

Tema 1

Estadística descriptiva:

Medidas de centralización y
dispersión

Curso 2017/18
Grados en biología sanitaria
Departamento de Física y Matemáticas
Marcos Marvá Ruiz



A partir de los valores de una **variable estadística**, vamos a considerar tres tipos de medidas (“parámetros”):

1. Medidas de **centralización** valor “resumen” que describa “qué podemos esperar”
2. Medidas de **dispersión** miden la representatividad del valor central anterior
3. Medidas de **posición** indica si un valor es “alto” o “bajo” comparado con el resto

Permiten intuir la distribución de los datos (forma del diagrama de barras o histograma)

MEDIDAS DE CENTRALIZACIÓN

Media aritmética: variables cuantitativas

Datos no agrupados: x_1, x_2, \dots, x_n

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Ejemplo: dados los valores: $X = 1, 4, 16, 11, 3, 6$, su media es

MEDIDAS DE CENTRALIZACIÓN

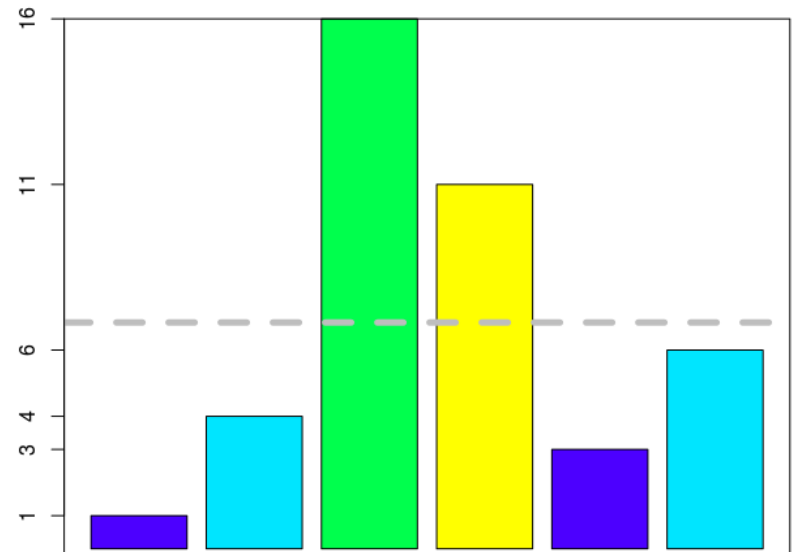
Media aritmética: variables cuantitativas

Datos no agrupados: x_1, x_2, \dots, x_n

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Ejemplo: dados los valores: $X = 1, 4, 16, 11, 3, 6$, su media es

$$\bar{X} = \frac{1+3+4+11+16}{5} = \frac{35}{5} = 7$$



Cada barra representa un valor de X, la línea gris es la media

MEDIDAS DE CENTRALIZACIÓN

Media aritmética: variables cuantitativas

Ejemplo: si tus calificaciones son 8, 9, 9, 9, 10, 10, la nota media es

$$\bar{x} = \frac{8+9+9+9+10+10}{6} = \frac{8 \cdot 1 + 3 \cdot 9 + 2 \cdot 10}{1+3+2} = \frac{55}{6} = 9.1666666667$$

Aprenderemos a manejar lo decimales

Datos no agrupados x_1, x_2, \dots, x_k con frecuencias absolutas f_1, f_2, \dots, f_k y relativas f'_1, \dots, f'_k

$$\bar{x} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \dots + f_k \cdot x_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k f_i \cdot x_i}{\sum_{i=1}^k f_i} = \sum_{i=1}^k f'_i \cdot x_i$$

MEDIDAS DE CENTRALIZACIÓN


Media aritmética: variables cuantitativas

Datos ya agrupados en clases

$$\begin{array}{ccccccc} [a_0, a_1] & (a_1, a_2] & (a_2, a_3] & \dots & (a_{k-1}, a_k] \\ f_1 & f_2 & f_3 & \dots & f_k \end{array}$$

La **marca de clase** (pto medio) $x_i := (a_i + a_{i+1})/2$ representa a los elementos de la clase.

