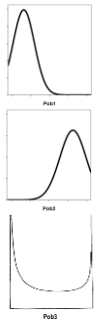
Las distribuciones de probabilidad en el análisis de los datos

1

Ejercicio:
Distribuciones

Las gráficas que se presentan a continuación muestran la distribución de probabilidad de los salarios de 3 universidades de un país AAAAA



2

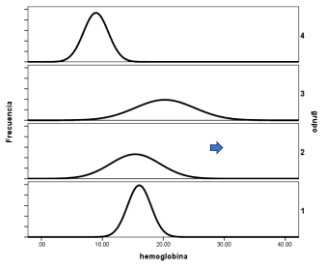
Problema: En un ensayo clínico con cuatro ramas se comparan tres productos que se espera normalicen los niveles de hemoglobina a los hombres anémicos

Valores normales:
14 a 18 g/100ml

3

Problema: En un ensayo clínico con cuatro ramas se comparan tres productos que se espera normalicen los niveles de hemoglobina a los hombres anémicos

Valores normales:
14 a 18 g/100mlc



¿Qué opina usted de los resultados?

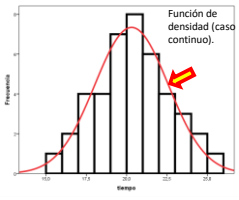
4

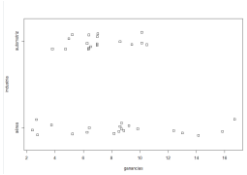
Representación gráfica de la distribución de las variables continuas

Histograma y diagramas de puntos

5

Representación de las distribuciones en el caso continuo: Histogramas y diagramas de puntos





6

7

Problema

Se sospecha que el si las adolescentes practican un ejercicio físico de alta intensidad esto puede afectar la edad de la menarquia (primera menstruación). Para analizar esta preocupación se realizó un estudio en el que se analizó la edad de la menarquia a adolescentes que practicaron el deporte antes o después de su primera menstruación

Represente la distribución (utilice el histograma y el diagrama de puntos)
Comente sobre el comportamiento de las edades y analice si se observa o no efecto del deporte sobre la edad

Base de datos menarquia

8

Una agencia de empleos hizo evaluaciones para caracterizar las habilidades para la captura de datos de cuatro personas. El número de datos digitados correctamente durante un minuto por cada una de las personas es:

Persona	datos					
A	63	66	68	64	69	72
B	68	67	66	67	68	
C	50	79	75	59	72	20
D	64	68	50	57	59	

Represente gráficamente los valores de cada persona y utilice el gráfico para describir su desempeño

9

Medidas para representar aspectos de la distribución de probabilidad

10

Medidas de posición

Posición central

Posición no central

11

Medidas de posición: No Centrales.
Los percentiles

105	135	180	210	245	250	265	380	395	420	430	615	690	830	1180

P25: 210 P50 : 322.5 P75: 615

25% 50% 75%

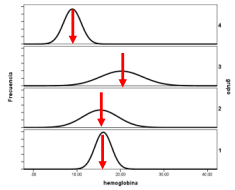
Percentiles, Cuartiles, Deciles ….

12

Los diagramas de cajas

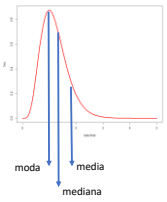
13

Medidas de posición



14

Las medidas de posición central



$X_1\ X_2\ \ldots\ldots\ldots\ X_n$

$$X = \frac{1}{n}\sum_i X_i$$

$X_{(1)}\ X_{(2)}\ \ldots\ldots\ldots\ X_{(n)}$

$\tilde{X}$

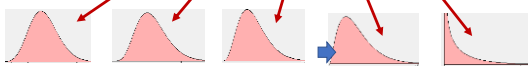
n par

n impar

15

Medidas de posición: ¿Media o mediana?

casos	23	23	23	23	23
	24	24	24	24	24
	22	22	22	22	22
	25	25	25	25	25
	28	28	28	28	28
	29	29	29	29	29
	31	31	31	31	31
	30	35	55	785	7882
promedio	26.5	27.1	29.8	120.9	1008.0



Cuando las distribuciones son muy asimétricas la media aritmética no identifica el puntaje "típico" por eso se debe usar otra medida de posición: LA MEDIANA

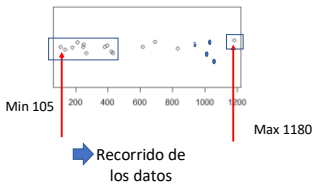
16

Medidas de dispersión (de escala)

17

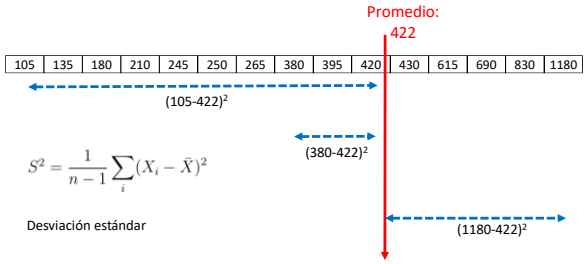
Medidas de dispersión:

Cuantifican homogeneidad / heterogeneidad



18

Medidas de dispersión: Varianza y Desviación estándar



$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$$

Desviación estándar

19

¿Por qué no puede ser de una forma más "racional"?

