INSTITUTO
Data Science

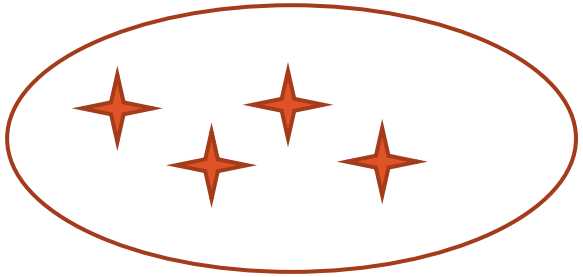
INTRODUCCIÓN A LA PROBABILIDAD Y LA ESTADÍSTICA

Conceptos fundamentales

Observación

Universo

Muestra



Variable aleatoria

Conceptos fundamentales

Medida: Resume las observaciones de un conjunto

Conceptos fundamentales

Medida: Resume las observaciones de un conjunto

- **Universo**
- **Muestra**

Conceptos fundamentales

Medida: Resume las observaciones de un conjunto

- **Universo**
- **Muestra**

Ejemplos:

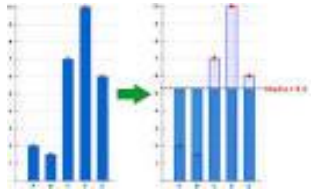
- Promedio
- Mediana
- Desvío estándar

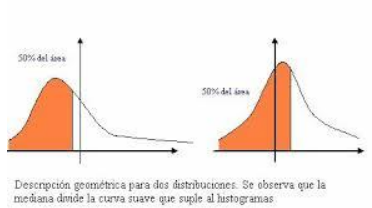
Conceptos fundamentales

Estadística Descriptiva	Conozco el universo y quiero las medidas que lo resumen
Inferencia Estadística	Conozco la muestra y quiero las medidas que resumen el universo

Estadística Descriptiva

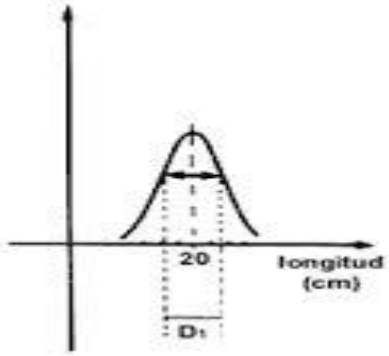
Medidas de Posición

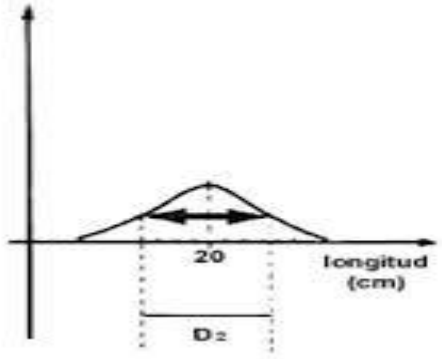




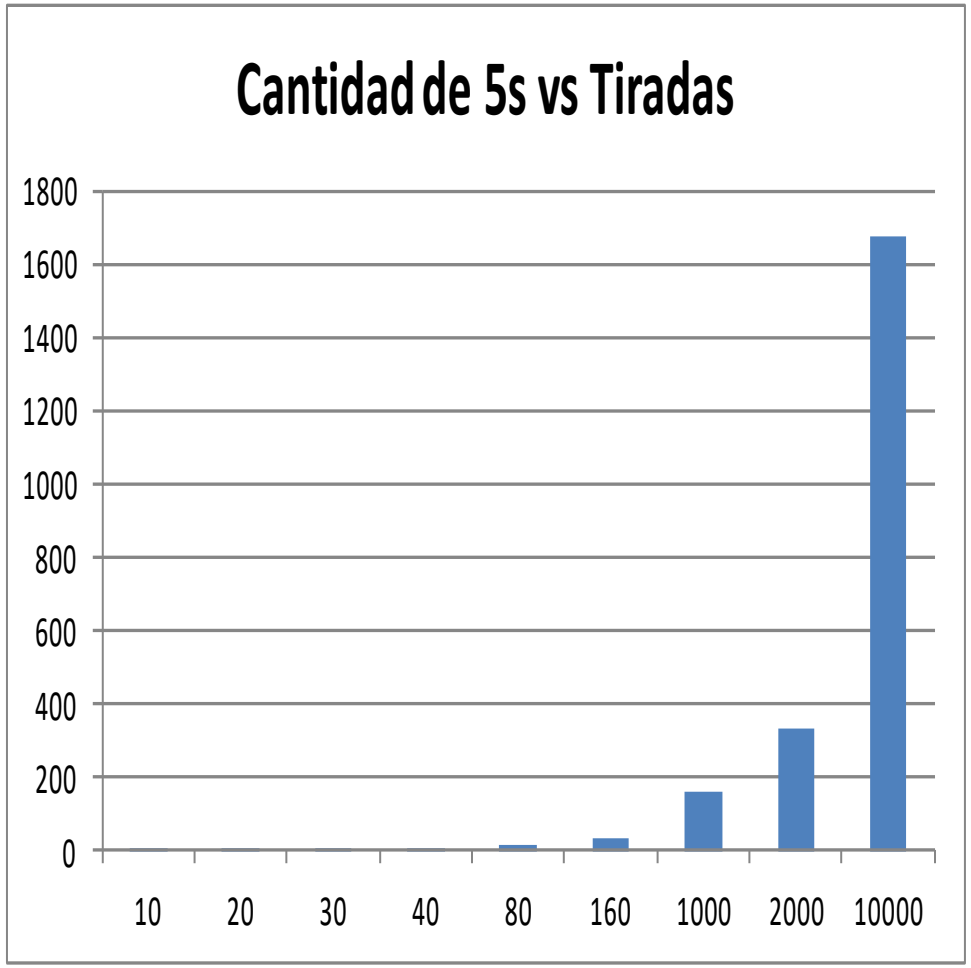
Descripción geométrica para dos distribuciones. Se observa que la mediana divide la curva suave que suple al histogramas

Medidas de Dispersión

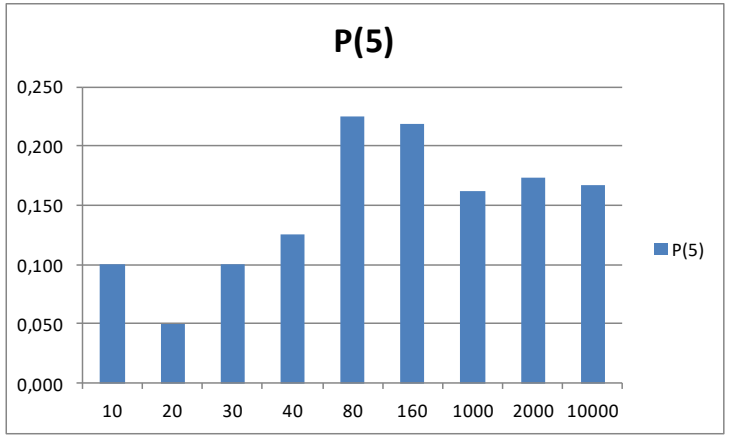




Variable Aleatoria



Tiradas	Cincos
10	1
20	1
30	3
40	5
80	18
160	34
1000	160
2000	330
10000	1670

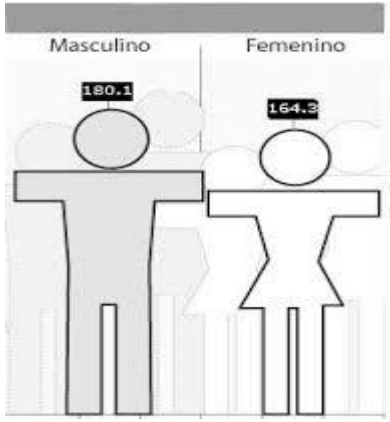


Promedio

- ¿Qué es?

