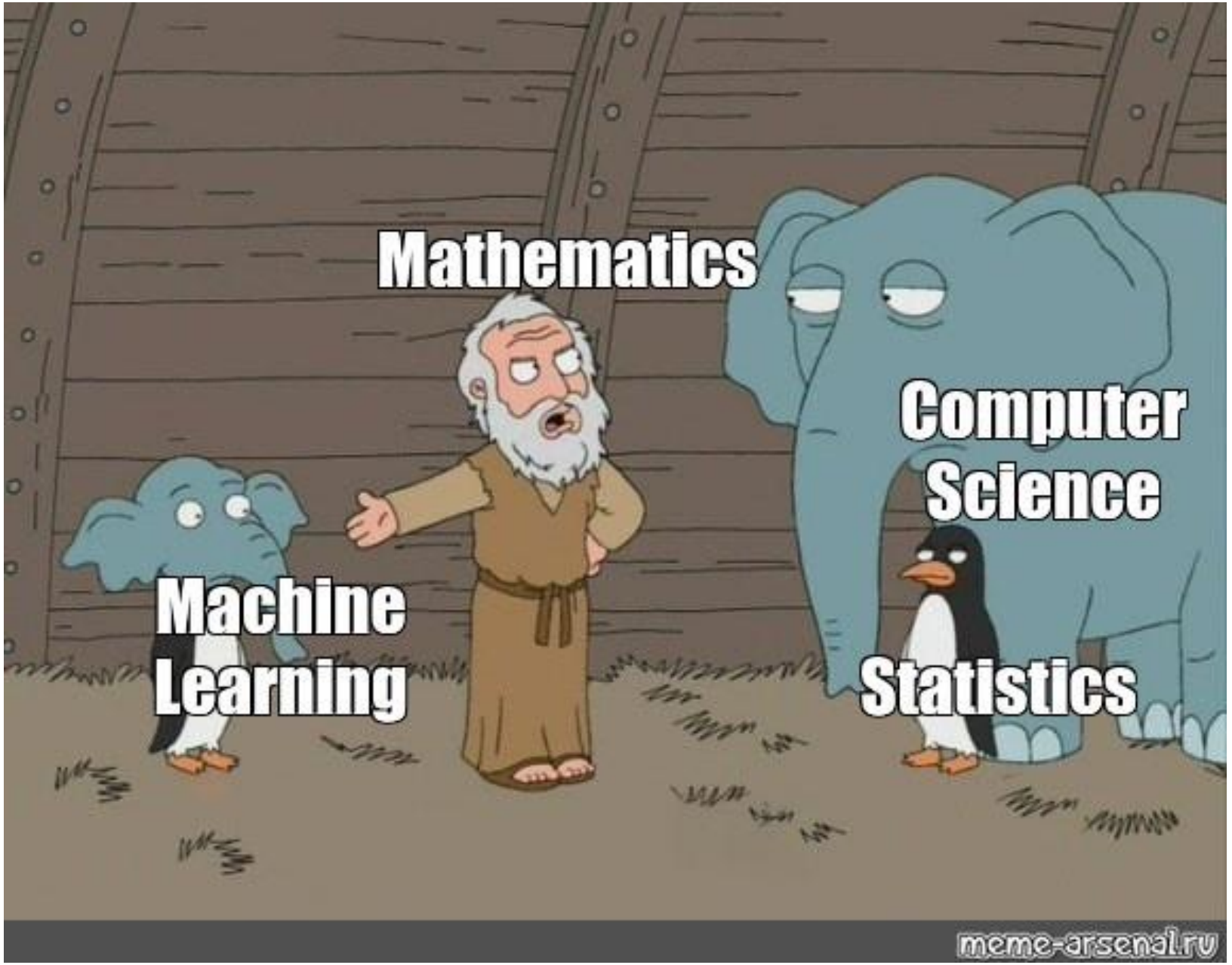
Inferencia estadística

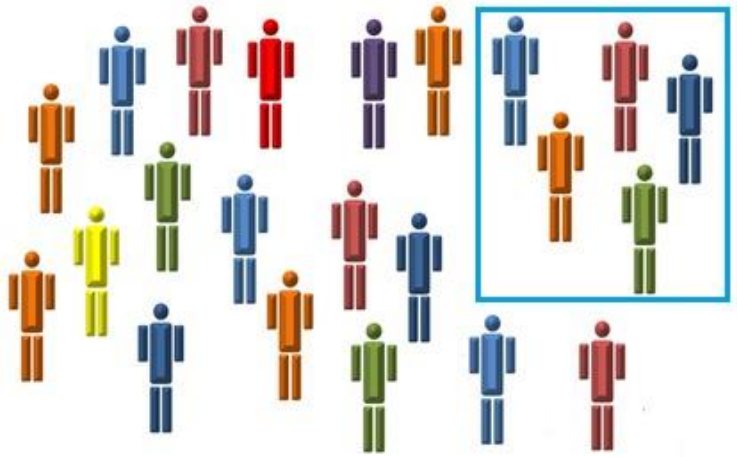
Hugo Andrés Dorado

Científico de datos

hugo.doradob@gmail.com


Mathematics
Computer Science
Machine Learning
Statistics
meme-arsenal.ru

Muestras



Inferencia estadística

Es el conjunto de métodos y técnicas que permiten sacar conclusiones de una **población** a partir de información de una **muestra**.

Homogeneidad
Tiempo
Espacio
Cantidad

Aleatorio
Estratificado
Sistemático

Diagnósticos y conclusiones basadas en evidencia.

Muestreo

- **Muestreo no probabilístico:** Obtiene muestras sin que todos los individuos de la población tengan posibilidades iguales de ser elegidos.
  - Muestreo intencional, opinático.
  - Muestreo por conveniencia
- **Muestreo probabilístico:** Todos los elementos a estudiar tiene una probabilidad conocida de formar parte de la muestra.
  - Muestreo aleatorio simple
  - Muestreo estratificado
  - Muestreo sistemático

Variable aleatoria

- Un valor numérico que está afectado por el azar.
- No es posible conocer con certeza el valor que tomará esta al ser medida o determinada.
- Se conoce que existe una distribución de probabilidad asociada al conjunto de valores posibles.
- Pueden ser discretas o continuas

