Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Unidad 1: Introducción

Andrés Peña M.

a.pena@rusersgroup.com

Julio 2021

X Seminario Internacional de Estadística Aplicada

Tabla de contenidos

1. Estructura del taller
2. Introducción
   - Visión general del método estadístico
3. Técnicas de conteo
4. Estadística Descriptiva

1. Estructura del taller

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.
2. Data management*

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.
2. Data management*
3. Probabilidad y variables aleatorias. Función de densidad y distribución. Principales Distribuciones Continuas y Discretas. Distribuciones asociadas a la normal.

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.
2. Data management*
3. Probabilidad y variables aleatorias. Función de densidad y distribución. Principales Distribuciones Continuas y Discretas. Distribuciones asociadas a la normal.
4. Distribuciones en el muestreo. Teorema del Límite Central.

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.
2. Data management*
3. Probabilidad y variables aleatorias. Función de densidad y distribución. Principales Distribuciones Continuas y Discretas. Distribuciones asociadas a la normal.
4. Distribuciones en el muestreo. Teorema del Límite Central.
5. Inferencia estadística. Estimación puntual: propiedades de los estimadores y principales métodos de estimación. Estimación por intervalos. Pruebas de hipótesis.

Estructura del taller

1. Visión general del método. Repaso de técnicas de conteo. Análisis descriptivo. Medidas de tendencia central, dispersión, posición y forma.
2. Data management*
3. Probabilidad y variables aleatorias. Función de densidad y distribución. Principales Distribuciones Continuas y Discretas. Distribuciones asociadas a la normal.
4. Distribuciones en el muestreo. Teorema del Límite Central.
5. Inferencia estadística. Estimación puntual: propiedades de los estimadores y principales métodos de estimación. Estimación por intervalos. Pruebas de hipótesis.
6. Análisis de encuestas por muestreo. Técnicas de muestreo. Factores de expansión. Exploración de variables categoricas y numéricas. Modelamiento.

2. Introducción

Introducción

- "The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language" Galileo Galilei.

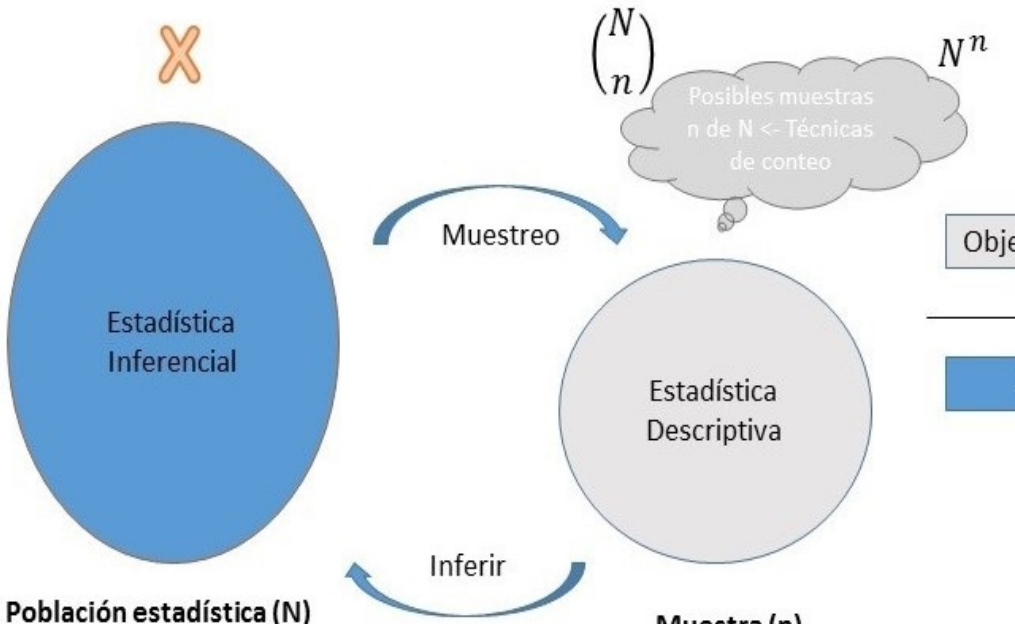
Introducción

- "The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language" Galileo Galilei.
- "**Statistics** is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, **better scientific investigation**. Whether in any given study this implies more or less mathematics is incidental" George E. P. Box.