- ¿Cómo calcularlo?

- ¿Para qué sirve?



Mediana

- ¿Qué es?

- ¿Cómo calcularla?

$$M_e = \begin{cases} X_{(\frac{n+1}{2})}, & \text{se "n" é impar} \\ \dfrac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & \text{se "n" é par} \end{cases}$$

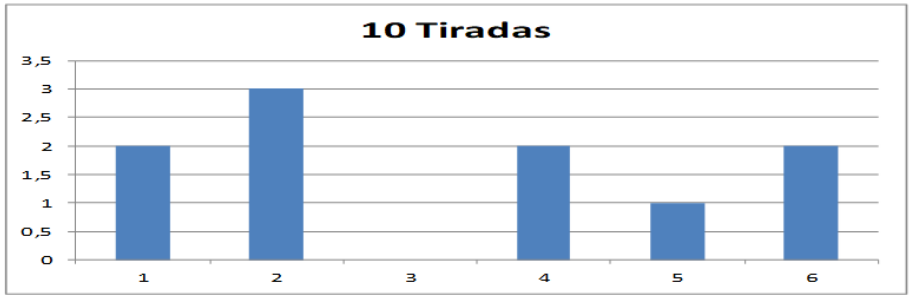
Desvío Estándar

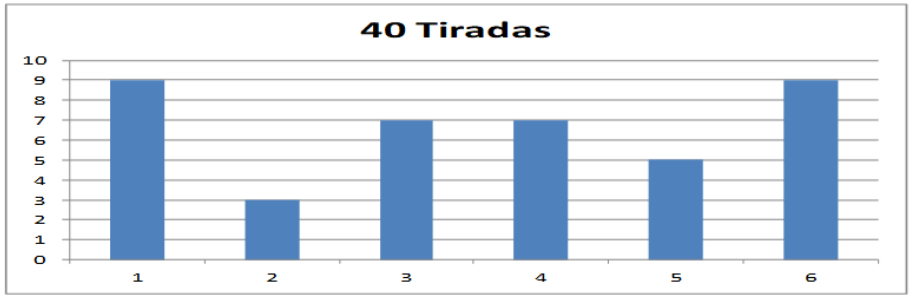
- ¿Qué es?

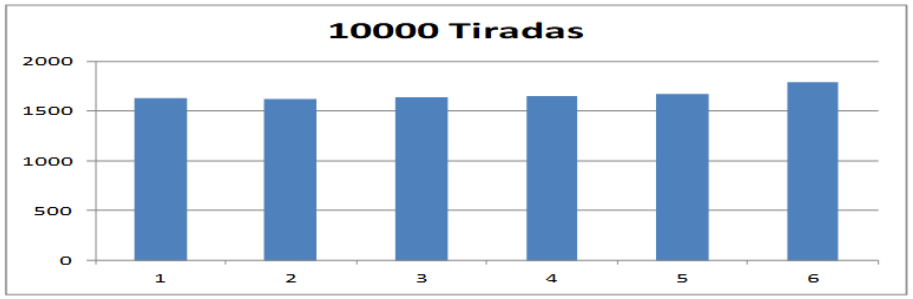
- ¿Cómo calcularlo?

