Análisis de Datos en Python: descriptivo e inferencial 3

Dora Suárez, Juan F. Pérez

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

juanferna.perez@urosario.edu.co

2018

Contenidos

Estimadores puntuales

Inferencia a partir de una muestra aleatoria

Población: X

- Valor esperado $\mu = E[X]$
- Varianza $\sigma^2 = V[X]$
- Desviación estándar $\sigma = \sqrt{V[X]}$

Inferencia a partir de una muestra aleatoria

Muestra aleatoria $\{X_1, \ldots, X_n\}$:

- Media muestral:

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

Inferencia a partir de una muestra aleatoria

Muestra aleatoria $\{X_1, \ldots, X_n\}$:

- Media muestral:

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

- Varianza muestral:

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

Estimadores puntuales

- Media muestral como estimador de la media poblacional:

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

Estimadores puntuales

- Media muestral como estimador de la media poblacional:

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

- Varianza muestral como estimador de la varianza muestral:

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

Estimadores puntuales

- Media muestral como estimador de la media poblacional:

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

- Varianza muestral como estimador de la varianza muestral:

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

- Obtengo un número que uso para estimar el valor del parámetro

Cargue y descripción de datos

```
# -*- coding: utf-8 -*-
import numpy as np
import scipy.stats as st
import pandas as pd

filename = "data/data_blood.txt"
datos = pd.read_csv(filename, header=None, sep="\s+",
          names = [u'Indice', u'Uno', u'Edad',
                                        u'Presión Sangre'])

datosPresion = datos[u'Presión Sangre']
ser = pd.Series(datosPresion)
print(ser.describe())
```

Cargue y descripción de datos

```
n = len(datosPresion)
print("n: ", n)
print("media: ", np.mean(datosPresion))
print("desv estandar (n): ", np.std(datosPresion))
print("desv estandar (n-1): ",
      np.std(datosPresion, ddof = 1))
```

Cargue y descripción de datos

	Indice	Uno	Edad	Presion Sangre
count	30.000000	30.0	30.000000	30.000000
mean	15.500000	1.0	45.133333	142.533333
std	8.803408	0.0	15.294203	22.581245
min	1.000000	1.0	17.000000	110.000000
25 %	8.250000	1.0	36.750000	125.750000
50 %	15.500000	1.0	45.500000	141.000000
75 %	22.750000	1.0	56.000000	157.000000
max	30.000000	1.0	69.000000	220.000000

Cargue y descripción de datos

```
n:  30
media:  142.533333333
desv estandar (n):  22.2017016365
desv estandar (n-1):  22.581245397
```

Estimadores de intervalo

Estimadores de intervalo

- $\bar{X}$ estimador puntual de μ

Estimadores de intervalo

- $\bar{X}$ estimador puntual de μ
- Intervalo: $[a, b]$

Estimadores de intervalo

- $\bar{X}$ estimador puntual de μ
- Intervalo: $[a, b]$
- Alta probabilidad de que μ esté en el intervalo

$$P(\mu \in [a, b]) = 0{,}95$$

Estimadores de intervalo

- $\bar{X}$ estimador puntual de μ
- Intervalo: $[a, b]$
- Alta probabilidad de que μ esté en el intervalo

$$P(\mu \in [a, b]) = 0{,}95$$

- Aprovechando $\bar{X}$:

$$[\bar{X} - c, \bar{X} + c]$$

Estimadores de intervalo: media

- ¿Cómo se comporta $\bar{X}$?