Introducción

- "The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language" Galileo Galilei.
- "**Statistics** is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, **better scientific investigation**. Whether in any given study this implies more or less mathematics is incidental" George E. P. Box.
- El carácter de las Matemáticas es fundamentalmente **deductivo**. Por su parte, la Inferencia Estadística hace uso del conocimiento matemático pero tiene una naturaleza **inductiva**. El vínculo entre estas dos áreas lo provee la **Probabilidad**.

Visión general del método estadístico



Objetos matemáticos — Variable aleatoria, Modelo probabilístico

Real — Variables experimentales, Característica de estudio

Teoría de Muestreo

- Estrategias de muestreo

Teoría de Probabilidad

- Fábrica de construcción de modelos.
- Distribuciones como sustituto de las poblaciones.
- Convergencia estocástica
  - Ley de los grandes números.
  - Teorema Central de Límite, desigualdad de Tchevysheff.

SOCIEDAD ECUATORIANA ESTADÍSTICA

3. Técnicas de conteo (repaso)

Variaciones con repetición

```
#Permutaciones y combinaciones
install.packages("gtools", dependencies = TRUE)
```

Usado en muestreo con reemplazo:

$$VR_n^N = N^n$$

```
(x<-1:4)

## [1] 1 2 3 4

4^2 #Duplas posibles con repetición de una población de 4

## [1] 16
```

Variaciones con repetición

```
library(gtools)
permutations(n=4,r=2,v=x,repeats.allowed=T)
```

```
##       [,1] [,2]
##  [1,]    1    1
##  [2,]    1    2
##  [3,]    1    3
##  [4,]    1    4
##  [5,]    2    1
##  [6,]    2    2
##  [7,]    2    3
##  [8,]    2    4
##  [9,]    3    1
## [10,]    3    2
## [11,]    3    3
## [12,]    3    4
## [13,]    4    1
## [14,]    4    2
## [15,]    4    3
```

Combinaciones sin repetición

Usado en muestreo sin reemplazo:

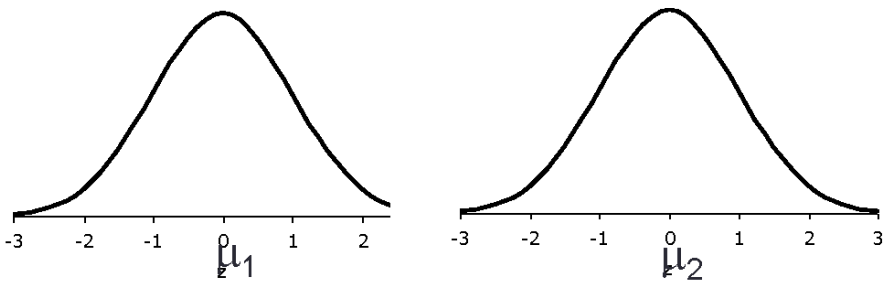
$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

```
choose(4,2) #Duplas posibles sin repetición

## [1] 6

combinations(4, 2, v=x)

##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    2    3
## [5,]    2    4
## [6,]    3    4
```

4. Estadística Descriptiva

Medidas de resumen de la información

Medidas
- Posición
- Dispersión
- Forma



Medidas de posición de tendencia central

Media aritmética

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

- Todo conjunto de escala medible tiene una media
- Un conjunto de datos solo tiene una media
- En su cálculo, se incluyen todos los valores de la variable, por lo cual es sensible a los valores extremos y puede No ser representativa