Así, tenemos $x_1, x_2, x_3, \dots, x_k$ con frecuencias $f_1, f_2, f_3, \dots, f_k$ y podemos aplicar


$$\bar{X} = \frac{\sum_{i=1}^k f_i \cdot x_i}{\sum_{i=1}^k f_i}$$

MEDIDAS DE CENTRALIZACIÓN


Media aritmética: variables cuantitativas

Datos agrupados en clases

$$\begin{array}{ccccccc} [a_0, a_1] & (a_1, a_2] & (a_2, a_3] & \dots & (a_{k-1}, a_k] \\ f_1 & f_2 & f_3 & \dots & f_k \end{array}$$

La **marca de clase** (pto medio) $x_i := (a_i + a_{i+1})/2$ representa a los elementos de la clase.

Así, tenemos $x_1, x_2, x_3, \dots, x_k$ con frecuencias $f_1, f_2, f_3, \dots, f_k$ y podemos aplicar


$$\bar{X} = \frac{\sum_{i=1}^k f_i \cdot x_i}{\sum_{i=1}^k f_i}$$

Ejemplo: creatinina en sangre (mg/dl)

[1.52,1.58]	(1.58,1.64]	(1.64,1.7]	(1.7,1.76]	(1.76,1.82]	(1.82,1.88]
12	16	30	28	11	2

Marcas de clase: $x_1 = 1.55$, $x_2 = 1.61$, $x_3 = 1.67$, $x_4 = 1.73$, $x_5 = 1.79$, $x_6 = 1.85$

Frecuencia: $f_1 = 12$, $f_2 = 16$, $f_3 = 30$, $f_4 = 28$, $f_5 = 11$, $f_6 = 2$

MEDIDAS DE CENTRALIZACIÓN

Media aritmética: variables cuantitativas: algunas consideraciones

Ejemplo: supón que tus calificaciones son 8, 9, 9, 9, 10,10

$$\bar{x} = \frac{8+9+9+9+10+10}{6} = \frac{8 \cdot 1 + 3 \cdot 9 + 2 \cdot 10}{1+3+2} = \frac{55}{6} = 9.1666666667$$

* ¿Y si hubiera “pinchado” en un examen?

$$\bar{x} = \frac{1+9+9+9+10+10}{6} = \frac{48}{6} = 8$$

La media es sensible
a valores extremos

MEDIDAS DE CENTRALIZACIÓN

Media aritmética: variables cuantitativas: algunas consideraciones

Ejemplo: supón que tus calificaciones son 8, 9, 9, 9, 10,10

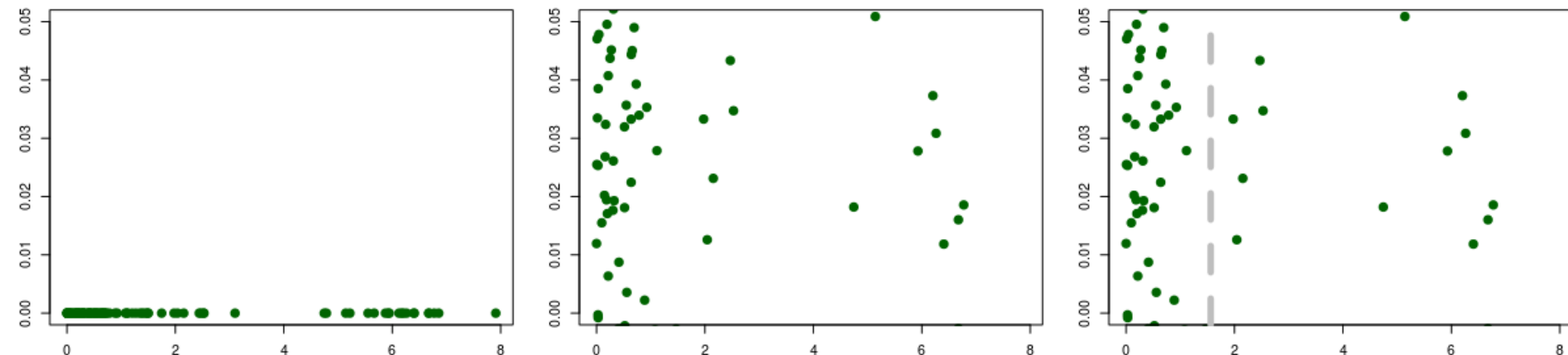
$$\bar{x} = \frac{8+9+9+9+10+10}{6} = \frac{8 \cdot 1 + 3 \cdot 9 + 2 \cdot 10}{1+3+2} = \frac{55}{6} = 9.1666666667$$