$$s = \sqrt{\frac{\sum_{i=1}^{N}(Xi - \bar{X})^2}{N - 1}}$$

Histograma



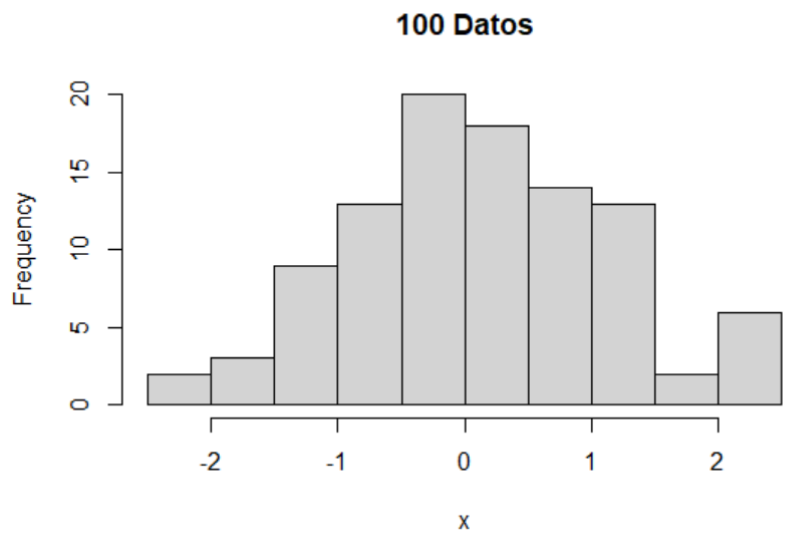


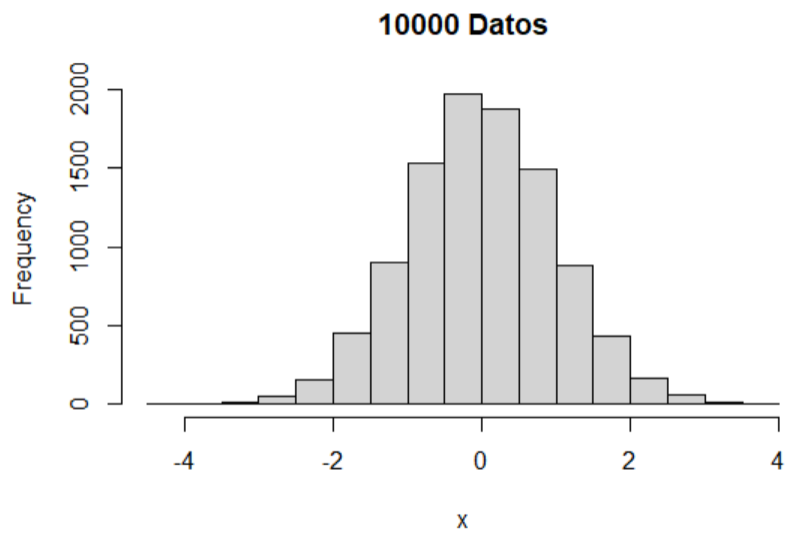


Tiradas	Cincos
10	1
20	1
30	3
40	5
80	18
160	34
1000	160
2000	330
10000	1670

Histogramas

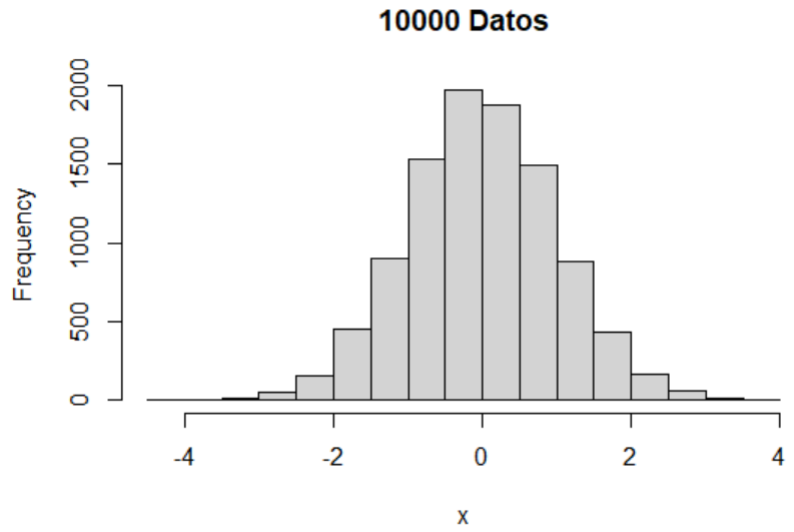
- Más datos veo mejor

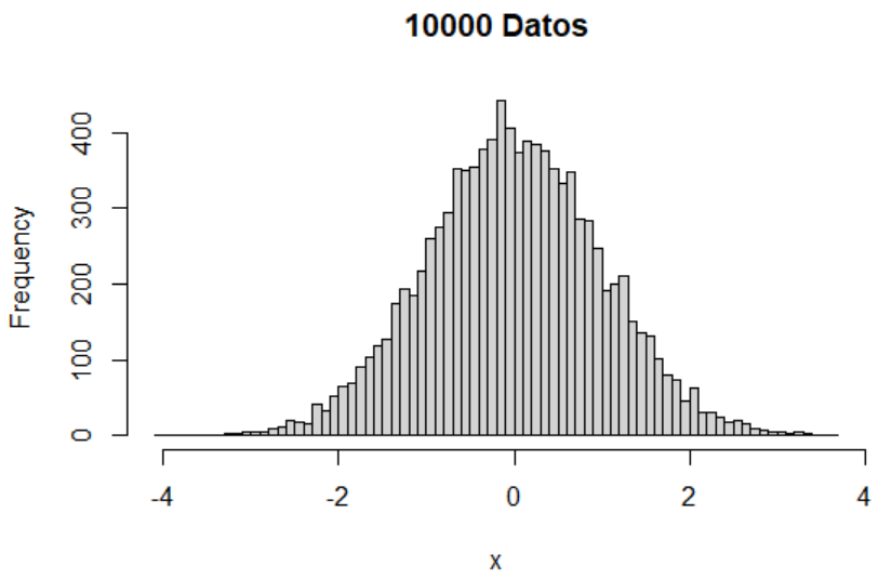




Histogramas

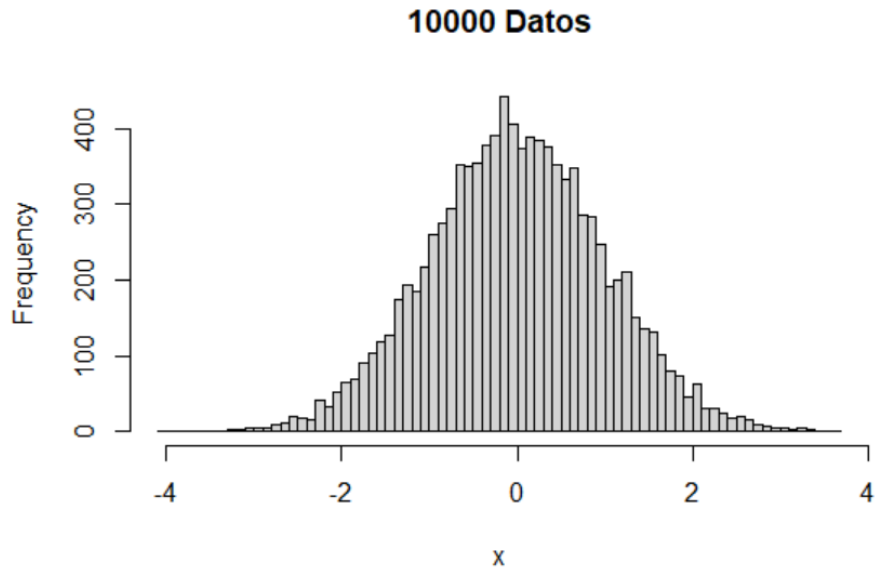
- Más divisiones veo peor

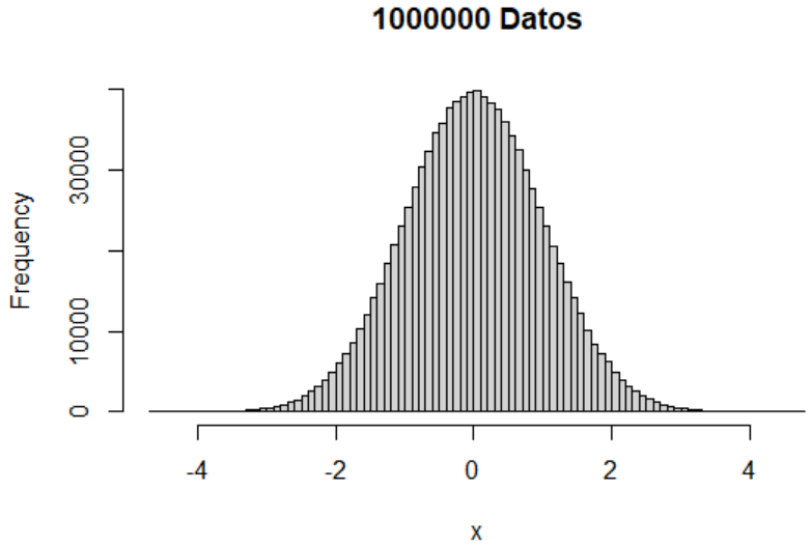




Histogramas -> Distribuciones

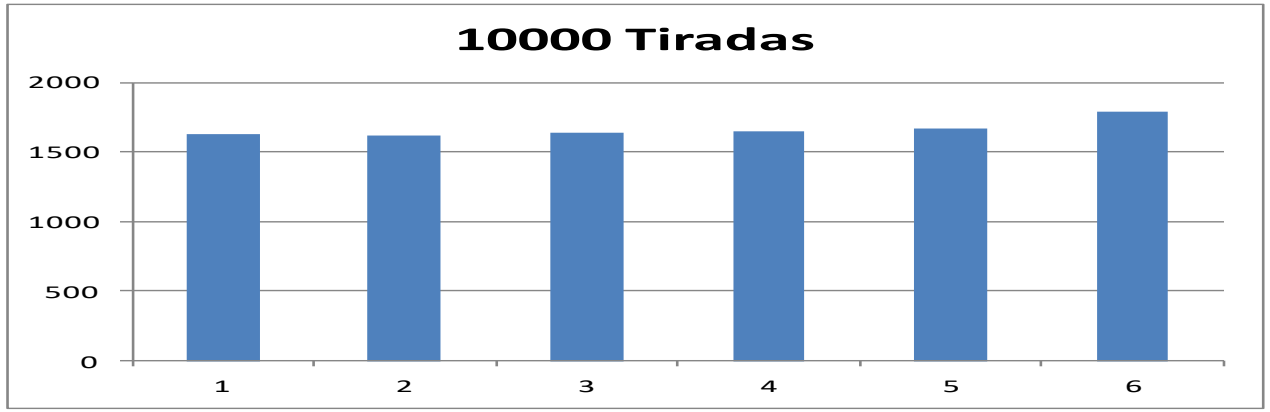
- La cantidad de datos crece
- La cantidad de divisiones también