Medidas de posición de tendencia central

Mediana

Es el valor de la variable que corresponde al lugar (n+1)/2

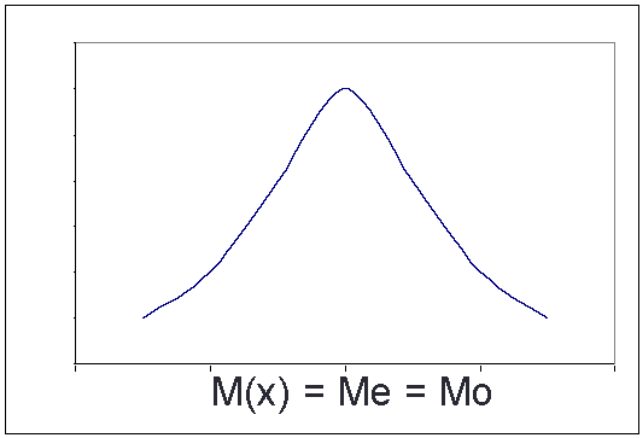
- No se ve afectada por observaciones extremas.

Modo

Es el valor de la variable más frecuente

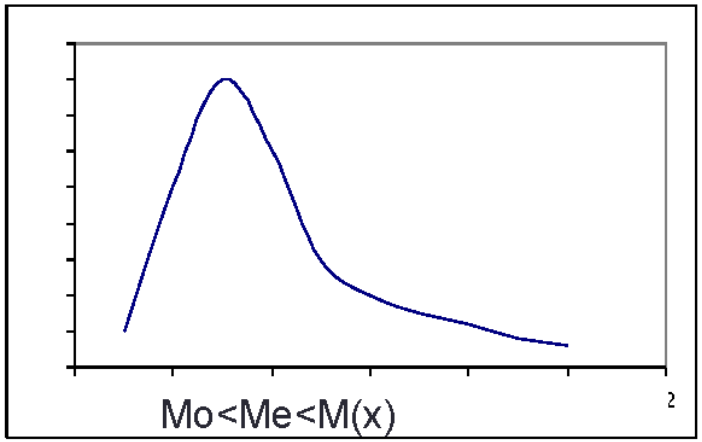
- Algunos conjuntos de datos no poseen modo, otros tienen dos o más.
- Se obtiene fácilmente a partir de un conjunto de datos.
- Suele ser la única medida de obtener en un conjunto de datos categóricos.

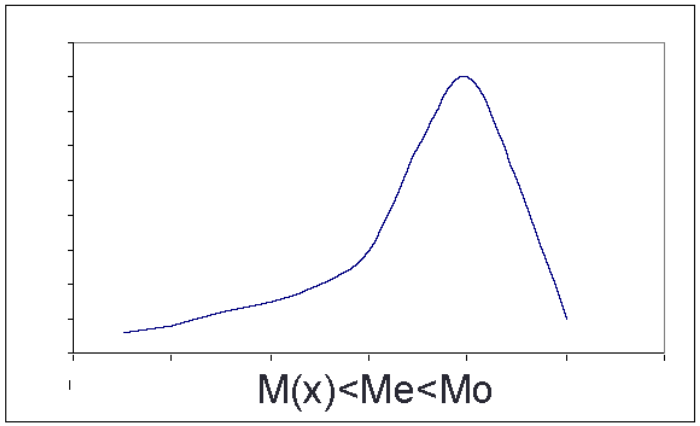
Relación entre media, mediana y moda



Distribución Simétrica

Distribución Asimétrica





Medidas de posición de tendencia no central

Cuartiles

Son medidas descriptivas que dividen los datos ordenados en cuartos.