https://es.wikipedia.org/wiki/Variable_aleatoria

Variable aleatoria ejemplos

X~ Valor obtenido al lanzar un dado

$Xi = \{ 1,2,3,4,5,6\}$

$P_i = \{1/6, 1/6, 1/6 , 1/6, 1/6, 1/6\}$

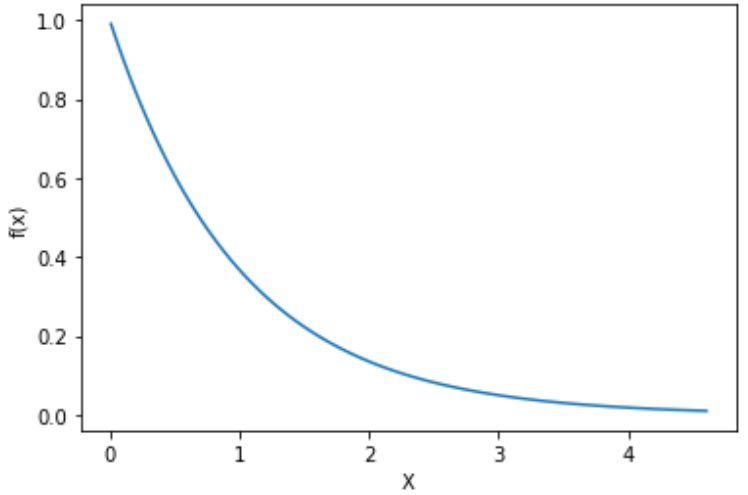
y~ Resultado de lanzar una moneda