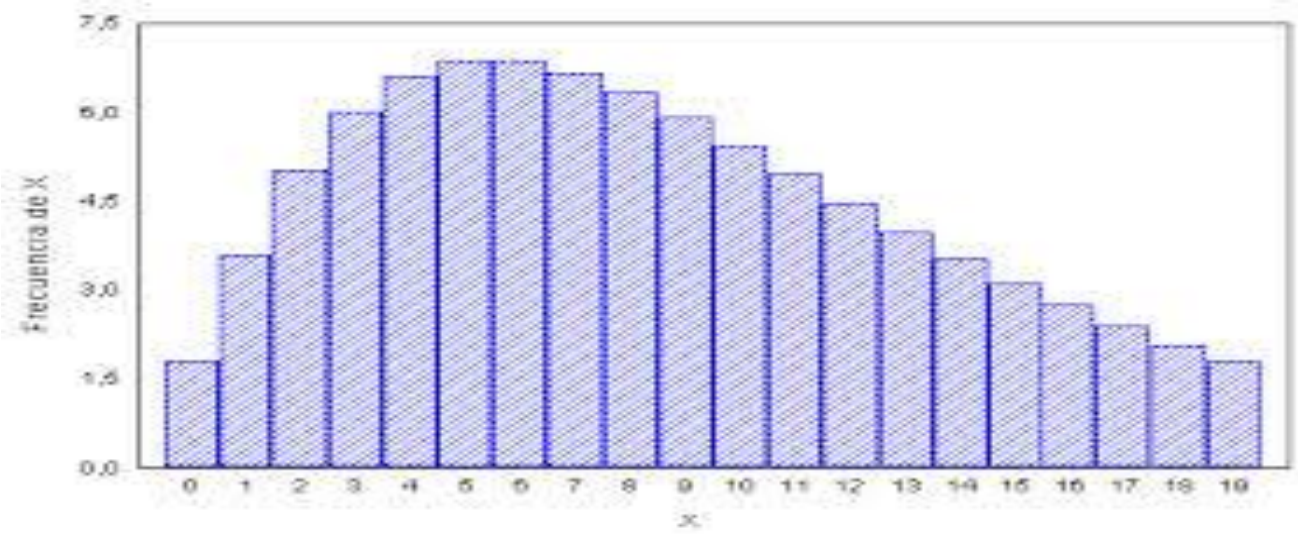


Distribuciones discretas

- Uniforme

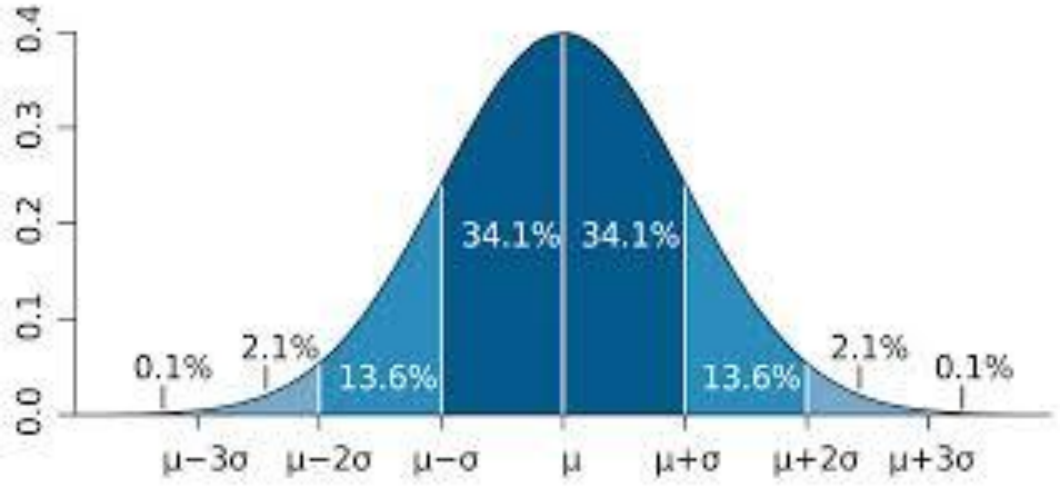


- Binomial

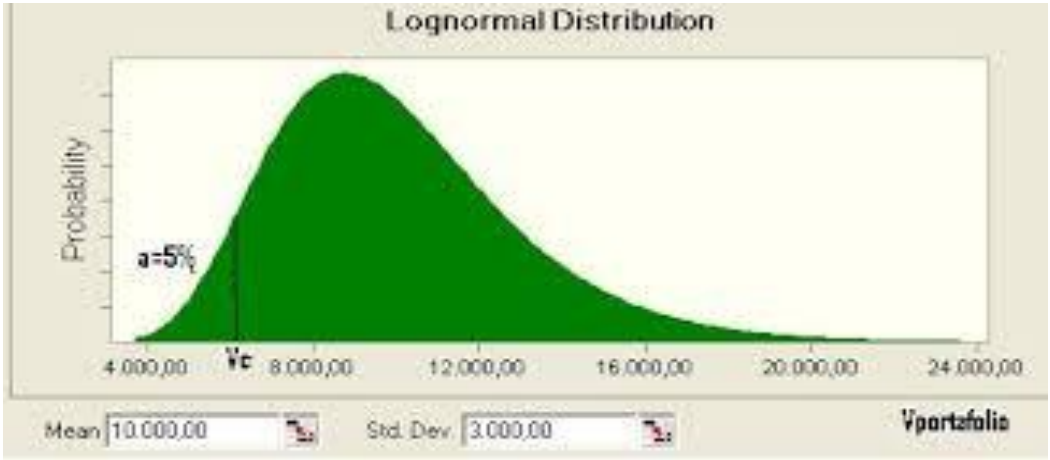


Distribuciones continuas

- Normal



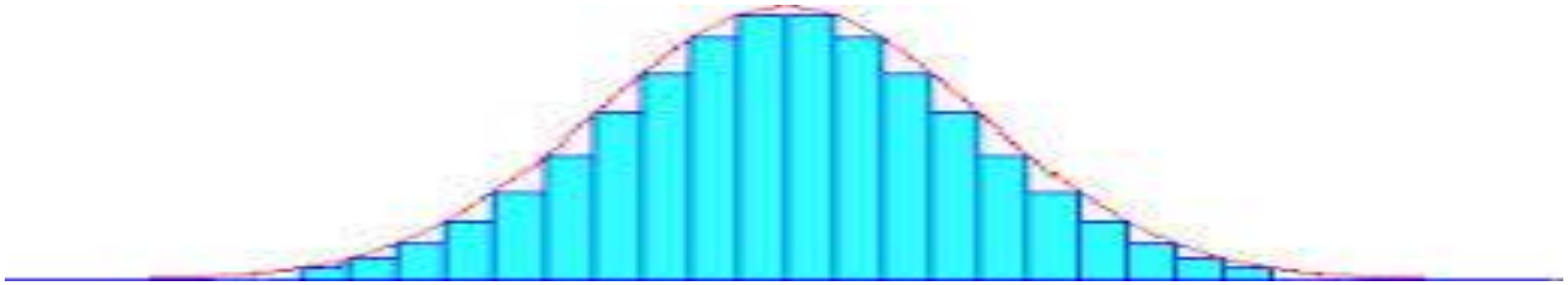
- Log Normal



Inferencia estadística

- ¿Qué es?

  - Conocemos una muestra
  - Queremos saber algo del universo



Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

¿Se puede?

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

¿Se puede?
Si, pero…

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

¿Se puede?
Si, pero…
Aparece la incertidumbre