Lo lógico sería definir la dispersión respecto a la media como el promedio de las desviaciones, esto es:

$$\frac{1}{n}\sum_i (X_i - \bar{X})$$

¡Pero esto no funciona!,

Entonces para seguir esa idea debe ser:

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$$

¿y qué le parece esta?

$$\frac{1}{n}\sum_i |X_i - X|$$

¿Por qué complicarse la vida y no usar mejor?:

$$\sum_i (X_i - \bar{X})^2 \quad \text{o} \quad \sum_i |X_i - \bar{X}|$$

20

La desviación estándar para caracterizar variabilidad

datos	105	110	125	200
	100	100	100	100
	95	90	75	0
media	**100**	**100**	**100**	**100**
DE	5	10	25	100

21

Uso de la varianza. Descripción de la alimentación de dos personas

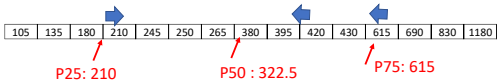
persona 1		persona 2
energía		energía
1759.8		887.6
3095.5		648.86
1612.4		614
1557.1		905.2
1236.96		883.7
1727.15		768.2
1448.35		629.56
1086.83		1038.4
1338.75		905.7
1011.41		810.2
1182.75		709.8
1294.16		1071.5
2019.06		636.4
2534.77		834.8
1072.45		873.5
1878.09		844.1
1069.15		800.3
1853.3		788.9
723.7		860.5
1684.01		746.5
1430.25		862.7
1167.68		767.5
1348.63		1244.75
899.35		618.1
1049.95		808.2
1032.88		916.28
1746.38		531.8
2183.72		758.7

característica	persona 1		persona 2
media	1500.2		811.3
DE	524.1		150.3

Recomienda para una persona ICBF: 2100Kcal/dia

22

Medidas de dispersión: Recorrido intercuartílico



Recorrido intercuartílico:
P75 – P25 = 615 - 210

23

Medidas de dispersión: Mediana de la desviación absoluta (MAD)

X_1 X_2 ..... X_n

$\tilde{X}$

Nueva variable

$Y_i = |X_i - \tilde{X}|, i = 1, \ldots n$

MAD (X) = $\tilde{Y}$

24

Problema: Tenemos 4 medidas de dispersión (escala). Comente las características de cada una y reflexiones sobre cuándo utilizaría o no utilizaría cada una

Recorrido = $X_{(n)} - X_{(1)}$

Recorrido intercuartílico: P75 – P25

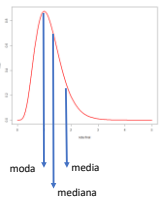
Varianza $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$

Mediana de la desviación absoluta: mediana (|X-$\tilde{X}$|)

25

Las medidas de forma

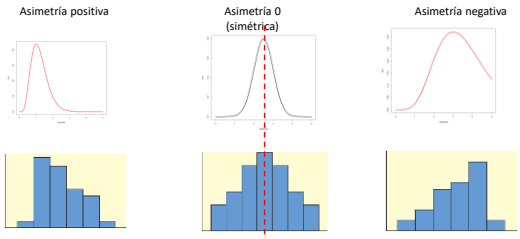
26



Medidas de forma: El coeficiente de asimetría (1er coeficiente de Pearson)

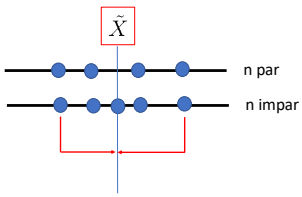
$$\frac{media - moda}{DE}$$

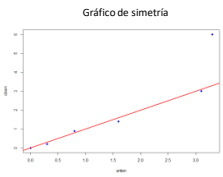
27



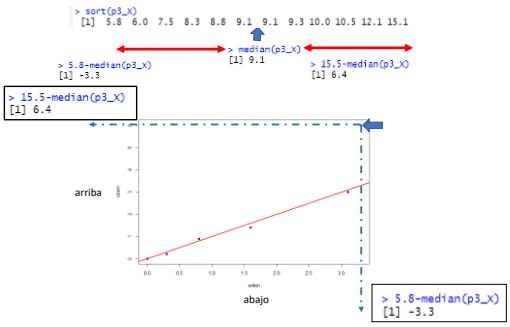
$$\frac{media - moda}{DE}$$

28





29



30

31