* ¿Y si hubiera “pinchado” en un examen?

$$\bar{x} = \frac{1+9+9+9+10+10}{6} = \frac{48}{6} = 8$$

La media es sensible
a valores extremos

Ejemplo: La media no siempre es “representativa”



MEDIDAS DE CENTRALIZACIÓN

Mediana: valor de la variable que, una vez ordenados de menor a mayor, deja la mitad de los datos por debajo de sí:

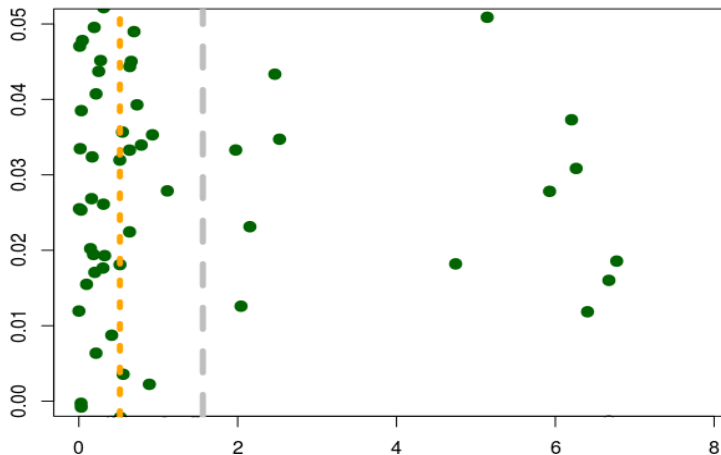
- * Si hay una cantidad **impar** de datos, se toma el valor del centro
- * Si hay una cantidad **par** de datos, se toma la media entre los dos centrales.

Es *robusta* frente a (unos pocos) valores extremos.

Ejemplos:

1 ¿Cuál es la mediana en los casos {8, 9, 9, 9, 10, 10} y {1, 9, 9, 9, 10, 10}?

2



Línea a puntos (izquierda) la mediana
Línea a guiones (derecha) la media

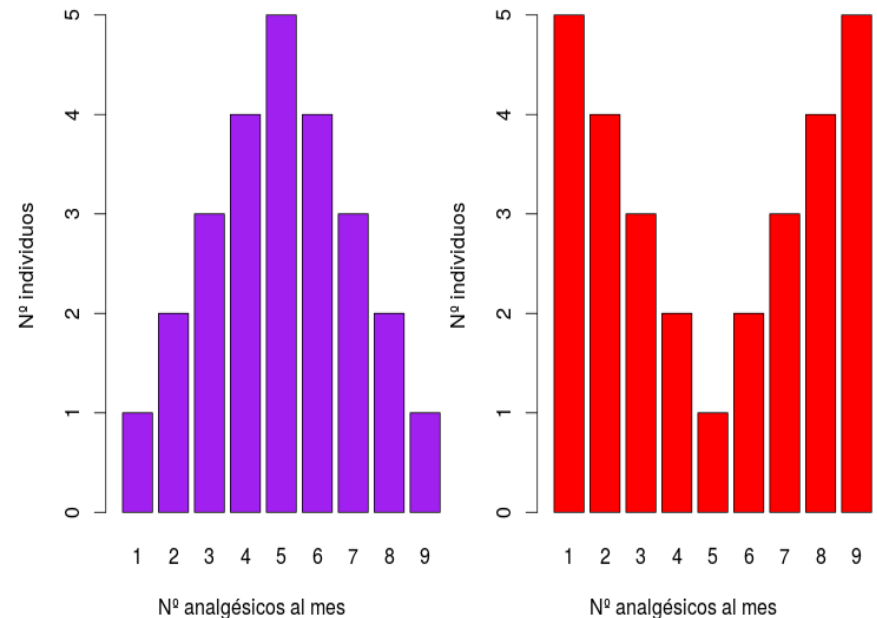
MEDIDAS DE CENTRALIZACIÓN

Moda

Valor de la variable estadística (o de la clase) con frecuencia más alta.
Hay muestras unimodales y multimodales (bimodal, trimodal,...)