$Xi = \{0,1\}$

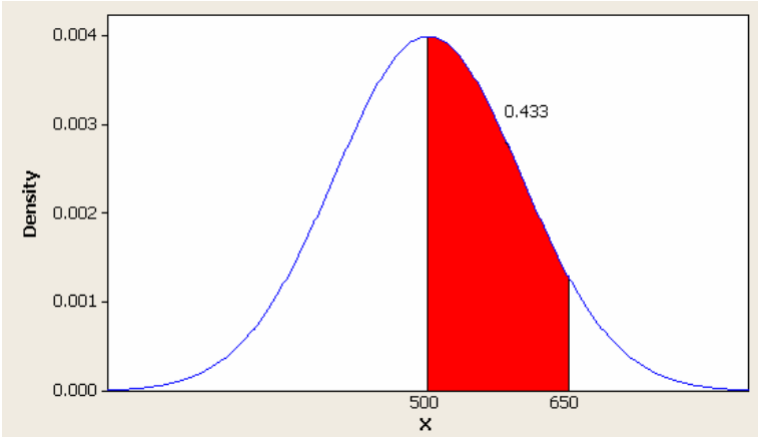
$P_i = \{1/2,1/2\}$

Variable aleatoria ejemplos

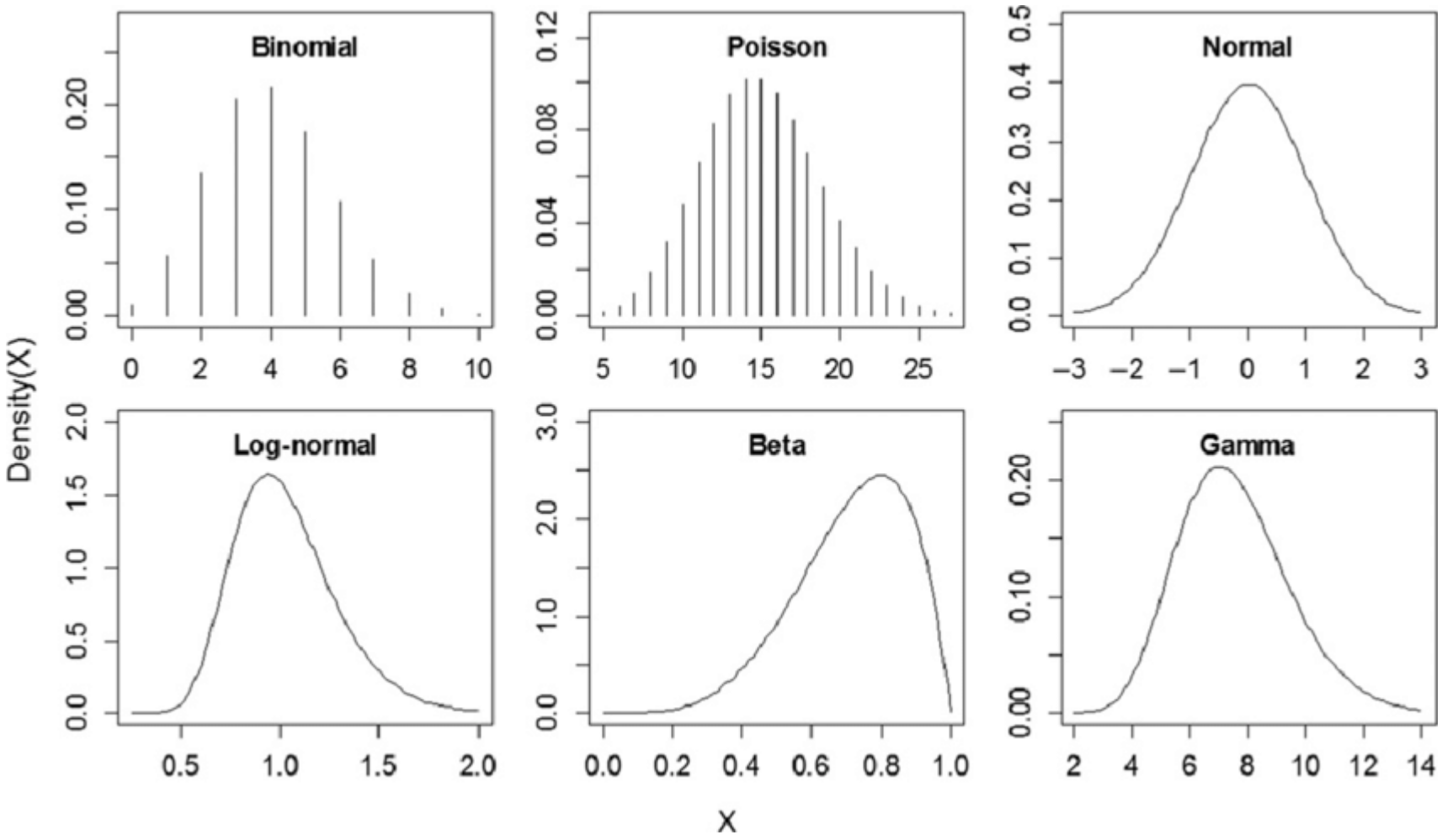
x ~ tiempo transcurrido en un centro de llamadas hasta recibir la primera llamada del día



