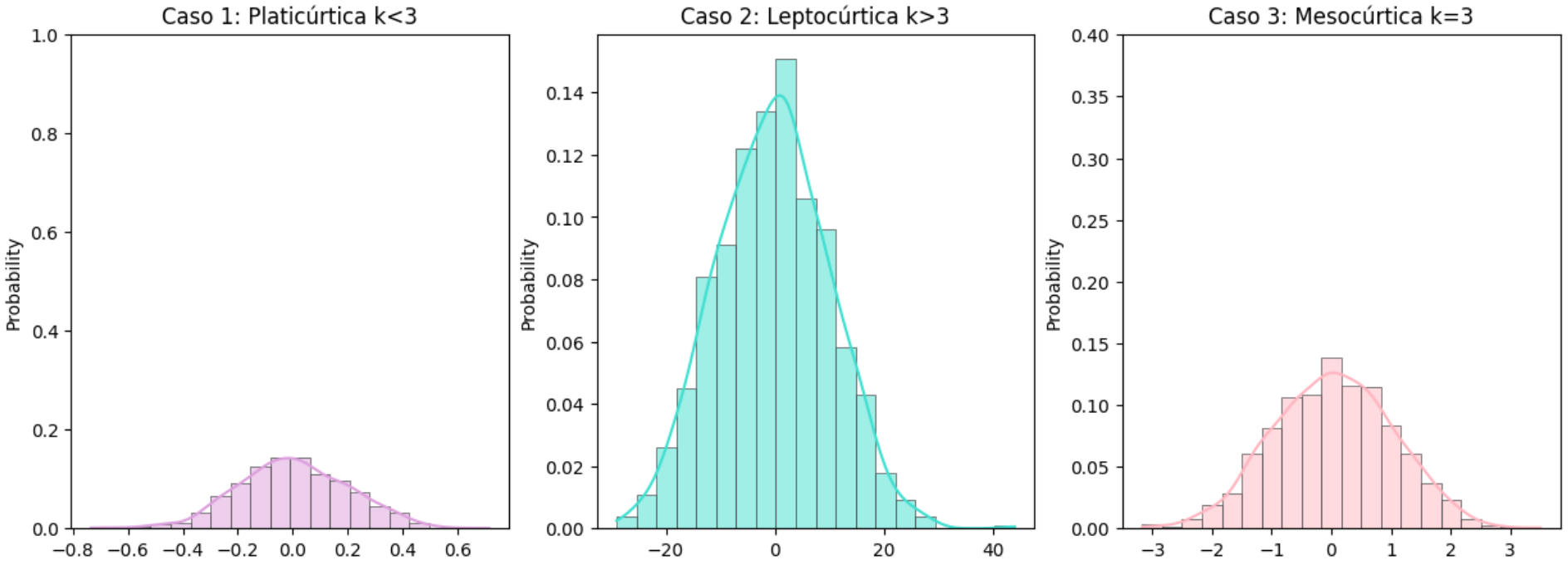
Si k=3 mesocúrtica
Si k >3 leptocúrtica
SI k <3 platicúrtica

Si la distribución es muy putiaguda entonces es leptocúrtica
Si la distribución es chata entonces es platicúrtica
Si la distribución no es ni muy chata ni muy puntiaguda entonces es mesocúrtica

Descripción

Histograma

En un histograma podemos observar la curtosis (respecto de la concentración de los datos alrededor de un valor central – generalmente la media de la distribución-)



Descripción → Medidas de Forma

Ejemplo

Continuando con el ejemplo de la clase anterior

x	$\left((x_i - \overline{X})/S\right)^3$
2	-0,786527082
2	-0,786527082
3	-0,003641329
4	0,233045061
5	2,654528903
16	1,310878471

$$A_s = \frac{n}{(n-1)*(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{S}\right)^3 = \frac{5}{(5-1)*(5-2)} * 1{,}310878471$$

$$= 0{,}546199362 > 0 \Rightarrow \text{asimétrica positiva}$$

x	$\left((x_i - \overline{X})/S\right)^4$
2	0,726024999
2	0,726024999
3	0,000560204
4	0,143412346
5	3,675501558
16	5,271524106