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

¿Cómo nos damos cuenta?

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

¿Conseguimos 10000 grupos de 100?

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

- Media1: 168.3
- Media 2: 174.7
- Media 3: 177.8
- Media 4: 174.4
- Media 5: 176.5

Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

- Media1: 168.3
- Media 2: 174.7
- Media 3: 177.8
- Media 4: 174.4
- Media 5: 176.5

¿Cómo interpretamos esto?

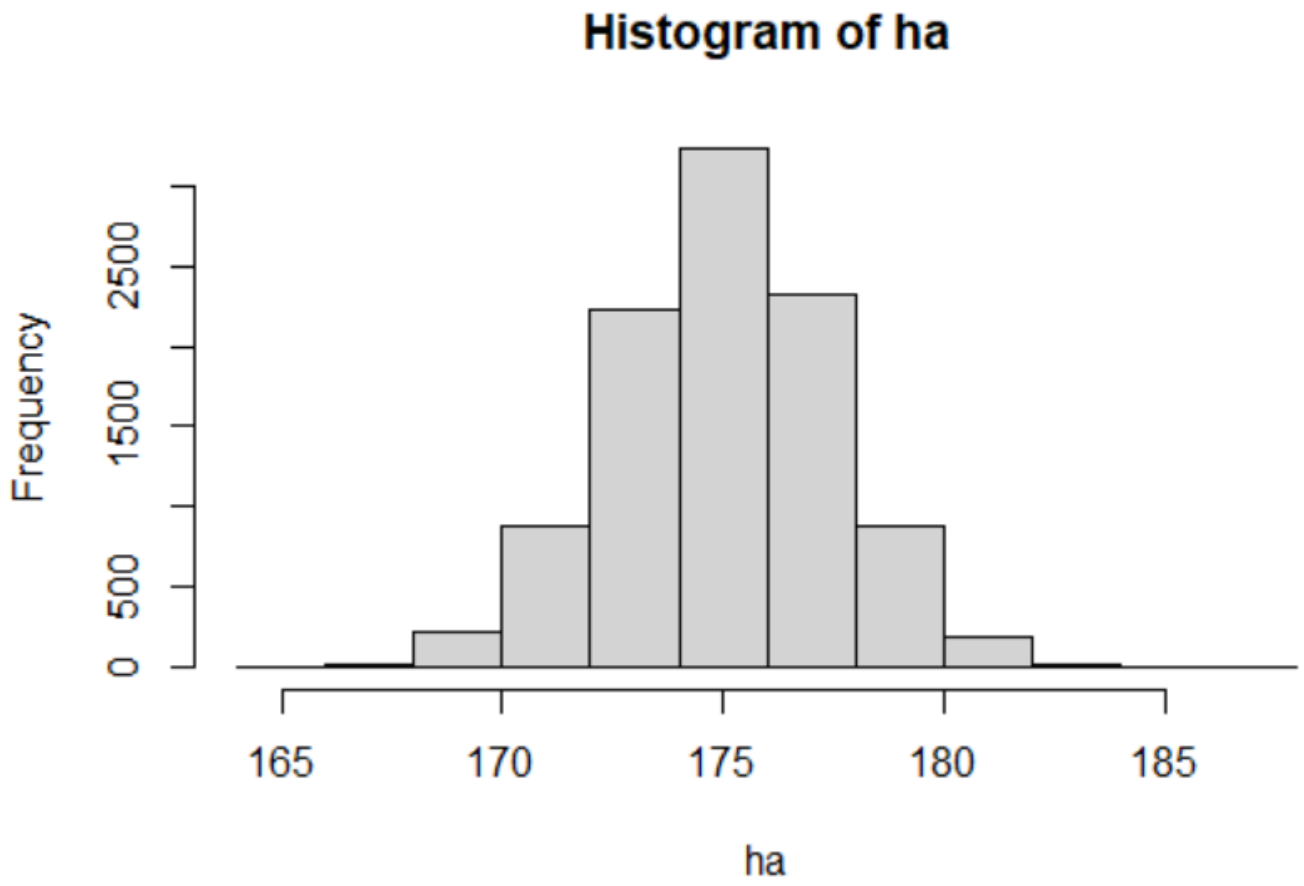
Ejemplo de inferencia estadística

- Conozco las alturas de 100 argentinos
- Quiero conocer el promedio de la altura de todos los argentinos