W~ La longitud en cms de piezas producidas por un máquina



Algunas distribuciones de probabilidad



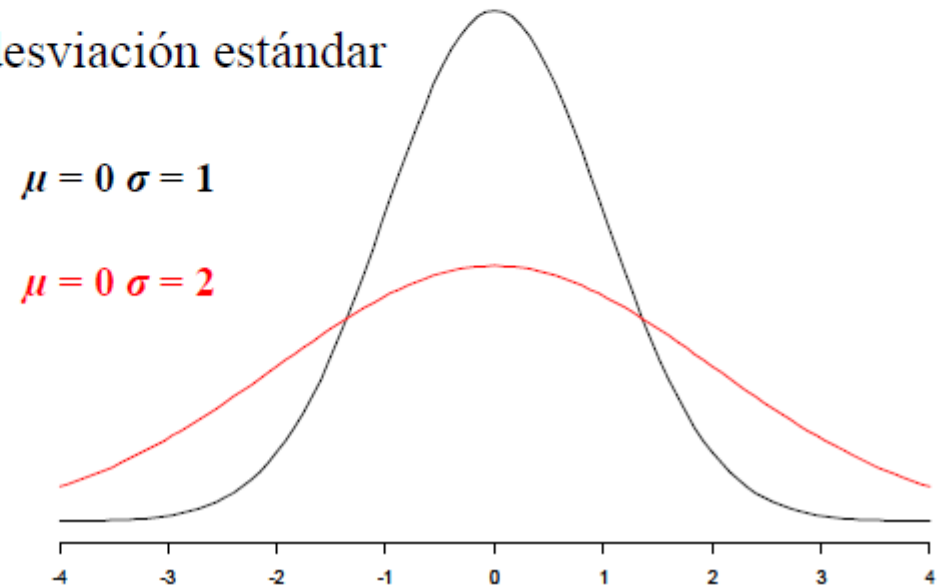
Distribución normal

Definición

Se dice que una variable aleatoria X tiene una distribución Normal, si su función de densidad es:

$$f(x) = N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

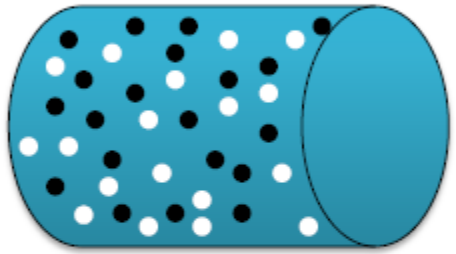
Donde μ es la media y σ la desviación estándar



Paramétros y estimadores

Parámetro

Característica medible sobre la población.

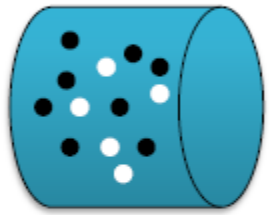


Ejemplo:

- Edad promedio de los estudiantes de ingeniería.
- Diámetro promedio de los tornillos fabricados en una empresa.