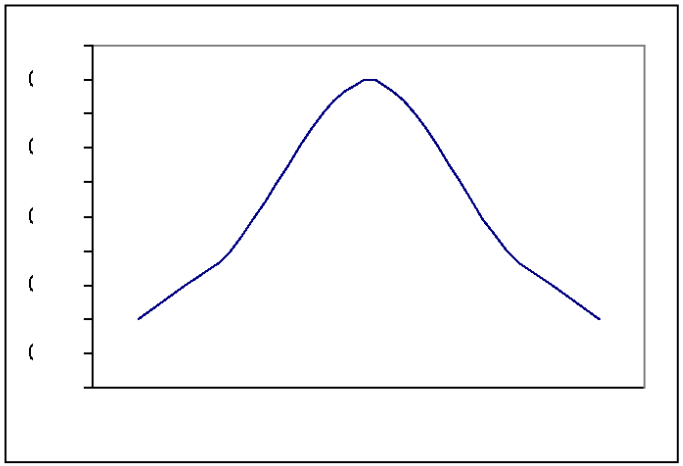
El Primer Cuartil (Q_1) es un valor de la variable que divide el 25 % de los valores más bajos del 75 % de los valores restantes.

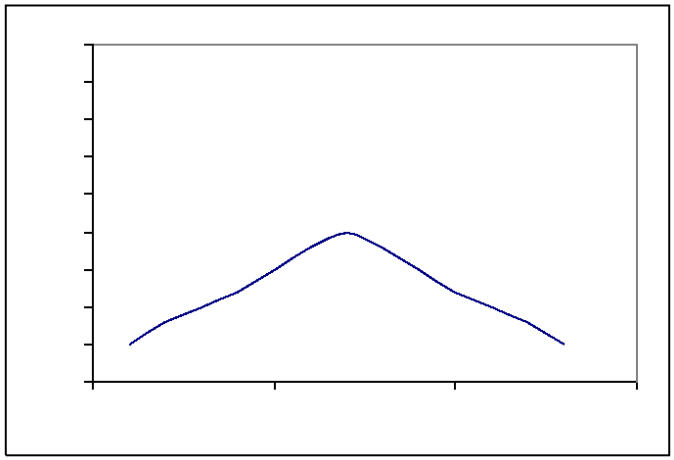
Me

El Tercer Cuartil (Q_3) es un valor de la variable que divide el 25 % de los valores más altos del 75 % de los valores restantes.

Medidas de dispersión

La variación de los valores de un conjunto de datos se llama dispersión y se refiere a la mayor o menor concentración de valores en torno a un valor particular, por lo general de tendencia central.





Medidas de dispersión

Varianza

Es la media aritmética del cuadrado de las desviaciones de cada observación respecto a su media.

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

Al calcularse como los desvíos al cuadrado de la variable respecto a la media está elevada a una magnitud superior que la variable original, por eso al interpretarla no resulta útil.

Desviación estándar o típica

Medidas de dispersión

Desviación estándar

- Se calcula con respecto a la media aritmética
- Cuanto mayor sea la dispersión, mayor será el valor de la varianza y desviación estándar

