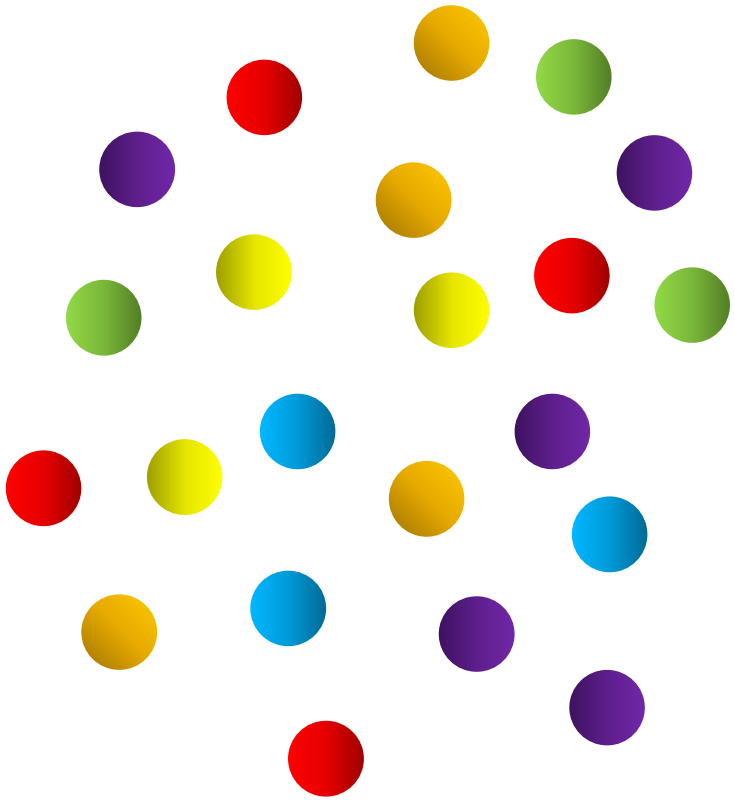
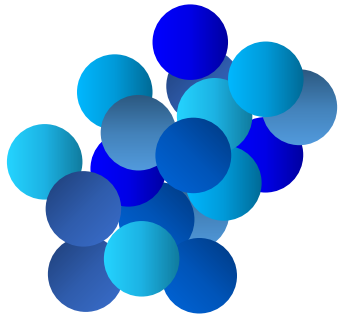
Diplomado en Ciencia de Datos
Fundamentos de Inferencia