Estimador

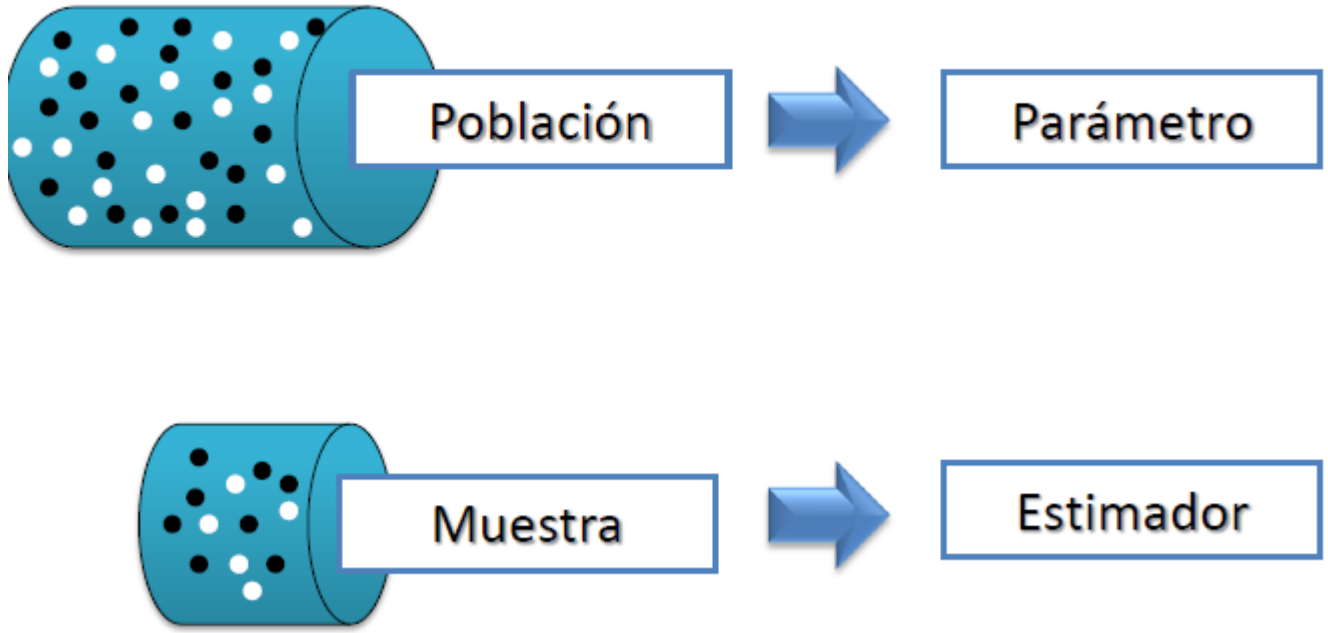
Característica medible sobre la muestra.



Ejemplo:

- Edad promedio de una muestra de los estudiantes de ingeniería.
- Diámetro promedio de una muestra de tornillos fabricados en una empresa.

Paramétros y estimadores



Estimación y error de muestreo

- **Estimación**: Procedimiento estadístico en el cual se logra una valoración del indicador asociado a una población no medida (Parámetro), mediante el estudio de una muestra aleatoria.

- **Error de muestreo:** la imprecisión que se comete al estimar una característica de la población de estudio (parámetro) mediante el valor obtenido a partir de una parte o muestra de esa población

$$\left|\bar{X} - \mu\right| \qquad \left|\hat{p} - P\right|$$

Tipos de estimación

Estimador puntual: Estadístico que estima el valor de un parámetro.

Intervalo de Confianza: Forma de estimar un parámetro en la cual se calcula un intervalo que indica con cierta seguridad un rango donde puede estar el parámetro.

Un estudio pretende estimar el porcentaje de hipertensos que hay entre las personas mayores de 65 años en la Comunidad Valenciana.