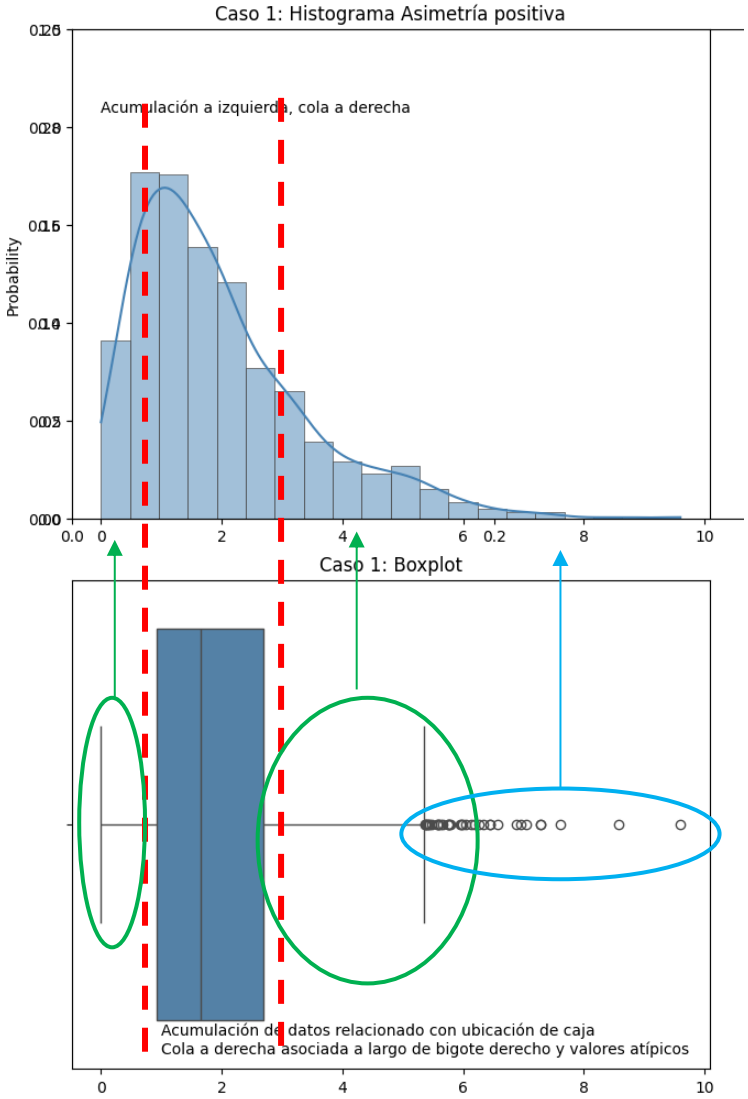
$$\left[\frac{n*(n+1)}{(n-1)*(n-2)*(n-3)} * \sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{S}\right)^4\right] - \frac{3*(n-1)^2}{(n-2)*(n-3)}$$

$$= \left[\frac{5*(5+1)}{(5-1)*(5-2)*(5-3)} * 5{,}271524106\right] - \frac{3*(5-1)^2}{(5-2)*(5-3)}$$

$$= 6{,}589405133 - 8 = -1{,}410594867 < 3 \Rightarrow \text{platicúrtica}$$

Descripción

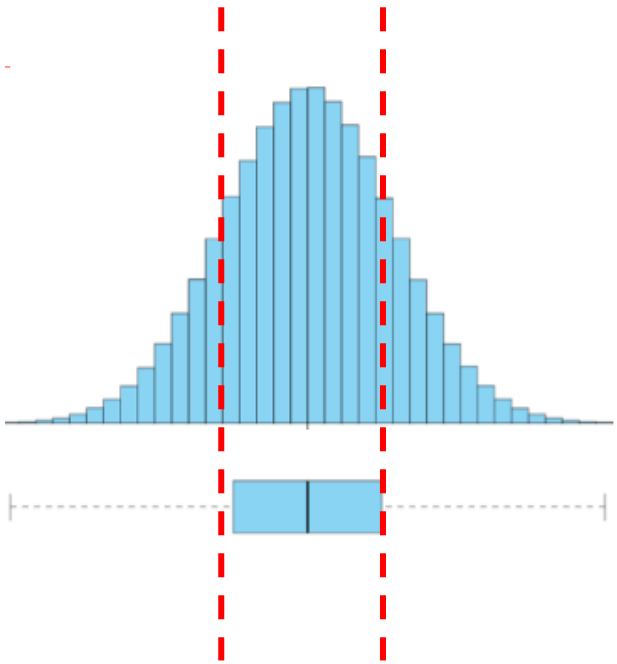
Podemos comparar ambos gráficos del siguiente modo:



El histograma muestra una asimetría a derecha. Es decir, mayor concentración de datos a izquierda y cola de distribución a derecha.