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

Estimadores de intervalo: media

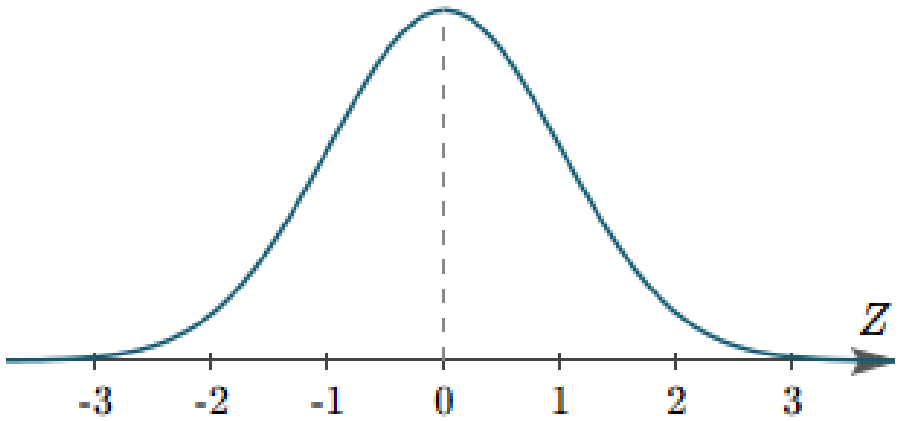
- ¿Cómo se comporta $\bar{X}$?

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

- Depende del comportamiento de X_i, es decir, de X

Variable aleatoria normal

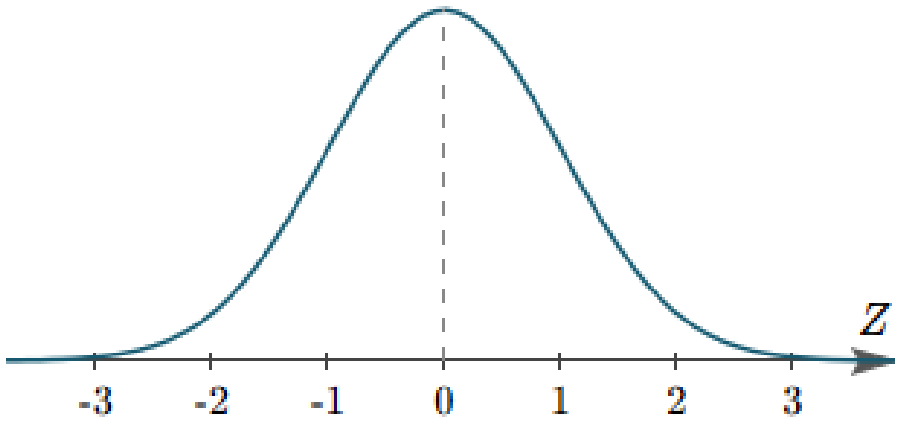
- Variable aleatoria continua



- Normal: `https://www.geogebra.org/m/QEayZCpM`

Variable aleatoria normal

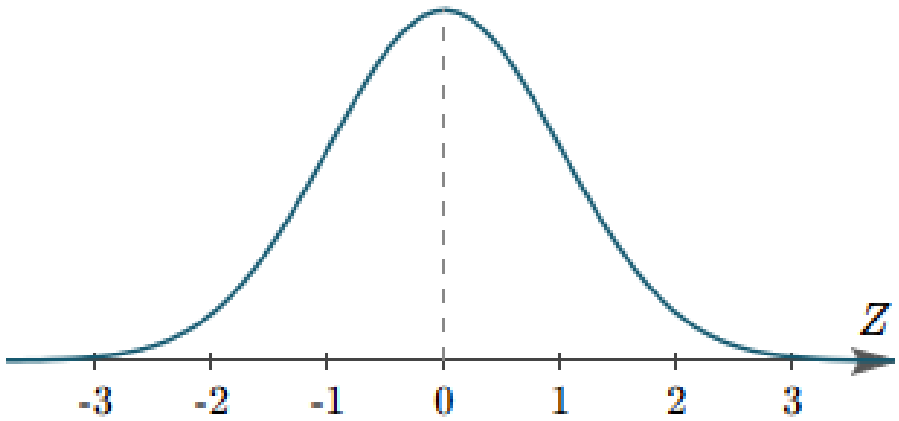
- Variable aleatoria continua
- Función de densidad de probabilidad (no de masa)