Estimador puntual: (167/350) x100=47.71%
Intervalo de confianza: **[42.48, 52.94]**

Paramétros y estimadores estadísticos

Valores puntuales

Comparaciones

θ	$\hat{\theta}$
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}$
$P = \frac{A}{N}$	$\hat{p} = \frac{a}{n}$
$\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2$
$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$
σ_1^2 / σ_2^2	S_1^2 / S_2^2

Distribuciones muestrales

Media muestral

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Proporción

$$\widehat{P} \sim N(P, \sqrt{\frac{P \cdot (100 - P)}{n}})$$

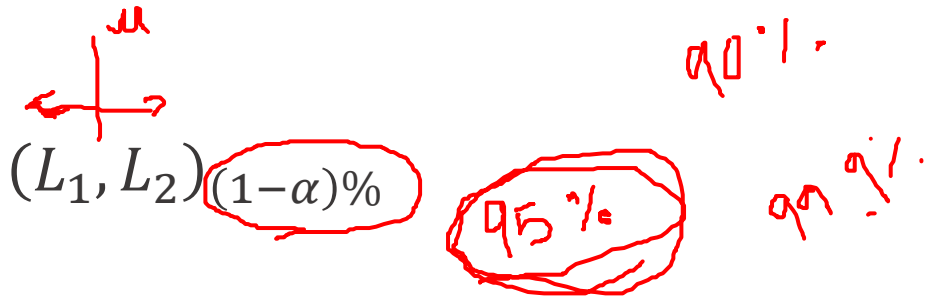
Varianza

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Comparación de varianzas

$$F = \frac{\widetilde{s_1}^2}{\widetilde{s_2}^2}$$

Intervalos de confianza



La construcción depende de:

- Un nivel de confianza definido.
- Una distribution de referencia.
- El tamaño de la muestra.
- La variabilidad.

Intervalos de confianza

Varianza

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right]$$

Media

$$\bar{X} \pm t_{n-1,1-\alpha/2}\, S/\sqrt{n}$$

Comparación de medias

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2}\, S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

Para tener en cuenta al construir intervalos

- Varianzas desiguales.
- Pocos datos.
- Muchos datos
- Requiere que ambas distribuciones sean normales.
- Datos pareados.

Pruebas de hipótesis

Pruebas de hipótesis

Hipótesis: Se refiere a una prueba formal para tomar decisiones usando datos.

- **Hipótesis nula Ho**. Afirmación acerca del valor de un parámetro poblacional que se considera válida para desarrollar el procedimiento de prueba.

- **Hipótesis Alternativa Ha.-** Afirmación que se aceptará si los datos muéstrales proporcionan evidencia de que la hipótesis nula es falsa

95%

$$H_0: \mu = 20$$
$$H_1: \mu \neq 20$$

$$H_0: \sigma = 8$$
$$H_1: \sigma \neq 8$$

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

$$H_0: \mu \geq 20$$
$$H_1: \mu < 20$$

Nivel sig $\Rightarrow 1 - 95\% = 0{,}05$

Pruebas de hipótesis

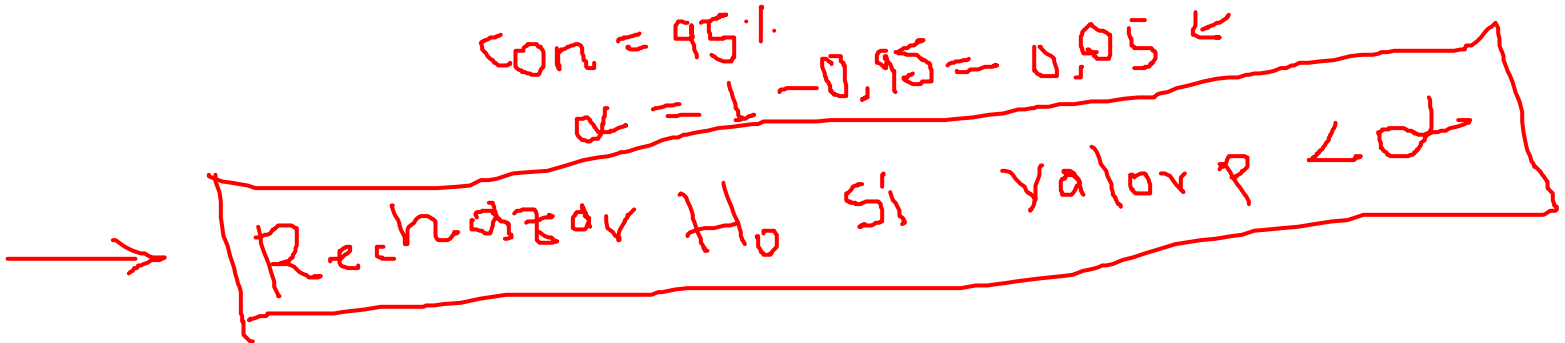
- La hipótesis alternativa normalmente se representa de la forma <, > ó ≠.