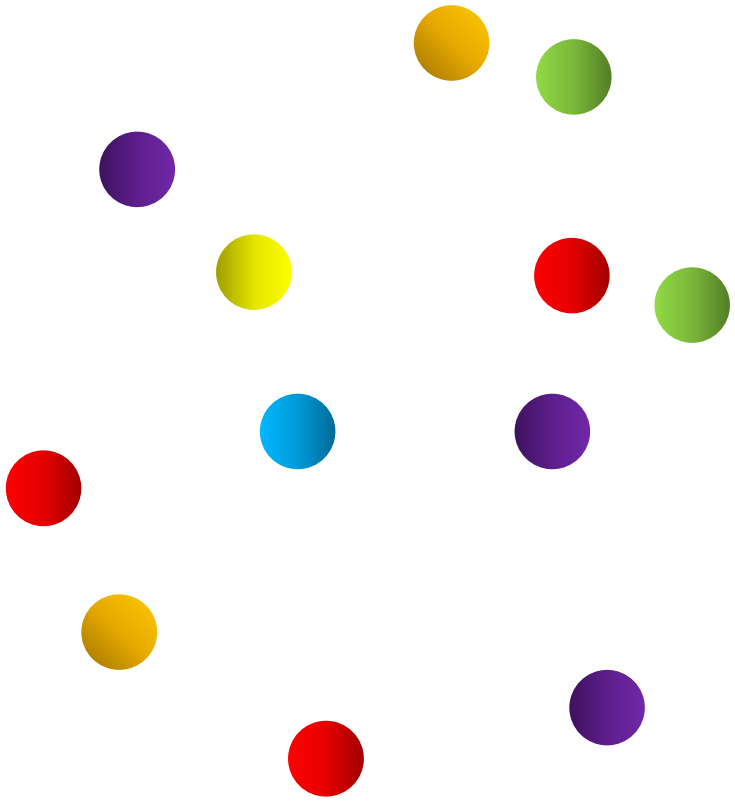
Docente: Dora Suárez

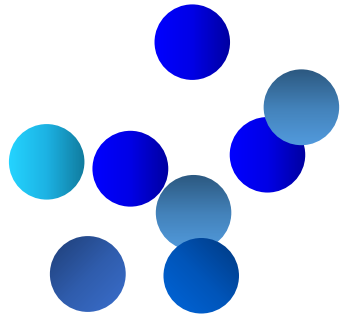
Población





Muestra

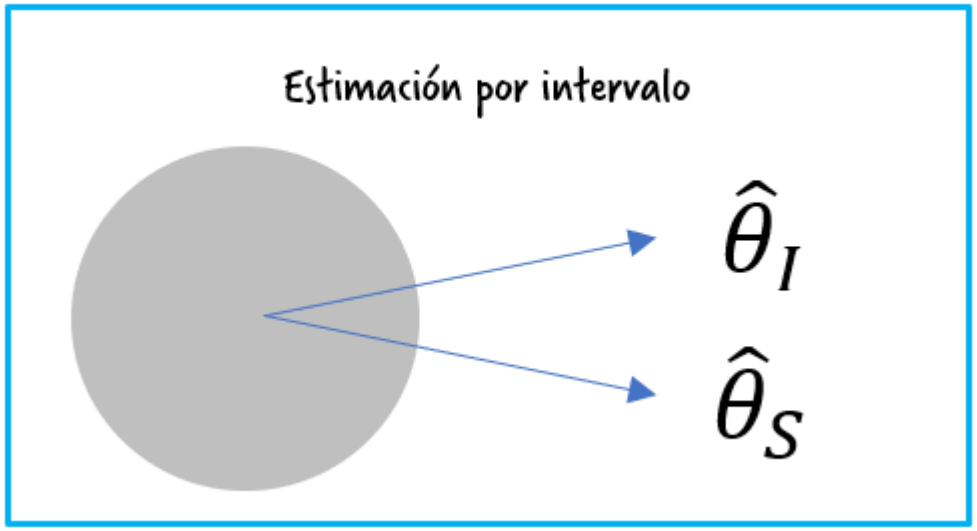




Estimador de intervalo

Un **estimador de intervalo** es una regla que especifica el método para calcular los números que hacen parte de los puntos extremos del intervalo.

Los extremos se denominan intervalo inferior y superior.



Coeficiente de confianza

La probabilidad de que un intervalo aleatorio contenga el parámetro se denomina coeficiente de confianza.