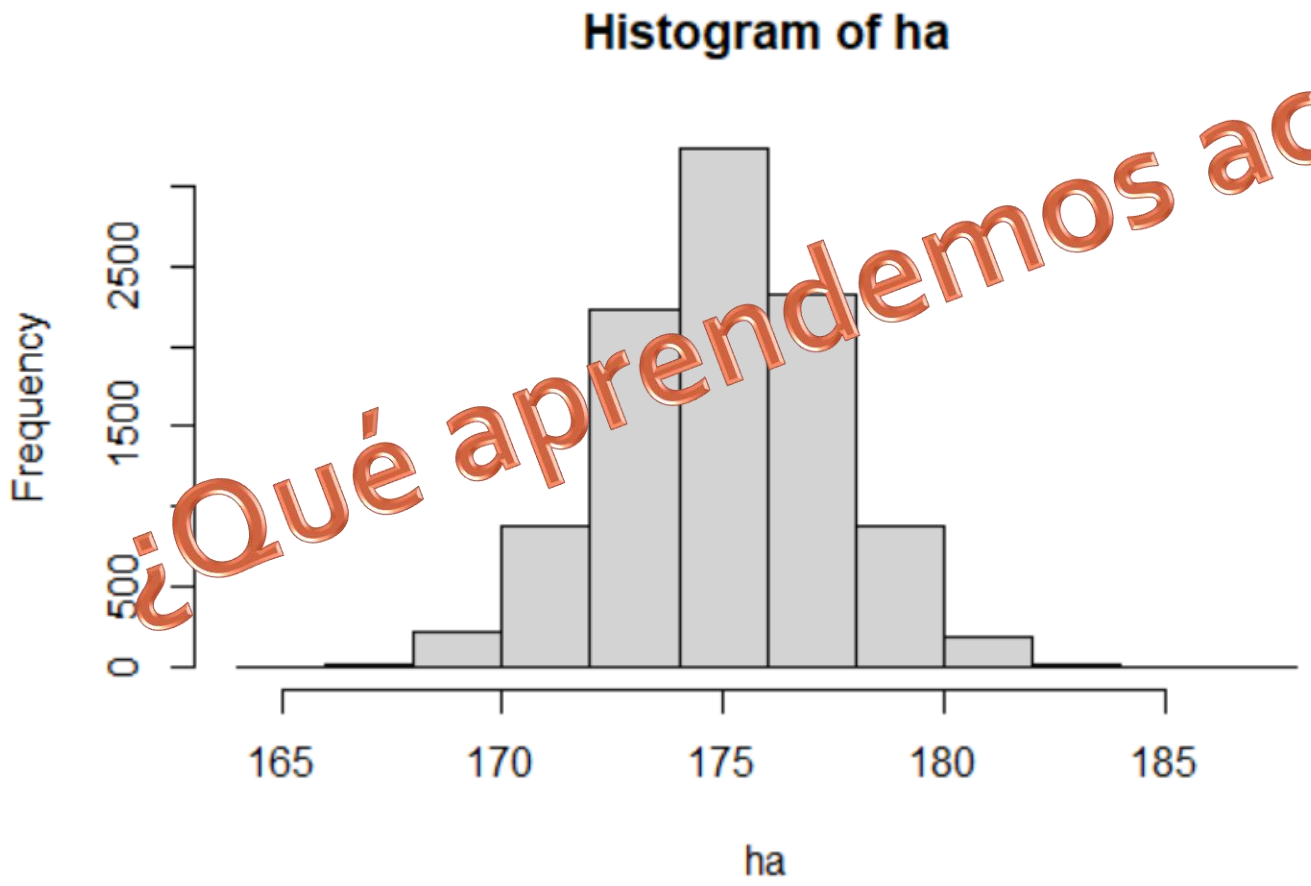
- Media1: 168.3
- Media 2: 174.7
- Media 3: 177.8
- Media 4: 174.4
- Media 5: 176.5

Las medias muestrales siguen una distribución

Distribución de las medias muestrales



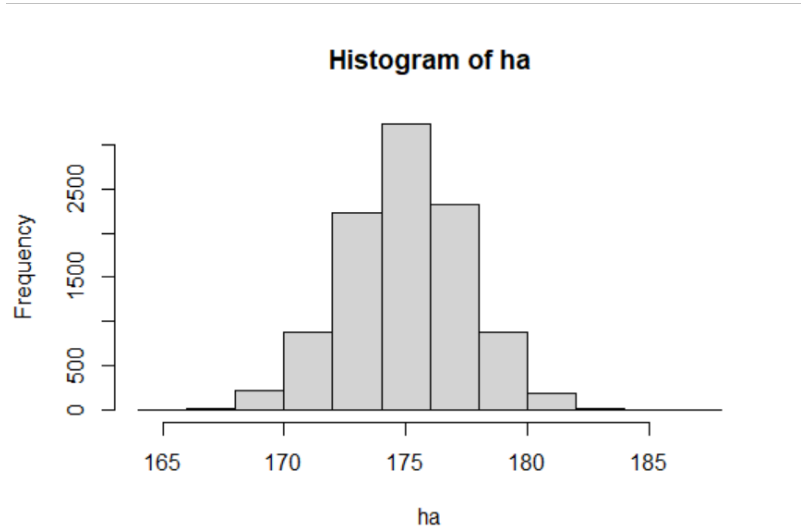
Distribución de las medias muestrales



¿Qué aprendemos acá?

¿Qué aprendemos?

- La media del universo debe estar en la zona de 175
- Sería raro que estuviera por debajo de 170
- Sería raro que estuviera por arriba de 180
- ¿Cuán raro?
- P(menor que 170) = 2.48%
- P(mayor que 180) = 2.15%



- Existe algo más del 95% de probabilidades de que la media universal esté entre 170 y 180

Estimadores

- Sirven para calcular medidas del universo desde los datos de la muestra
- Siempre pagan el precio de la incertidumbre