Tales características también las vemos representadas en el boxplot: la concentración del 50% de los datos (caja del boxplot) se encuentra a izquierda (concentración de datos a izquierda)). Y el bigote superior es más largo a derecha (cola de distribución a derecha). Y además se encuentran los valores atípicos, que en el histograma se ven representados en las últimas barras de la derecha (de menor tamaño)

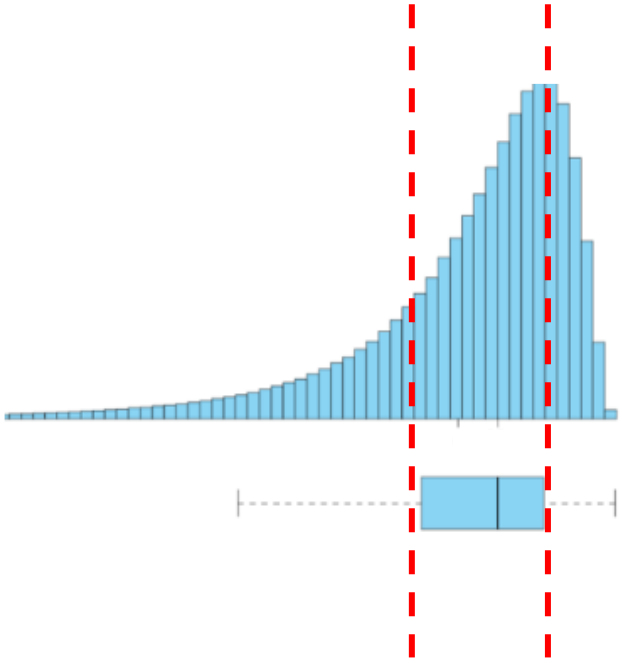
Descripción



En el histograma se observa simetría.

En el bloxplot también, ya que la caja se encuentra en el medio de todo el largo (desde bigote inferior hasta bigote superior) del boxplot.

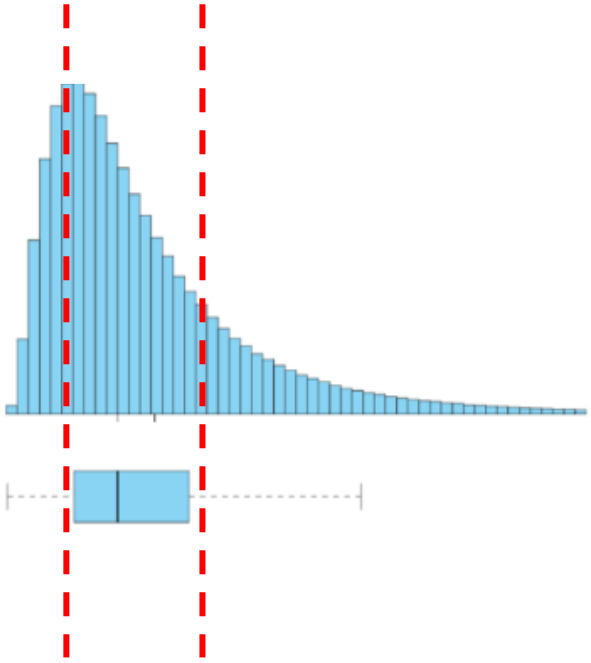
Además, la mediana se encuentra en el centro de la caja de boxplot lo que indica simetría.



En el histograma se observa asimetría a izquierda.

En el bloxplot también, ya que la caja se encuentra hacia la derecha y el bigote inferior es más largo que el superior.

Además, la mediana se encuentra más cercana al límite superior de la caja (Q3) lo que indica asimetría a izquierda.