Ejemplo: dos muestras de 29 individuos. Se pregunta por el nº analgésicos que Ingieren al mes.

Por cierto: ¡Los dos conjuntos de datos tienen la misma media y mediana: 5!
¿Qué diferencia ambas situaciones?



MEDIDAS DE CENTRALIZACIÓN

Moda

Distribuciones multimodales: posible(s) variable(s) oculta(s)

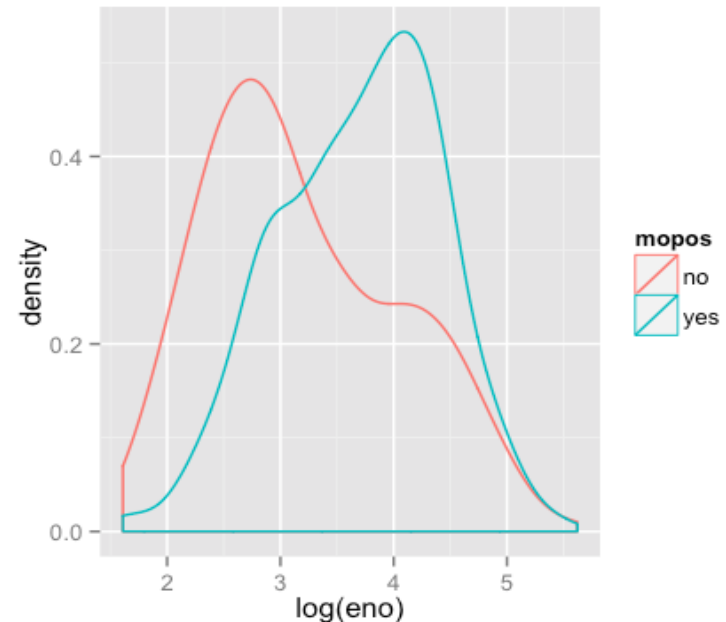
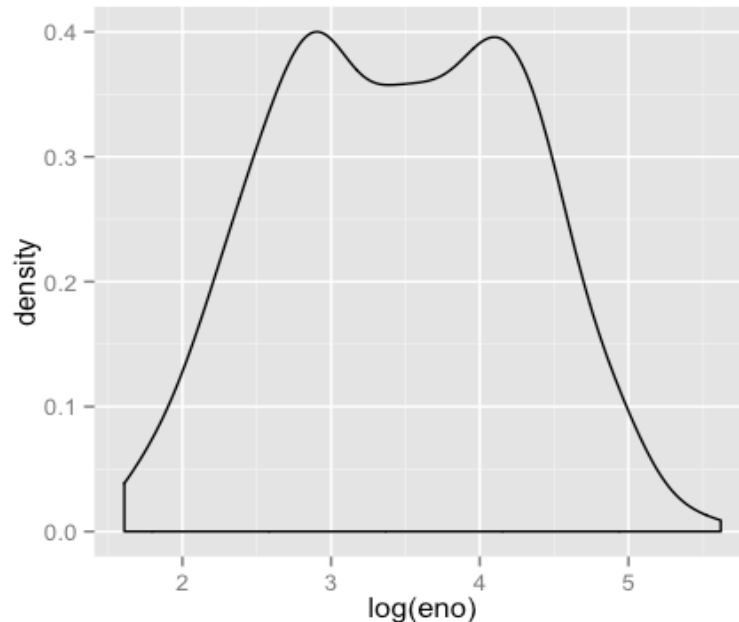
Mouse Allergen and Asthma Cohort Study.

