M2 - Análisis de Datos 2: descriptivo e inferencial

Dora Suárez, Juan F. Pérez

Departamento MACC
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario

juanferna.perez@urosario.edu.co

Primer Semestre de 2019

Contenidos

Media, varianza y desviación estándar

Análisis descriptivo de datos: valor esperado

- X: característica de la población (variable aleatoria)

Análisis descriptivo de datos: valor esperado

- X: característica de la población (variable aleatoria)
- **Valor esperado** de X:

$$E[X] = \sum_j j \times p_j$$

Análisis descriptivo de datos: valor esperado

- X: característica de la población (variable aleatoria)
- **Valor esperado** de X:

$$E[X] = \sum_j j \times p_j$$

  - Promedio ponderado de los valores que toma X

Análisis descriptivo de datos: valor esperado

- X: característica de la población (variable aleatoria)
- **Valor esperado** de X:

$$E[X] = \sum_j j \times p_j$$

  - Promedio ponderado de los valores que toma X
  - Medida de localización de X

Análisis descriptivo de datos: varianza

- X: característica de la población (variable aleatoria)

Análisis descriptivo de datos: varianza

- X: característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

Análisis descriptivo de datos: varianza

- X: característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

  - Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado

Análisis descriptivo de datos: varianza

- X: característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

  - Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado
  - Medida de la variabilidad de X respecto a $E[X]$

Análisis descriptivo de datos: varianza

- X: característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

  - Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado
  - Medida de la variabilidad de X respecto a $E[X]$
  - Desviación estándar:

$$\sigma_X = \sqrt{V[X]}$$

Análisis descriptivo de datos (ejemplo)

- $$P(X=x)=\begin{cases}1/3, & x=1,\\ 1/3, & x=2,\\ 1/3, & x=3.\end{cases}$$
- $$P(Y=y)=\begin{cases}1/4, & y=1,\\ 1/2, & y=2,\\ 1/4, & y=3.\end{cases}$$
- $$P(Z=z)=\begin{cases}1, & z=2.\end{cases}$$
- $$P(W=x)=\begin{cases}1/2, & x=1,\\ 1/2, & x=3.\end{cases}$$

Análisis descriptivo de datos (ejemplo)

- $E[X] = E[Y] = E[Z] = E[W] = 2$
- $V(X) = (1-2)^2(1/3) + (2-2)^2(1/3) + (3-2)^2(1/3) = 2/3$
- $\sigma_X = \sqrt{2/3}$
- $V(Y) = (1-2)^2(1/4) + (2-2)^2(1/2) + (3-2)^2(1/4) = 1/2$
- $\sigma_Y = \sqrt{1/2}$
- $V(Z) = (2-2)^2(1) = 0$
- $\sigma_Z = \sqrt{0} = 0$
- $V(W) = (1-2)^2(1/2) + (3-2)^2(1/2) = 1$
- $\sigma_W = \sqrt{1} = 1$

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Media muestral**: $\bar{X}$

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

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Media muestral**: $\bar{X}$

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

- Promedio de los datos de la muestra

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Media muestral**: $\bar{X}$

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

- Promedio de los datos de la muestra
- Mean, media, promedio

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Varianza muestral**: S^2

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

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Varianza muestral**: S^2

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

- Promedio de la diferencia (al cuadrado) entre los datos de la muestra y el promedio muestral

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \ldots, X_n\}$
- **Varianza muestral**: S^2

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

- Promedio de la diferencia (al cuadrado) entre los datos de la muestra y el promedio muestral
- **Desviación estándar muestral** (std): $S = \sqrt{S^2}$

Cuantiles, percentiles, cuartiles y mediana

Cuantiles

- X: característica de la población (variable aleatoria)

Cuantiles

- X: característica de la población (variable aleatoria)
- p: número entre 0 y 1

Cuantiles

- X: característica de la población (variable aleatoria)
- p: número entre 0 y 1
- El **cuantil** p es el número c más pequeño tal que la probabilidad de que X sea menor que c es al menos p, i.e.,:

$$P(X \leq c) = \geq p$$

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & dlc. \end{cases}$$

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & dlc. \end{cases}$$

- El cuantil 0.25 es 1

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & dlc. \end{cases}$$

- El cuantil 0.25 es 1
- El cuantil 0.5 es 2.

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & dlc. \end{cases}$$

- El cuantil 0.25 es 1
- El cuantil 0.5 es 2.
- Los cuantiles 0.3 y 0.75 también son 2.

Cuantiles especiales

- **Percentiles**: el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99

Cuantiles especiales

- **Percentiles**: el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75

Cuantiles especiales

- **Percentiles**: el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75
- La **mediana** es el percentil 50 (cuantil 0.5)

Cuantiles especiales

- **Percentiles**: el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75
- La **mediana** es el percentil 50 (cuantil 0.5)
- Los tres **cuartiles** son los percentile 25, 50 y 75

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \ldots, X_n\}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \ldots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \ldots, X_{(n)}\}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \ldots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \ldots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \ldots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \ldots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$
- Justo en $X_{(k)}$ se han observado k muestras menores o iguales a $X_{(k)}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \ldots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \ldots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$
- Justo en $X_{(k)}$ se han observado k muestras menores o iguales a $X_{(k)}$
- A partir de $X_{(n)}$ se han observado todas las n muestras

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $c \times n$ muestras menores o iguales.

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $c \times n$ muestras menores o iguales.
- Cuantil c es la muestra $c \times n$-ésima en la lista ordenada

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $c \times n$ muestras menores o iguales.
- Cuantil c es la muestra $c \times n$-ésima en la lista ordenada
- Ejemplo: si tenemos una muestra de tamaño $n = 1000$, el cuantil 0,2 es la muestra 200 de la lista ordenada

Cuantiles muestrales (ejemplo)

```
data(mtcars)
summary(mtcars)
summary(mtcars$mpg)

quantile(mtcars$mpg)
quantile(mtcars$mpg, 0.5)
quantile(mtcars$mpg, c(0.5, 0.9))

median(mtcars$mpg)
x <- sort(mtcars$mpg)
x[16:17]
hist(mtcars$mpg)
```