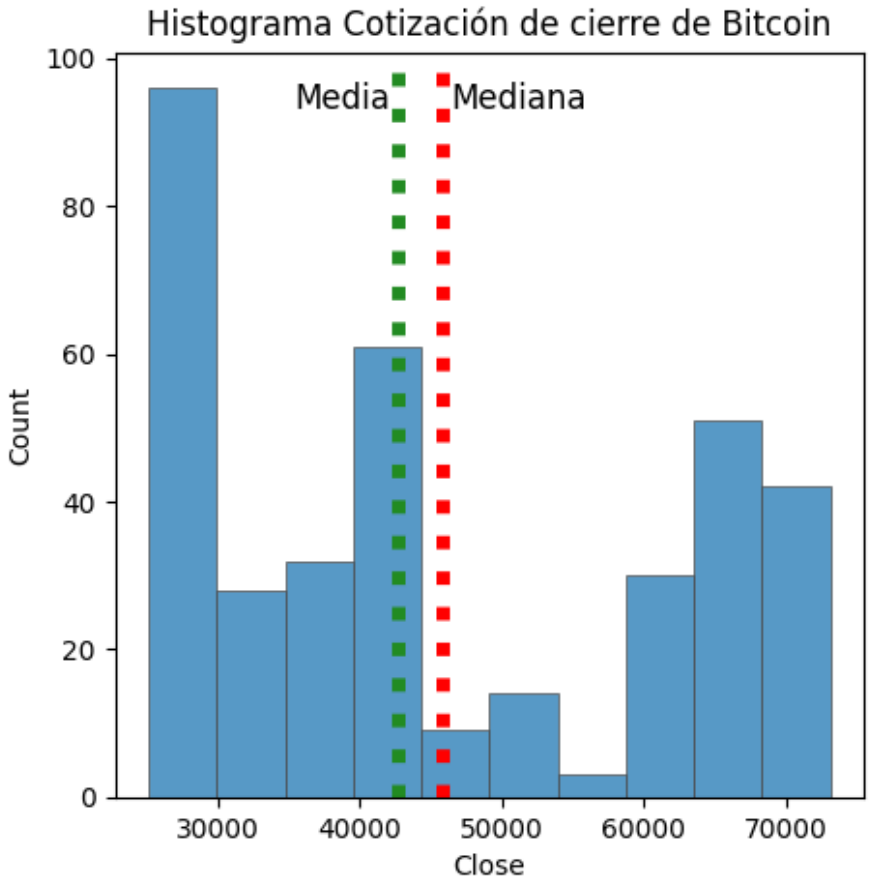
En el histograma se observa asimetría a derecha.

En el bloxplot también, ya que la caja se encuentra hacia la izquierda y el bigote superior es más largo que el inferior.

Además, la mediana se encuentra más cercana al límite inferior de la caja (Q1) lo que indica asimetría a derecha.

Descripción

EJEMPLO: Observando el siguiente histograma de la cotización diaria de cierre del Bitcoin entre 1/7/2023 y 1/7/2024 en dólares, responder acerca de las siguientes afirmaciones **justificando** adecuadamente su respuesta.



A) “Puede afirmarse que la distribución de la cotización del BTC es asimétrica positiva”

Falso, justificar

B) “Dado un valor del coeficiente de curtosis de -1,48 entonces la distribución es platicúrtica”