Publicacion: <http://bit.ly/2cuYVyw>

Se representa el log(ENO) frente a la tproporción de episodios nocturnos de asma

ENO: oxido nítrico exhalado. El oxido nítrico (NO) es una molécula gaseosa producido por cierto tipo de células como respuesta a un proceso inflamatorio.

Mopo: individuos sensibilizados a los alérgenos del ratón



MEDIDAS DE DISPERSIÓN

Miden la cercanía de los datos a la media, de forma global. En particular, permiten evaluar la representatividad de la media.

Dispersión alta → datos alejados de la media → datos heterogéneos → media poco representativa.

Dispersión baja → datos próximos a la media → datos homogéneos → media muy representativa.

MEDIDAS DE DISPERSIÓN

Para variables cuantitativas.

Miden lo agrupados que están los datos en torno a una medida de centralización o su grado de desagregación

Recorrido (o rango) de una variable:

Resta entre los valores máximo y mínimo de la variable

Ejemplo: valores 6, **13**, 5, 8, **2**, 4 → recorrido : $13 - 2 = 11$

Varianza poblacional: Datos no agrupados: x_1, x_2, \dots, x_n con media \bar{x}

$$Var(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

MEDIDAS DE DISPERSIÓN

Para variables cuantitativas.

Miden lo agrupados que están los datos en torno a una medida de centralización o su grado de desagregación

Recorrido (o rango) de una variable:

Resta entre los valores máximo y mínimo de la variable

Ejemplo: valores 6, **13**, 5, 8, **2**, 4 → recorrido : $13 - 2 = 11$

Varianza poblacional: Datos no agrupados: x_1, x_2, \dots, x_n con media \bar{x}

$$Var(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Ejemplo: *Para el conjunto de valores*

$$9, 6, 19, 10, 17, 3, 28, 19, 3, 5, 19, 2, \quad \bar{x} = \frac{140}{12}$$

$$Var(x) = \frac{(9 - \frac{140}{12})^2 + (6 - \frac{140}{12})^2 + \dots + (19 - \frac{140}{12})^2 + (2 - \frac{140}{12})^2}{12} =$$

$$= \frac{\frac{2360}{3}}{12} = \frac{2360}{36} \approx 65.56$$

MEDIDAS DE DISPERSIÓN

Varianza poblacional datos agrupados

Ejemplo: cálculo de la varianza a partir de la tabla de frecuencias absolutas (la media es 43)

x _i	f _i
40	2
42	1
45	3

$$Var(X) = \frac{(40-43)^2 + (40-43)^2 + (42-43)^2 + (45-43)^2 + (45-43)^2 + (45-43)^2}{6}$$

$$Var(X) = \frac{(40-43)^2 \cdot 2 + (42-43)^2 \cdot 1 + (45-43)^2 \cdot 3}{2+1+3} = 7.4$$

Desviación típica poblacional

$$DT(X) = \sqrt{Var(X)}$$

$$Var(X) = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^k f_i}$$

MEDIDAS DE DISPERSIÓN

Varianza muestral o cuasivarianza

Datos sin agrupar:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Datos agrupados:

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^k f_i - 1}$$

Desviación típica muestral o cuasidesviación típica: $s = \sqrt{s^2}$

Muchos libros hablan de la cuasivarianza incluso sin definir la varianza. La cuasivarianza aparecerá en el bloque de inferencia. Si usas software o una función de la calculadora, es **importante** que sepas si el número que se obtienes es la varianza o la cuasivarianza muestra

MEDIDAS DE DISPERSIÓN

Varianza muestral o cuasivarianza

Ejemplo: Cálculo de varianza y cuasivarianza

Para el conjunto de valores

$$9, 6, 19, 10, 17, 3, 28, 19, 3, 5, 19, 2, \quad \bar{x} = \frac{140}{12}$$

$$\begin{aligned} Var(x) &= \frac{(9 - \frac{140}{12})^2 + (6 - \frac{140}{12})^2 + \dots + (19 - \frac{140}{12})^2 + (2 - \frac{140}{12})^2}{12} = \\ &= \frac{\frac{2360}{3}}{12} = \frac{2360}{36} \approx 65.56 \end{aligned}$$