Decisiones Posibles	Situaciones Posibles	
	La hipótesis nula es verdadera	La hipótesis nula es falsa
Aceptar la Hipótesis Nula	Se acepta correctamente	Error tipo II
Rechazar la Hipótesis Nula	Error tipo I	Se rechaza correctamente

Nivel de significancia

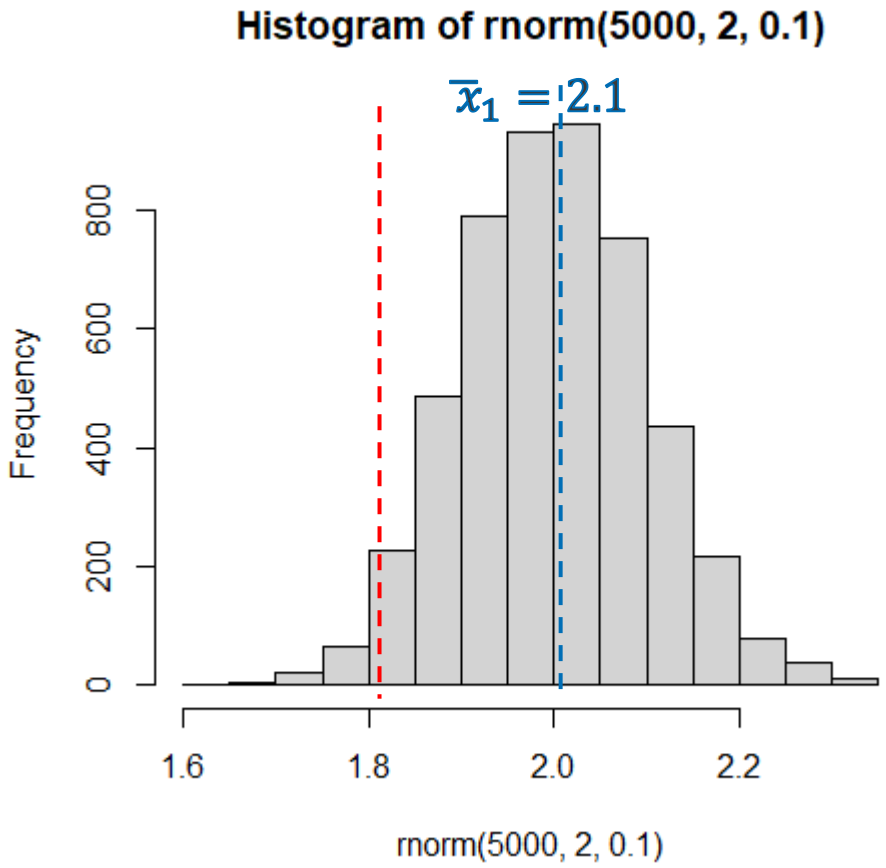
Valor p

- Los valores p evalúan qué tan bien los datos de la muestra apoyan el argumento del investigador de que la hipótesis nula es verdadera. Mide qué tan compatibles son los datos con la hipótesis nula.
  - Valores p altos: los datos son probables con una hipótesis nula verdadera.
  - Valores p bajos: los datos son poco probables con una hipótesis nula verdadera.
- Un valor p bajo sugiere que la muestra provee suficiente evidencia de que se puede rechazar la hipótesis nula para toda la población.
- Se rechaza en base a un nivel de significancia, generalmente entre



Máquina de producir puntilla de dos pulgadas, ¿necesito calibrarla?

¿Qué tan probable es el efecto observado en los datos de la muestra si la hipótesis nula es verdadera?