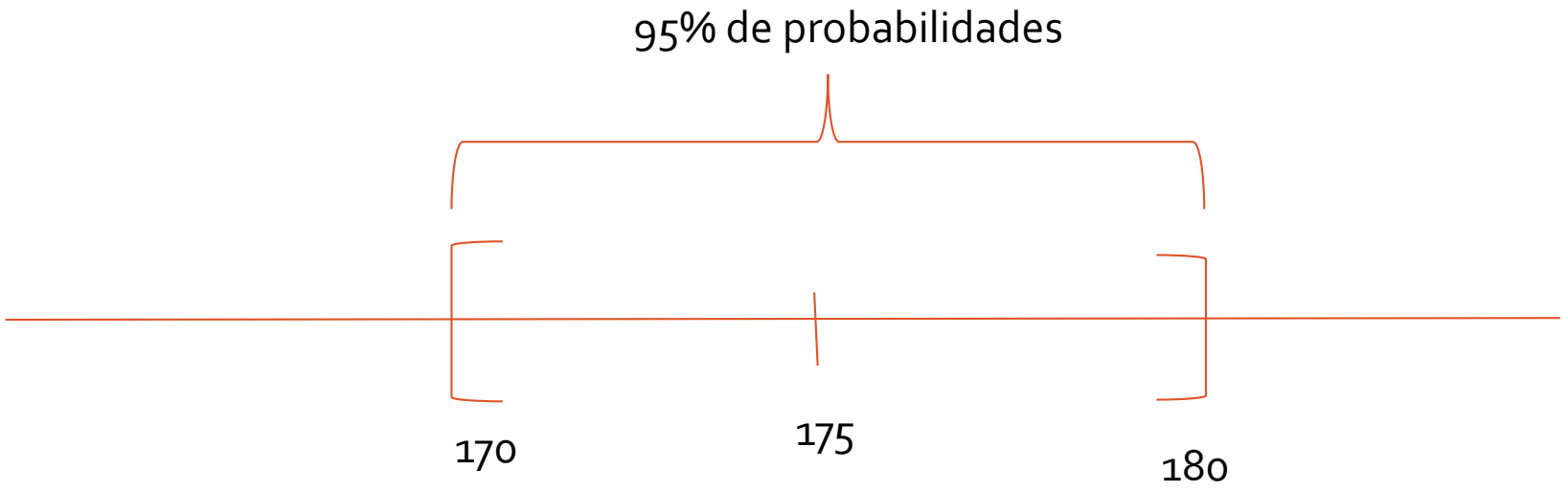
- Pueden ser sesgados o no sesgados:
  - Media de la muestra: estimador no sesgado de la media del universo
  - Desvío estándar de la muestra: estimador sesgado del desvío estándar del universo

Estimador no sesgado del desvío estándar:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(Xi - \bar{X})^2}{N-1}}$$

Intervalo de confianza

- Es el conjunto de números donde existe una probabilidad definida de encontrar la medida del universo.
- Esa probabilidad definida se conoce como **nivel de confianza**

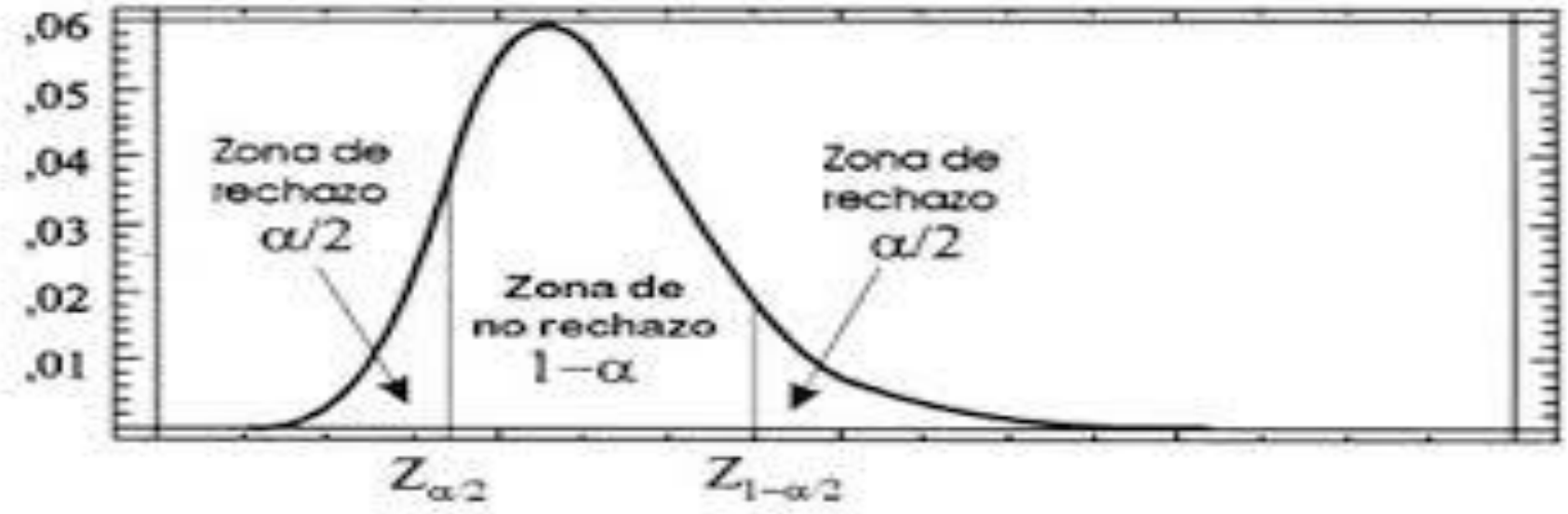


Test de Hipótesis

- ¿En que consiste?

Tabla 1. Decisiones en la prueba de hipótesis

		Verdad en la población	
		Hipótesis nula falsa	Hipótesis nula verdadera
Resultado de la prueba de hipótesis	Rechazar hipótesis nula	Potencia $1-\beta$	Error tipo I α
	No rechazar hipótesis nula	Error tipo II β	$1-\alpha$



+distr.binom(exitos; casos; probabilidad; acumulado)

Ejemplo dado

Cara	Veces que salió	alfa	1-alfa
1	170	0,6307	0,3693
2	235	1,0000	0,0000
3	163	0,3975	0,6025
4	105	0,0000	1,0000
5	168	0,5654	0,4346
6	159	0,2735	0,7265
	1000		

Mecanismo del test de hipótesis

- Identificar las hipótesis nula y alternativa
- Identificar el nivel de confianza
- ¿Test de medias, proporciones o desvíos?
- Calcular la variable normalizada
- ¿Es un test a una o dos colas?
- Calcular la zona de aceptación

Ejemplo de test de hipótesis

Tenemos una pizzería en internet.

Tenemos dudas por el color del fondo:

- Rojo
- Amarillo

Se hizo un experimento:

A los que se conectaron en:

- Segundo par: fondo rojo
- Segundo impar: fondo amarillo

Ejemplo de test de hipótesis

Datos obtenidos:

Fondo	Casos	Pedidos	Monto	Desvío estándar
Rojo	1.215	350	351.500	332
Amarillo	1.195	365	340.000	352

Ejemplo de test de hipótesis

Hipótesis nula:

El ticket promedio es el mismo

Hipótesis alternativa:

El ticket promedio es distinto

Ejemplo de test de hipótesis

Calculo de la variable normalizada:

X medio:351.500/350

Y medio:340.000/365

S1:332

S2:352

N1:350

N2:365

$$T_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Da aproximadamente 2.84

Ejemplo de test de hipótesis

Nivel de confianza:

95%

Test de medias con distinto desvío estándar desconocido:

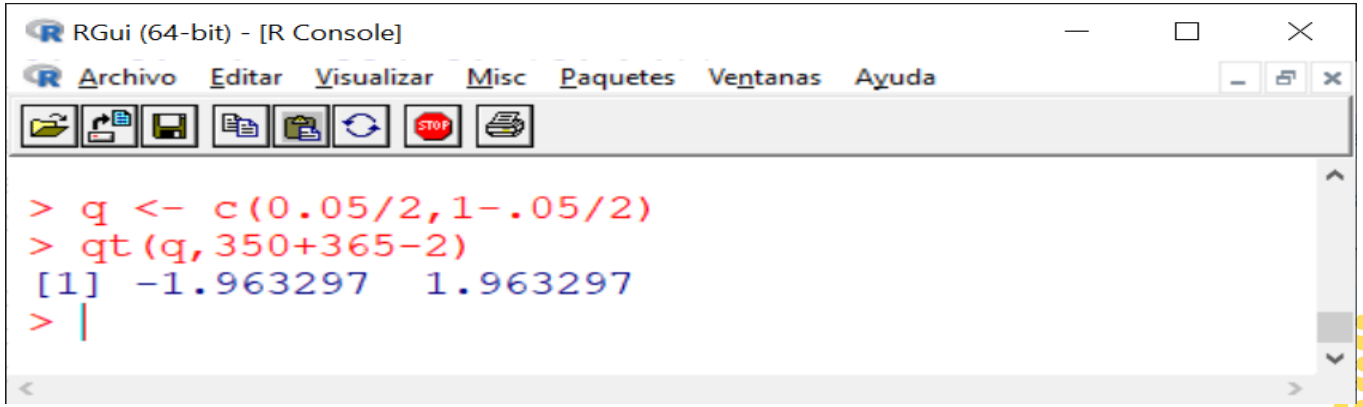
$$T_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Ejemplo de test de hipótesis

Es un test a dos colas:

Porque el monto medio de la venta con fondo rojo puede ser tanto mayor como menor que con fondo amarillo:

Calculo la zona de aceptación



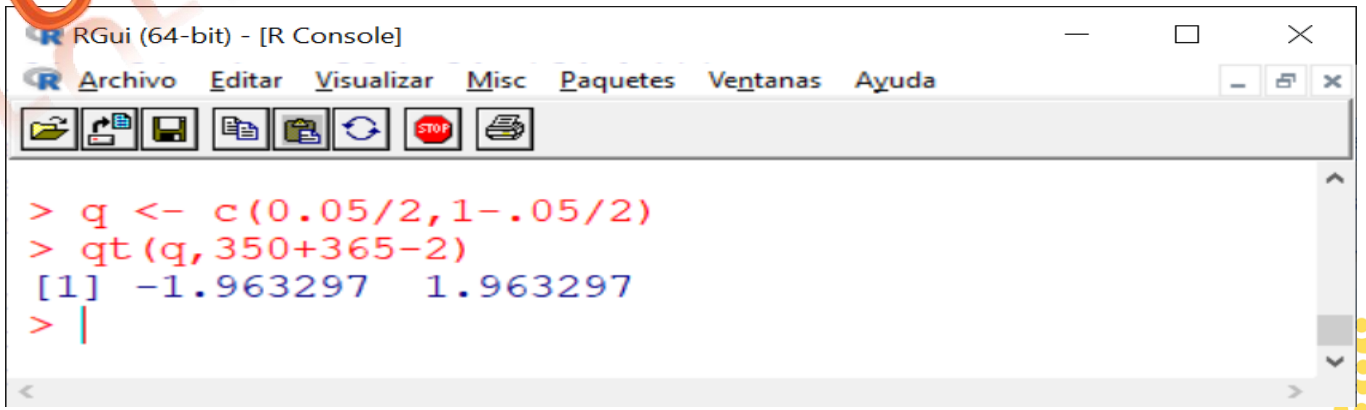
Ejemplo de test de hipótesis

Es un test a dos colas:

Porque el monto medio de la venta con fondo rojo puede ser tanto mayor como menor que con fondo amarillo:

2.84 CAE FUERA DE (-1.96,1.96)

Calculo la zona de aceptación



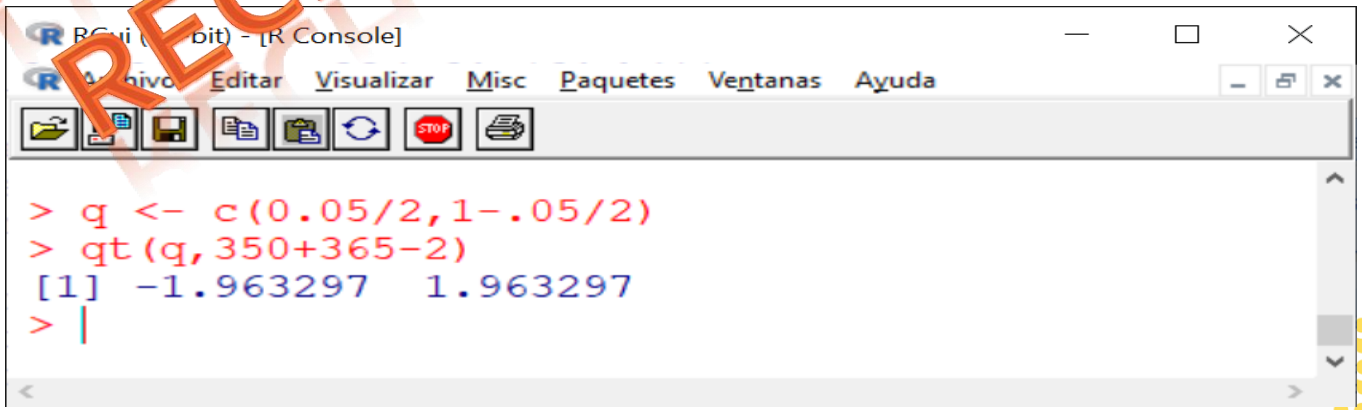
Ejemplo de test de hipótesis

Es un test a dos colas:

Porque el monto medio de la venta con fondo rojo puede ser tanto mayor como menor que con fondo amarillo:

2.84 CAE FUERA DE (-1.96,1.96) RECHAZO!!!

Calculo la zona de aceptación



Caso	Variable normalizada
Comparar medias con el mismo desvío estándar previamente conocido	$z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Comparar medias con el mismo desvío estándar pero desconocido	$T_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Comparar medias con distinto desvío estándar previamente conocido	$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Comparar medias con distinto desvío estándar pero desconocidos	$T_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Comparar proporciones	$z = \frac{\frac{x_1}{n_1} + \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
Comparar varianzas	$F = \frac{s_1^2}{s_2^2}$

Test de hipótesis de práctica

- Se sospecha que dos máquinas llenadoras de garrafas trabajan en forma distinta
- Tome los datos que están en el archivo DatosGarrafas.xlsx en el campus virtual.
- Diseñe un test de hipótesis que discuta si el llenado medio del equipo 1 es igual al del equipo 2 con un nivel de confianza del 90%
- Ejecute ese test para los 10, 20 y 50 primeros casos. ¿Qué conclusión obtiene en cada caso?
- Diseñe un test de hipótesis que discuta si el desvío estándar del llenador 1 es el mismo que el del llenador 2 para los 100 casos.

MUCHAS GRACIAS

INSTITUTO
Data Science