$$P\left(\hat{\theta}_I \leq \theta \leq \hat{\theta}_S\right) = 1 - \alpha$$

$\hat{\theta}_I$: Límite inferior

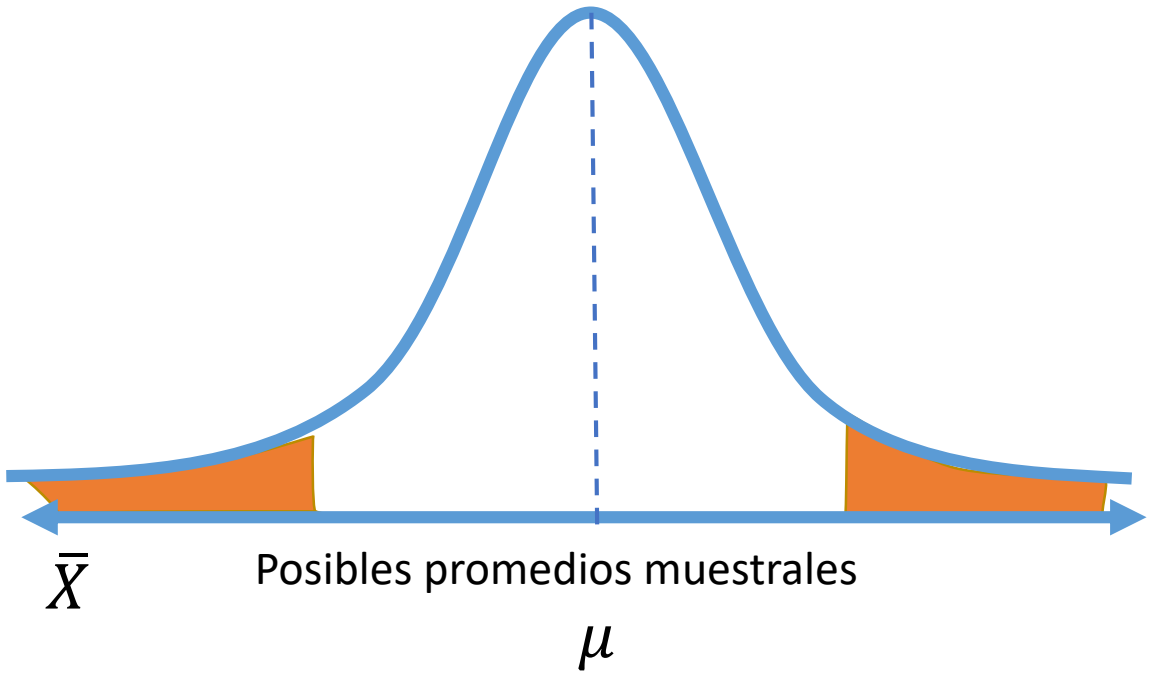
$\hat{\theta}_S$: Límite superior

$1 - \alpha$: Coeficiente de confianza

Intervalos de confianza

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Estimador

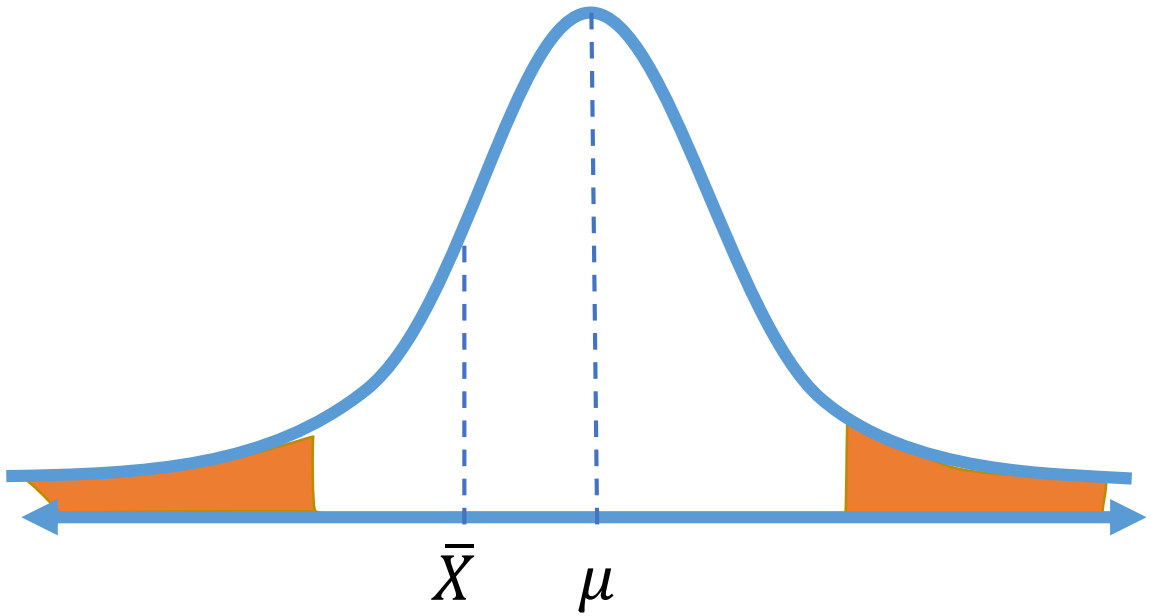


Intervalos de confianza

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Estimador

$$\frac{\bar{X} - \mu}{\sqrt{n}} \sim N(0,1)$$

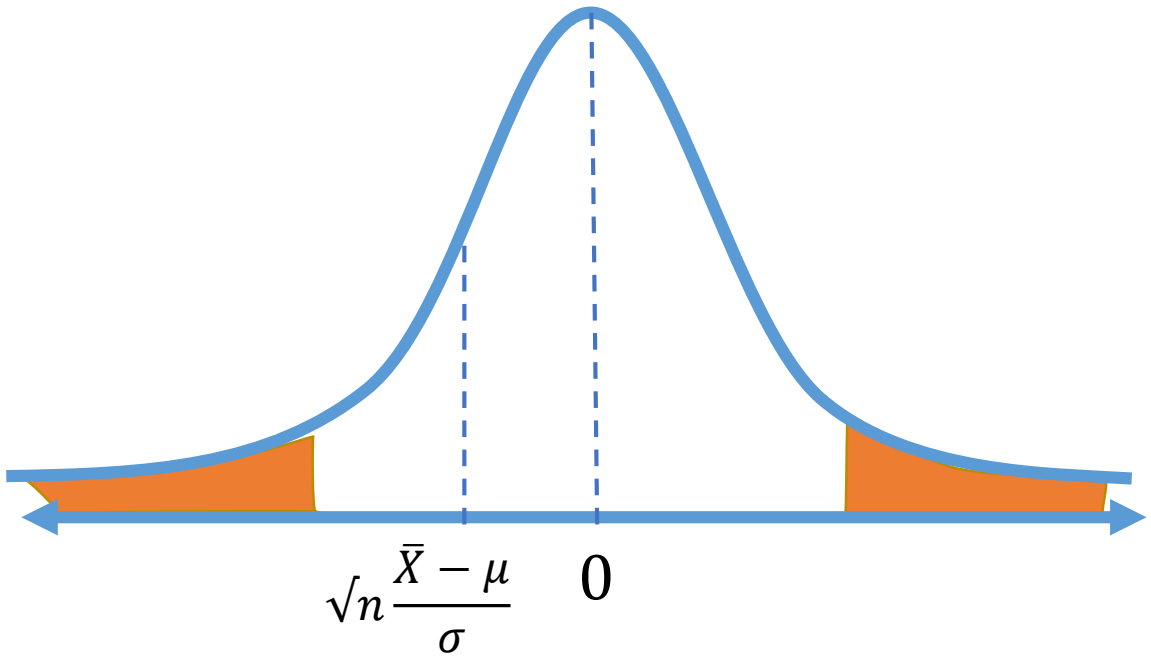


Intervalos de confianza

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