Ho: El promedio de las puntillas es dos pulgadas $\mu = 2$

Ha: El promedio de las puntillas no es dos pulgadas $\mu \neq 2$

$n = 10$

Caso 1. $\bar{x}_1 = 2.1$ → 0.9

Caso 2. $\bar{x}_2 = 1.65$ → 0.005

Ejemplo pruebas de hipótesis

El gerente de una fábrica de tuberías desea determinar si el diámetro promedio de los tubos es diferente de 5 cm. El gerente sigue los pasos básicos para realizar una prueba de hipótesis.

1. Especificar las hipótesis.

   En primer lugar, el gerente formula las hipótesis. La hipótesis nula es: la media de la población de todos los tubos es igual a 5 cm. Formalmente, esto se escribe como: H_0: $\mu = 5$

   Luego, el gerente elige entre las siguientes hipótesis alternativas:

Condición que se probará	Hipótesis alternativa
La media de la población es menor que el objetivo.	unilateral: $\mu < 5$
La media de la población es mayor que el objetivo.	unilateral: $\mu > 5$
La media de la población es diferente del objetivo.	bilateral: $\mu \neq 5$

   Como tiene que asegurarse de que los tubos no sean más grandes ni más pequeños de 5 cm, el gerente elige la hipótesis alternativa bilateral, que indica que la media de la población de todos los tubos no es igual a 5 cm. Formalmente, esto se escribe como H_1: $\mu \neq 5$

https://support.minitab.com/es-mx/minitab/20/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/example-of-a-hypothesis-test/

Ejemplo pruebas de hipótesis

2. Elegir un nivel de significancia (también denominado alfa o α).

   El gerente selecciona un nivel de significancia de 0.05, que es el nivel de significancia más utilizado.

3. Determinar la potencia y el tamaño de la muestra para la prueba.

   El gerente utiliza un cálculo de potencia y tamaño de la muestra para determinar cuántos tubos tiene que medir para tener una buena probabilidad de detectar una diferencia de 0.1 cm o más con respecto al diámetro objetivo.

4. Recolectar los datos.

   Recoge una muestra de tubos y mide los diámetros.

5. Comparar el valor p de la prueba con el nivel de significancia.

   Después de realizar la prueba de hipótesis, el gerente obtiene un valor p de 0.004. El valor p es menor que el nivel de significancia de 0.05.

6. Decidir si rechazar o no rechazar la hipótesis nula.

   El gerente rechaza la hipótesis nula y concluye que el diámetro medio de todos los tubos no es igual a 5 cm.

Pruebas estadísticas no paramétricas

Cómo decidir si optamos por un método no paramétrico

- Los datos de la población no tienen una distribución normal.
- La media y la desviación estándar no son confiables de la población.
- Las relaciones no son lineales.

Rankiar datos y agregar signos

Raw Data	Sorted Data	Ranked Data	Ranked Data
15	8	1	1
8	10	2	2 → 3
27	10	3	3 → 3
25	10	4	4 → 3
10	12	5	5
23	14	6	6
12	15	7	7 → 7.5
18	15	8	8 → 7.5
14	18	9	9
10	23	10	10
15	25	11	11
10	27	12	12

#	Set 1	Set 2	Set 1 – Set 2	Sign
1	443	57	386	+
2	421	352	69	+
3	436	587	-151	-
4	376	415	-39	-
5	458	458	0	NA
6	408	424	-16	-
7	422	463	-41	-
8	431	583	-152	-
9	459	432	27	+
10	369	379	-10	-
11	360	370	-10	-
12	431	584	-153	-
13	403	422	-19	-
14	436	587	-151	-
15	376	415	-39	-
16	370	419	-49	-
17	443	57	386	+

Gracias!

Hugo Andrés Dorado.

Científico de datos

hugo.doradob@gmail.com

Conocimiento generado a partir de proyectos de:

Alianza

Bioversity International

CIAT
Centro Internacional de Agricultura Tropical
Desde 1967 Ciencia para cultivar el cambio

CLIMATE CHANGE AGRICULTURE AND FOOD SECURITY