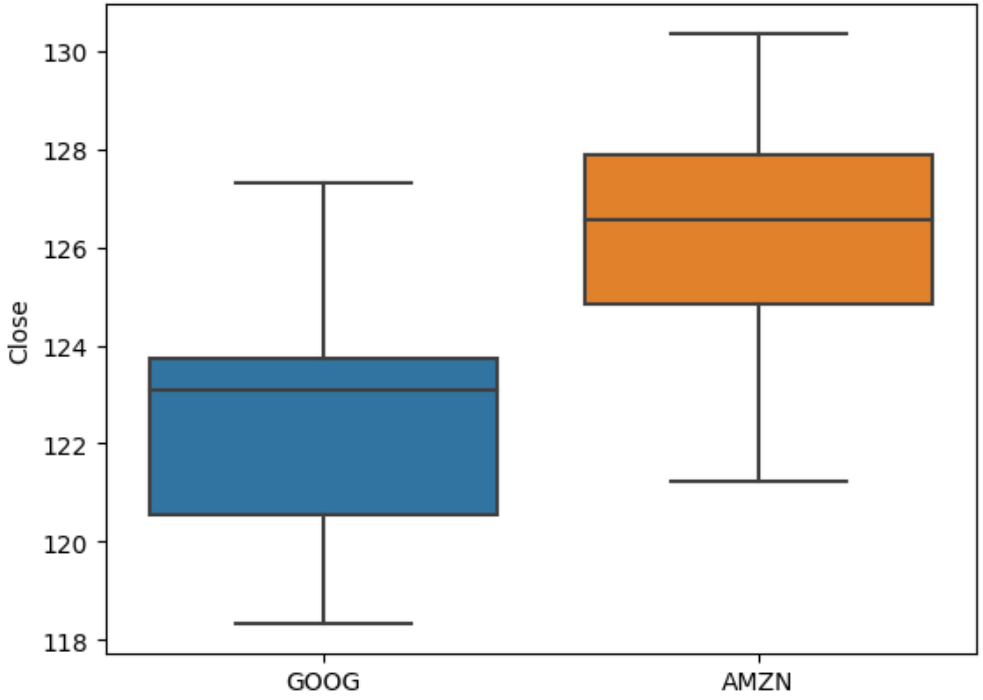
Verdadero, justificar

C) “Podría decirse que no hay valores atípicos”

Verdadero, justificar

Descripción

EJEMPLO: Se tomaron las cotizaciones de Google (GOOG) y Amazon (AMZN) para el período 1/6/2023 a 1/7/2023. Los boxplots correspondientes se encuentran en la siguiente imagen. Se dispone además de algunas medidas estadísticas.



Estadística	N	MEDIA	MEDIANA	DESVÍO	MIN	MAX
AMZN	21.0	126.4	126.6	2.45	121.2	130.4
GOOG	21.0	122.5	123.1	2.43	118.3	127.3

Estadística	P05	P10	P25	P50	P75	P90
AMZN	122.8	123.4	124.8	126.6	127.9	129.3
GOOG	118.3	119.1	120.6	123.1	123.7	125.1

Descripción

Analizar si las siguientes afirmaciones son verdaderas o falsas. Justificar

a. El nivel medio de la cotización de AMZN es mayor al nivel medio de la cotización de GOOG.
Verdadero porque el nivel medio de la cotización de AMZN es 126,4; mientras que el nivel medio de la cotización de GOOG es 122,5.

b. El nivel de cotización mínima de la mitad de las cotizaciones de AMZN es mayor que el nivel de cotización mínima de la mitad de las cotizaciones de GOOG.
Verdadero porque el nivel de cotización mínima de la mitad (P50) es de 126,6 para AMZN y de 123.1 para GOOG

c. El nivel de cotización máxima del 25 % de GOOG que tienen menores cotizaciones es mayor al nivel de cotización máxima del 25 % de AMZN que tienen menores cotizaciones.
Falso porque el nivel cotización máxima del 25% (P25) de GOOG que tiene menores cotizaciones es 120.6; mientras que, en el caso de AMZN, este indicador es 124.8