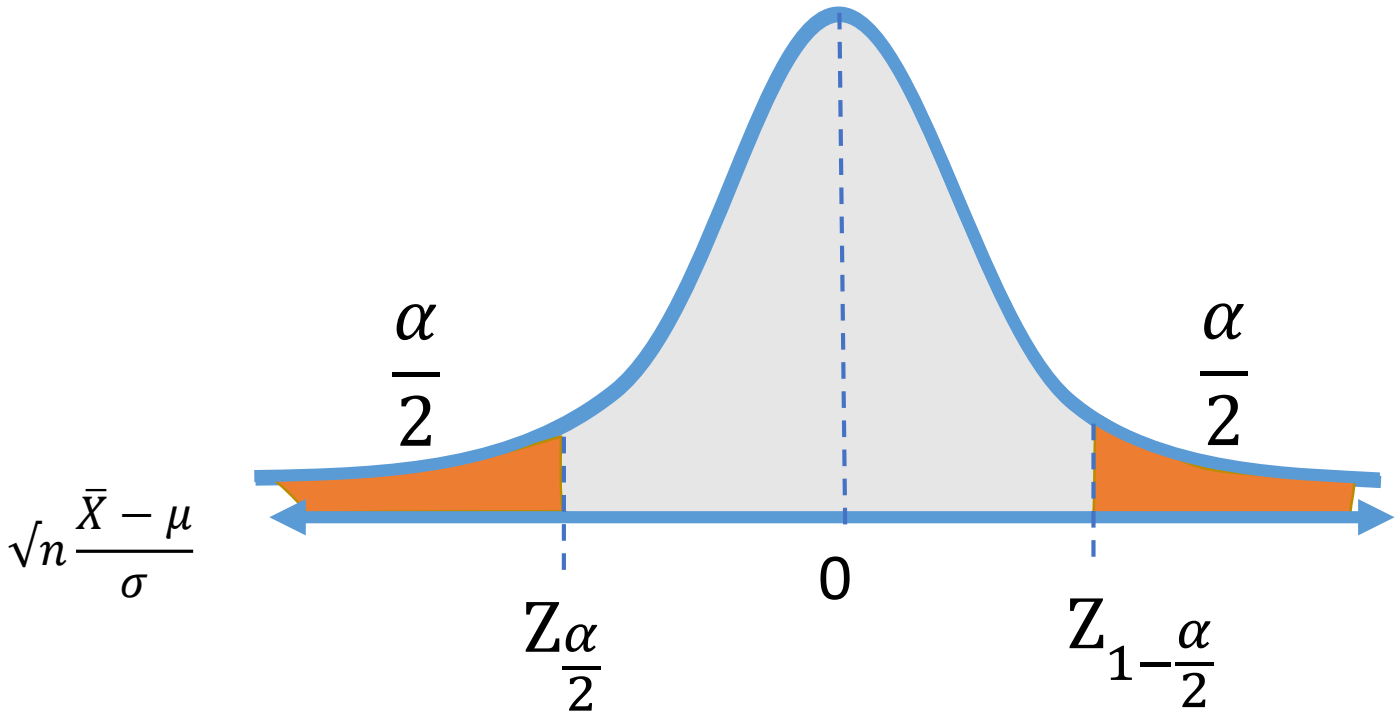
Estimador

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$$



Intervalos de confianza

$$P\left(Z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$



Confianza: $1-\alpha$

Intervalos de confianza

$$P\left(Z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Intervalos de confianza

$$P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

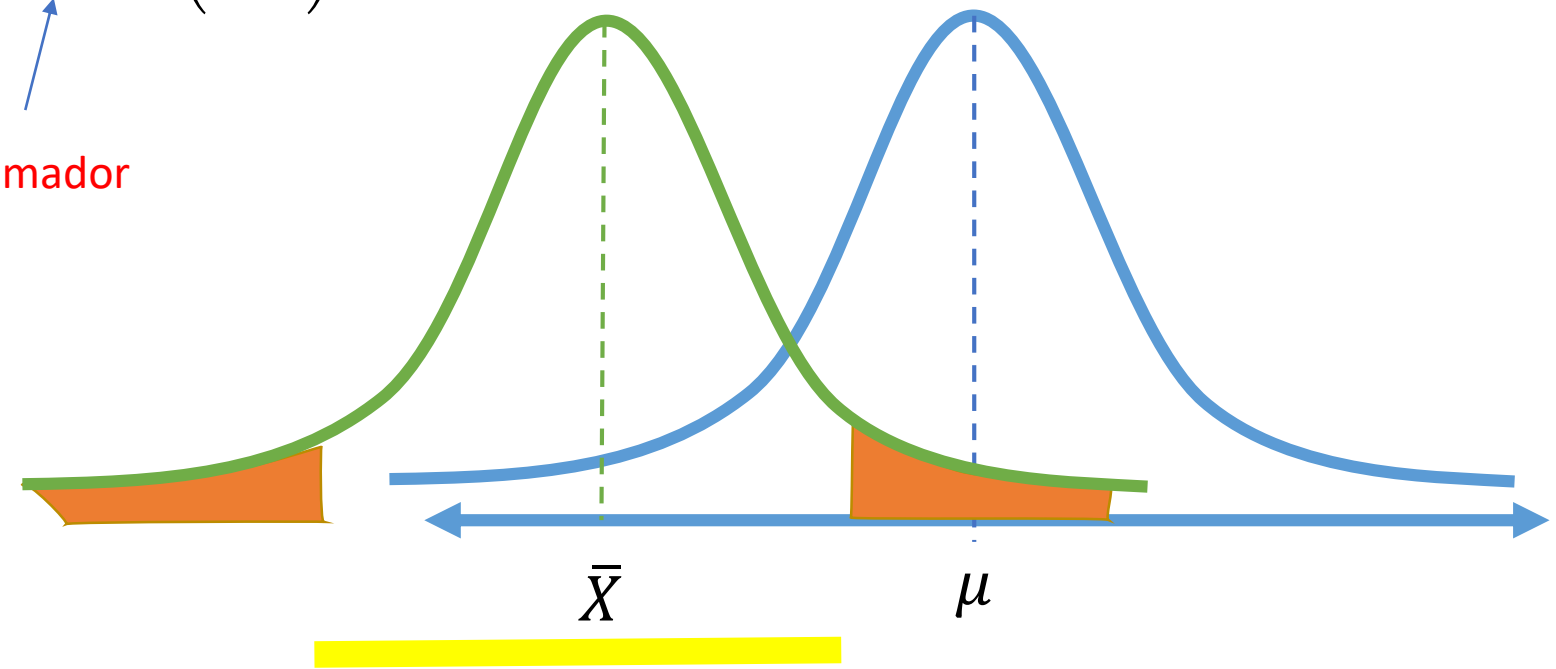
$$P\left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$IC(\mu) = \left(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \;\; ; \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Intervalos de confianza