- Normal: `https://www.geogebra.org/m/QEayZCpM`
- Normal estándar ($\mu = 0$, $\sigma^2 = 1$): `https://www.geogebra.org/m/Xhp5vB98`

Variable aleatoria normal

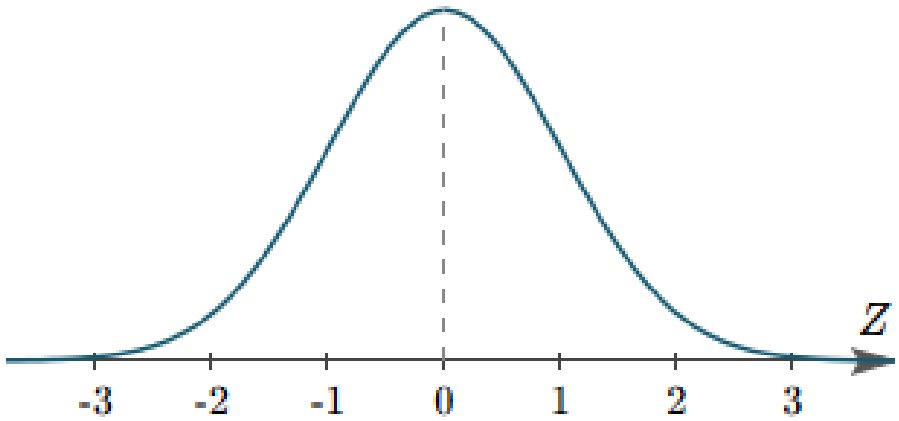
- Variable aleatoria continua
- Función de densidad de probabilidad (no de masa)
- Probabilidad: área bajo la curva



- Normal: `https://www.geogebra.org/m/QEayZCpM`
- Normal estándar ($\mu = 0$, $\sigma^2 = 1$): `https://www.geogebra.org/m/Xhp5vB98`

Variable aleatoria normal

- Variable aleatoria continua
- Función de densidad de probabilidad (no de masa)
- Probabilidad: área bajo la curva
- Parámetros: media μ y varianza σ^2



- Normal: `https://www.geogebra.org/m/QEayZCpM`
- Normal estándar ($\mu = 0$, $\sigma^2 = 1$): `https://www.geogebra.org/m/Xhp5vB98`

Estimadores de intervalo: media

- X sigue una distribución normal (μ, σ^2)

Estimadores de intervalo: media

- X sigue una distribución normal (μ, σ^2)
- Cada muestra X_i sigue la misma distribución normal

Estimadores de intervalo: media

- X sigue una distribución normal (μ, σ^2)
- Cada muestra X_i sigue la misma distribución normal
- La media muestral

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

sigue una distribución normal $\left(\mu, \frac{\sigma^2}{n}\right)$

Estimadores de intervalo: media

- Estimador de intervalo para la media μ:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Estimadores de intervalo: media

- Estimador de intervalo para la media μ:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$

Estimadores de intervalo: media

- Estimador de intervalo para la media μ:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$
- $\frac{\sigma}{\sqrt{n}}$: error estándar (variabilidad de $\bar{X}$)

Estimadores de intervalo: media

- Estimador de intervalo para la media μ:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$
- $\frac{\sigma}{\sqrt{n}}$: error estándar (variabilidad de $\bar{X}$)
- $z_{\alpha/2}$: factor que depende de la distribución normal

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)
- Resultado:

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)
- Resultado:

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)
- Resultado:

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$
- $\frac{S}{\sqrt{n}}$: error estándar (variabilidad estimada de $\bar{X}$)

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)
- Resultado:

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$
- $\frac{S}{\sqrt{n}}$: error estándar (variabilidad estimada de $\bar{X}$)
- $t_{\alpha/2}$: factor que depende de la **distribución T**