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

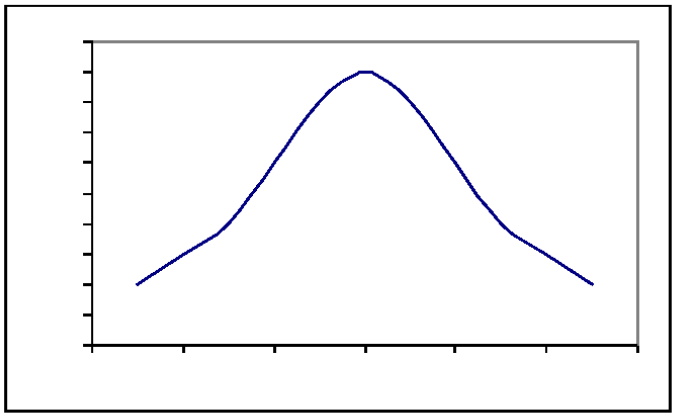
Coeficiente de variación

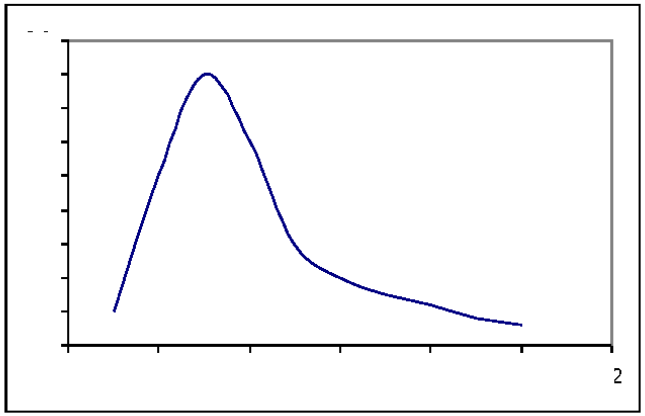
$$CV(x) = \frac{S}{\bar{X}}$$

Dos conjuntos de datos son comparables a través de este coeficiente. Se trata de una medida adimensional de dispersión

Es una medida de dispersión relativa, refleja la desviación estándar como porcentaje de la media.

Medidas de Forma



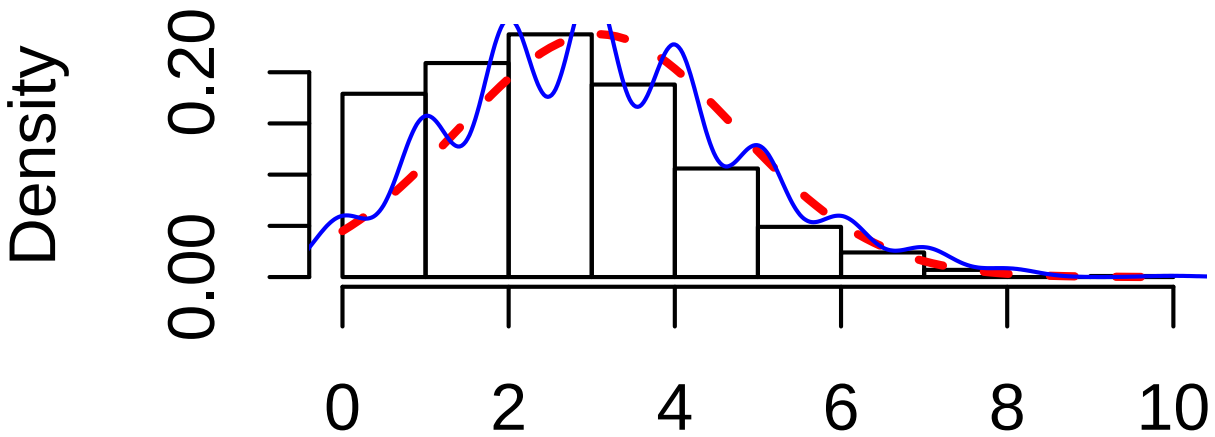


Tienen que ver con la **puntiagudez o curtosis** (deformación vertical) y la **asimetría** (deformación horizontal) del conjunto de datos

Descriptivos en R

$$f(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \textit{para } x = 0, 1, 2, 3... \quad \lambda \in (0, \infty) \\ 0 & \textit{en caso contrario} \end{cases}$$

```
X<-rpois(1000, 3)
```



Descriptivos en R

```
#Medidas de posición

mean(X)

## [1] 3.1

quantile(X, probs = 0.5)

## 50%
##   3

quantile(X)

##   0%  25%  50%  75% 100%
##    0    2    3    4   10
```

Descriptivos en R

```
#Medidas de dispersión
var(X)

## [1] 2.8

sd(X)

## [1] 1.7

sd(X)/mean(X)

## [1] 0.55
```

Descriptivos en R

```
#Medidas de forma
install.packages("e1071", dependencies = TRUE)
```

```
library(e1071)

skewness(X)

## [1] 0.41

kurtosis(X)

## [1] 0.05
```

Gracias!!!