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

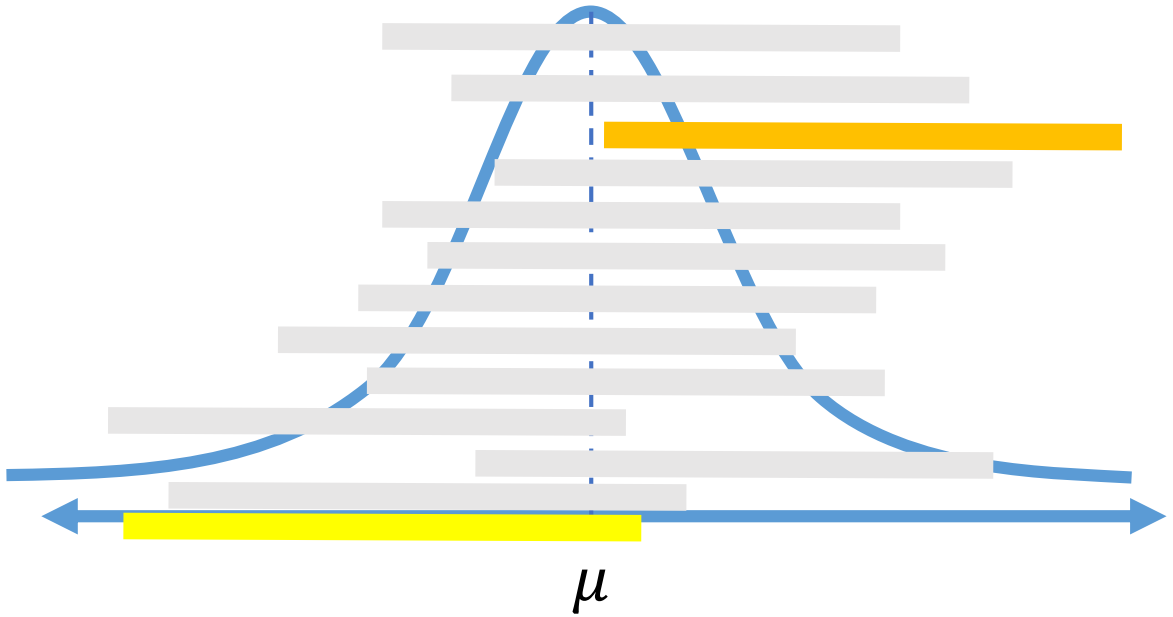
Estimador



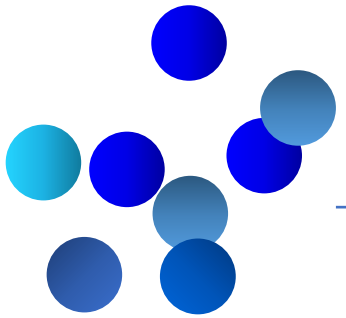
Intervalos de confianza

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

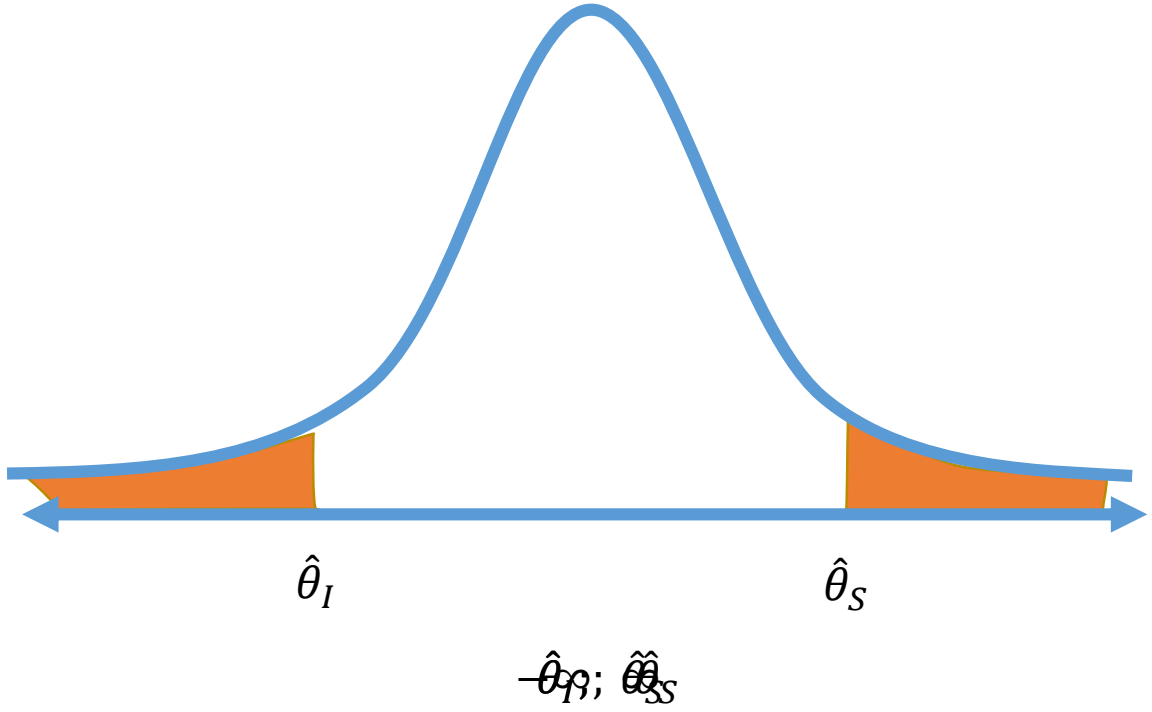
Estimador



Intervalos de confianza



$\hat{\theta}_I \ ; \ \hat{\theta}_S$



$\hat{\theta}_I ; \hat{\theta}_S$

¿Qué es una prueba de hipótesis?