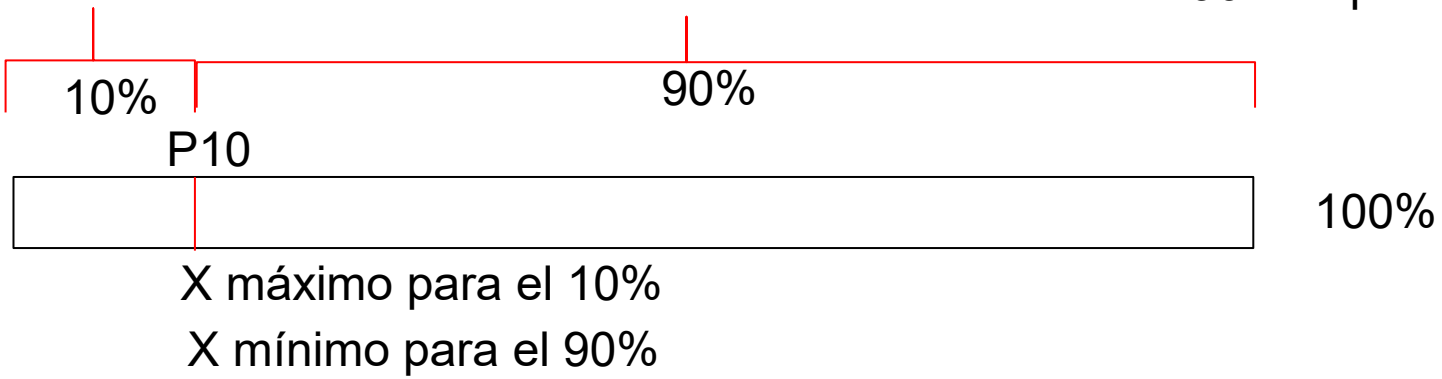
d. En las cotizaciones de AMZN se observa una mayor variabilidad.
Verdadero porque el desvío (2,45) de AMZN es mayor que en el caso GOOG (2,43).

Descripción

Analizar si las siguientes afirmaciones son verdaderas o falsas. Justificar

e. El nivel de cotizaciones **mínimas** del 25 % superior de las cotizaciones de GOOG es casi similar al nivel de cotizaciones mínimo del 90 % superior de las cotizaciones de AMZN.
Verdadero porque el valor del percentil 10 (123,4) para AMZN es similar al valor del percentil 75 (123,7) de GOOG

P10 me da el valor máximo del 10% o el valor mínimo del 90% superior

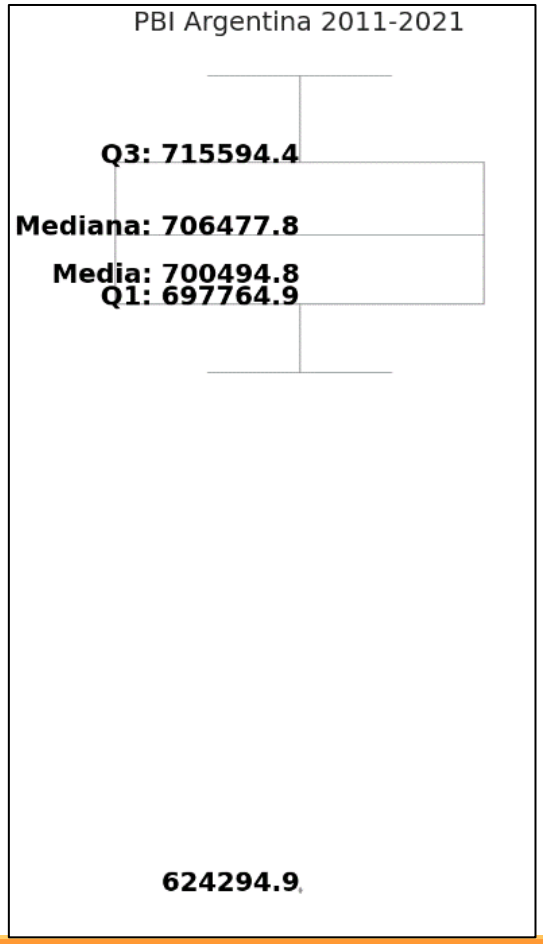


f. No existen valores atípicos (ni moderados ni severos) en ambas series.

Verdadero porque en los boxplots no se observan ningún valor atípico.

Descripción

EJEMPLO: Observando el siguiente boxplot del PBI anual de Argentina entre 2011-2021 en millones de pesos. Indicar si las siguientes afirmaciones son verdaderas o falsas **justificando** adecuadamente su respuesta.



A) “Puede afirmarse que existe una alta variabilidad en la distribución del PBI de Argentina”

Rta: FALSO, justificar

B) “El valor atípico que se observa es moderado”

Rta: FALSO, justificar

C) “La distribución del PBI es asimétrica positiva”

Rta: FALSO, justificar

D) “Sabiendo que el coeficiente de curtosis es igual a 3,42, es posible afirmar que la distribución es leptocúrtica”

Rta: VERDADERO, justificar

Pueden realizar la práctica 5