Estimadores de intervalo: media

- Problema: intervalo depende de σ (desconocido)
- Solución: reemplazar σ por su estimador puntual S (desviación estándar muestral)
- Resultado:

$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

- Punto medio: $\bar{X}$
- $\frac{S}{\sqrt{n}}$: error estándar (variabilidad estimada de $\bar{X}$)
- $t_{\alpha/2}$: factor que depende de la **distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`

Estimadores de intervalo: media

- **Distribución T**

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`
- Parámetro adicional (grados de libertad):
  - Cercano a uno: más variable/dispersa que la normal estándar

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`
- Parámetro adicional (grados de libertad):
  - Cercano a uno: más variable/dispersa que la normal estándar
  - Al llegar a 40: similar a la normal estándar

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`
- Parámetro adicional (grados de libertad):
  - Cercano a uno: más variable/dispersa que la normal estándar
  - Al llegar a 40: similar a la normal estándar
- Grados de libertad: asociados al número de observaciones

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`
- Parámetro adicional (grados de libertad):
  - Cercano a uno: más variable/dispersa que la normal estándar
  - Al llegar a 40: similar a la normal estándar
- Grados de libertad: asociados al número de observaciones
  - Pocas observaciones: más incertidumbre sobre el valor del parámetro

Estimadores de intervalo: media

- **Distribución T**
- `https://www.geogebra.org/m/RPGjU7Vz`
- Parámetro adicional (grados de libertad):
  - Cercano a uno: más variable/dispersa que la normal estándar
  - Al llegar a 40: similar a la normal estándar
- Grados de libertad: asociados al número de observaciones
  - Pocas observaciones: más incertidumbre sobre el valor del parámetro
  - Muchas observaciones: más certeza sobre el valor del parámetro

Estimadores de intervalo: media

- Intervalo de **confianza** para la media μ:

$$\left[\bar{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Estimadores de intervalo: media

- Intervalo de **confianza** para la media μ:

$$\left[\bar{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Garantiza que μ está en el intervalo con probabilidad $1 - \alpha$ (nivel de confianza)

Estimadores de intervalo: media

- Intervalo de **confianza** para la media μ:

$$\left[\bar{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Garantiza que μ está en el intervalo con probabilidad $1 - \alpha$ (nivel de confianza)
- Probabilidad de que esté por fuera del intervalo: α

Estimadores de intervalo: media

- Intervalo de **confianza** para la media μ:

$$\left[\bar{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Garantiza que μ está en el intervalo con probabilidad $1 - \alpha$ (nivel de confianza)
- Probabilidad de que esté por fuera del intervalo: α
- A mayor confianza $1 - \alpha$, más grande el intervalo

Estimadores de intervalo: media

- Intervalo de **confianza** para la media μ:

$$\left[\bar{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Garantiza que μ está en el intervalo con probabilidad $1 - \alpha$ (nivel de confianza)
- Probabilidad de que esté por fuera del intervalo: α
- A mayor confianza $1 - \alpha$, más grande el intervalo
- `https://www.geogebra.org/m/Xhp5vB98`

Calculando intervalos de confianza en python (versión 1)

Después de cargar los datos y tenerlos almacenados en la lista *datosPresion*

```
import numpy as np
import scipy.stats as st

intervalo = st.t.interval(0.95,
                len(datosPresion)-1,
                loc = np.mean(datosPresion),
                scale=st.sem(datosPresion) )
print(intervalo)
```

Calculando intervalos de confianza en python (versión 2)

Después de cargar los datos y tenerlos almacenados en la lista *datosPresion*

```
import statsmodels.stats.api as sms
intervalo = sms.DescrStatsW(datosPresion).
                tconfint_mean(0.05)
print(intervalo)
```

Calculando intervalos de confianza en python (resultado)

```
(134.10135772646 43, 150.96530894020236)
```

Probemos ahora con otros datos

- `https://goo.gl/xt3LJp`
- Y otros: `https://www.kaggle.com/jessicali9530/honey-production/data`