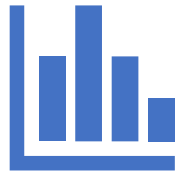
Una prueba de hipótesis es un procedimiento a través del cual es posible especificar el hecho de aceptar o rechazar una afirmación. En general estas afirmaciones se realizan sobre la población y la elección entre aceptar o rechazar la afirmación es hecha con base en la muestra de datos.

El método científico

Definir una pregunta de interés

Plantear una hipótesis acerca de una posible respuesta a la pregunta planteada

Experimentar y recolectar datos



Analizar los datos y con base en ellos ver si la hipótesis nula se contradice o no

Establecer conclusiones y de ser posible dar respuesta a la pregunta de investigación

Pruebas de hipótesis y el método científico

- Plantear una hipótesis nula o una pregunta de interés
- Tomar una muestra representativa que refleje el comportamiento de la población frente a la pregunta de interés
- Si lo que se observa en los datos de la muestra contradice la hipótesis que fue planteada se rechaza la hipótesis
- Si lo que se observa en los datos no contradice la hipótesis que fue planteada no se rechaza la hipótesis

Un ejemplo más

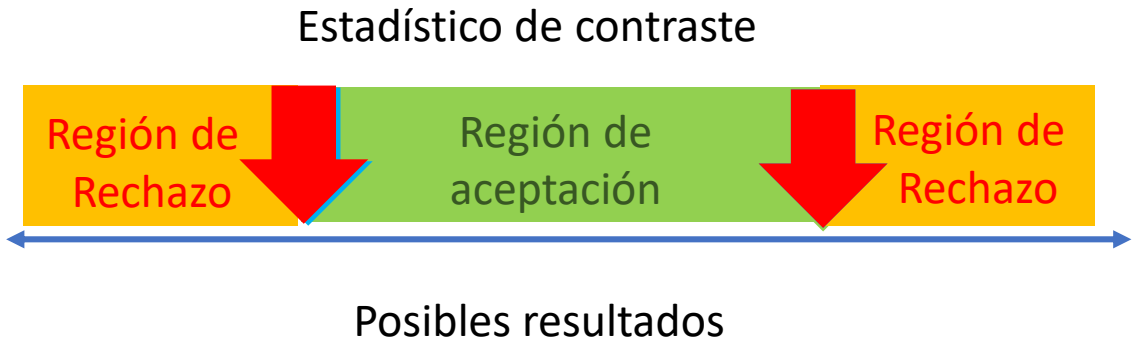
15 veces

$H_0: p = 0.5$
$H_a: p \neq 0.5$

10 caras

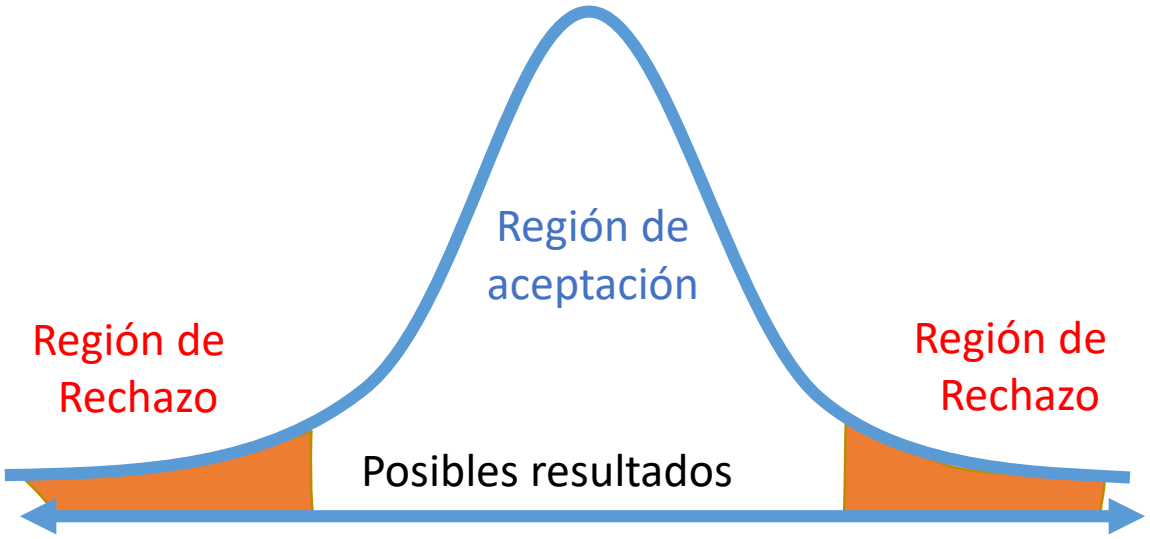
Estadístico de prueba

El estadístico de prueba es, al igual que los estimadores, una función de las mediciones muestrales en las que la decisión de la prueba estará basada. (Cumple la labor de ser una cantidad de referencia)