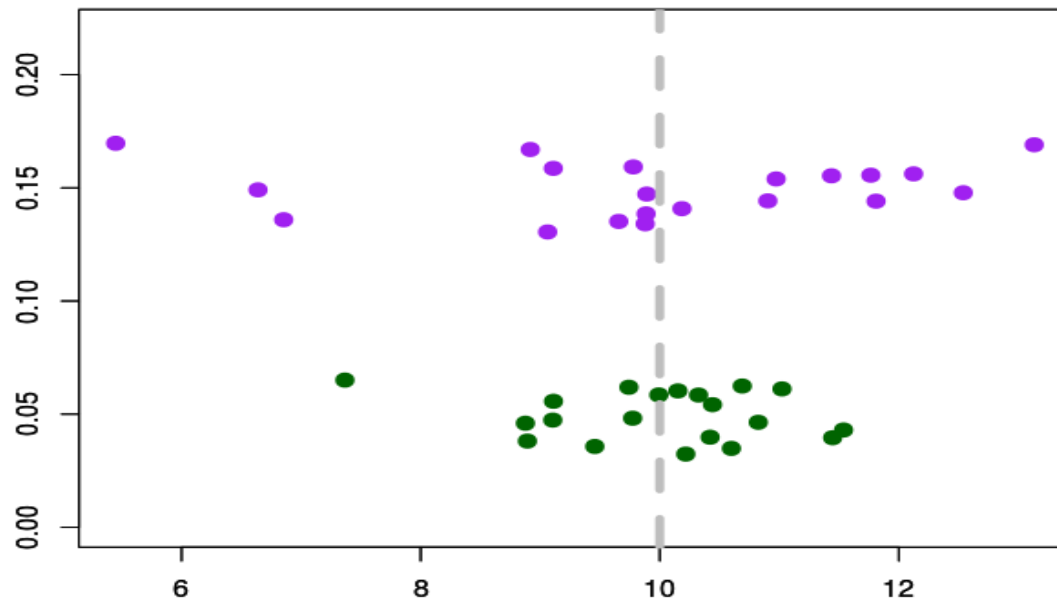
$$\begin{aligned} s^2 &= \frac{(9 - \frac{140}{12})^2 + (6 - \frac{140}{12})^2 + \dots + (19 - \frac{140}{12})^2 + (2 - \frac{140}{12})^2}{11} = \\ &= \frac{\frac{2360}{3}}{11} = \frac{2360}{33} \approx 71.52, \end{aligned}$$

MEDIDAS DE DISPERSIÓN

Varianza muestral o cuasivarianza

Propiedades de la (cuasi)varianza y la (cuasi)desviación típica:

- 1.- La varianza no puede ser negativa.
- 2.- A igualdad de medias, mayor dispersión implica mayor varianza.
- 3.- De dos muestras con medias similares, es más dispersa la que tenga mayor varianza.



Ambas muestras tienen media 10

Muestra morada (arriba) tiene desviación típica = 4

Muestra verde (abajo) tiene desviación típica = 1

MEDIDAS DE DISPERSIÓN

Varianza muestral o cuasivarianza

Propiedades de la (cuasi)varianza y la (cuasi)desviación típica:

4.- PERO, si dos muestras tienen medias diferentes, mayor varianza NO implica mayor dispersión **la varianza depende del tamaño (unidades) de los datos.**

Adimensionalizar



Coeficiente de variación (CV)

$$CV = \frac{s_X}{\bar{x}}$$

A mayor CV, mayor dispersión, y viceversa.



También útil para comparar variables diferentes

Presenta problemas cuando la media es próxima a cero

MEDIDAS DE DISPERSIÓN

Coeficiente de variación

Ejemplo: En el experimento de Framinham quiere saber si en la primera medición las variables sysbp, diabp y bmi presentan dispersiones similares

sysbp

Media = 132.89

s = 22.36

CV = 0.1683

diabp

Media = 83.09

S = 12.05

CV = 0.1450

bmi

Media = 25.85

S = 4.10

CV = 0.1587