Región de rechazo

El estadístico de prueba es, al igual que los estimadores, una función de las mediciones muestrales en las que la decisión de la prueba estará basada. (Cumple la labor de ser una cantidad de referencia)

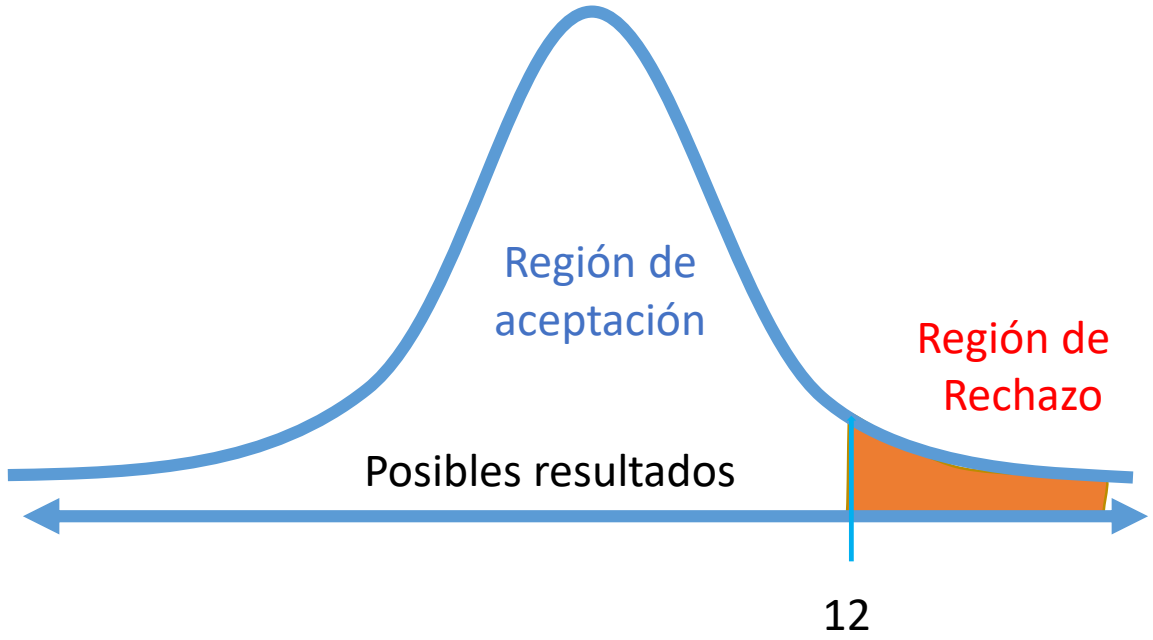


Región de rechazo

Si el valor del estadístico de prueba cae en la región de rechazo, se rechaza la hipótesis nula

Región de rechazo - ejemplo

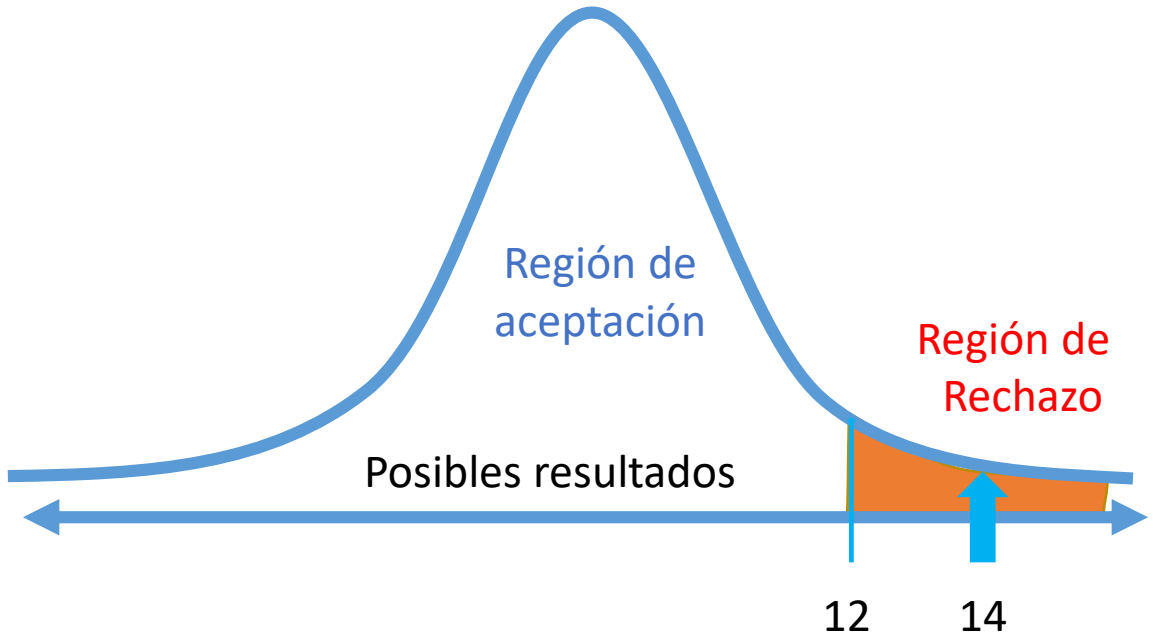
Nuevamente en el lanzamiento de las monedas, supongamos que si en mas de 12 ocasiones la moneda cae cara, diremos que esta está cargada.



Región de rechazo - ejemplo

Como se obtuvieron 14 caras entonces el estadístico de prueba sería 14.

Por lo cual, **se rechaza la hipótesis nula** de que la moneda está equilibrada a favor de que está cargada a la ocurrencia de caras.



Elementos de una prueba de hipótesis

1. Hipótesis nula Ho
2. Hipótesis alternativa Ha
3. Estadístico de prueba
4. Región de rechazo

Error tipo 1 y tipo 2

Se comete un error tipo I si **Ho es rechazada** cuando **Ho es verdadera**.

Se comete un error tipo II si **Ho es aceptada** cuando **Ho es falsa.**

Error tipo 1 y tipo 2

Se comete un error tipo I si **Ho es rechazada** cuando **Ho es verdadera**.

$$P(\text{Error tipo I}) = \alpha$$

Se establece a priori α como nivel de significancia o error máximo aceptable para la conclusión.

La robustez de un método estadístico es una determinada situación se calcula como (1−β), lo que corresponde con la situación de haber rechazado correctamente H0 ya que esta era falsa.

Se comete un error tipo II si **Ho es aceptada** cuando **Ho es falsa.**

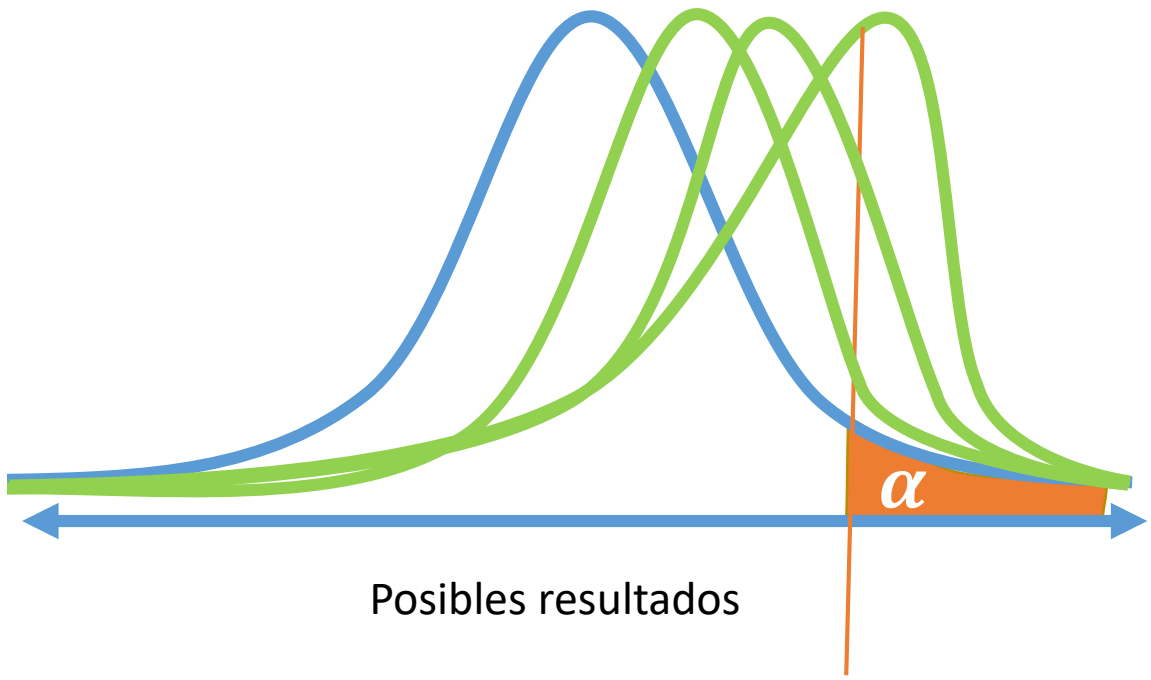
$$P(\text{Error tipo II}) = \beta$$

Error tipo 1 y tipo 2 – Ejemplo 1

Se comete un error tipo I si **Ho es rechazada** cuando **Ho es verdadera**.

Se comete un error tipo II si **Ho es aceptada** cuando **Ho es falsa.**

Error tipo I y tipo II



Error tipo 1 y tipo 2

Hipótesis Nula	Aceptamos	Rechazamos
Verdadera	CORRECTO	Error tipo I α
Falsa	Error tipo II β	CORRECTO

Ejemplo 1 – Media

Hipótesis	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu \neq \mu_0 \end{cases}$	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	$\begin{cases} Ho: \mu = \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$
Región de rechazo	$\{\lvert Z\rvert > z_{\frac{\alpha}{2}}\}$	$\{Z > z_{1-\alpha}\}$	$\{Z < z_{\alpha}\}$

Estadístico de Prueba: $\boldsymbol{Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}$

P-valor

El p-valor es la probabilidad de obtener un resultado más extremo o igual que el del estadístico de contraste bajo la hipótesis nula

SI el p-valor es menor que el nivel de significancia, se rechaza la hipótesis nula

Casos comunes

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Media Poblacional	$\mu = \mu_0$	$\sqrt{n}\frac{(\bar{X} - \mu_0)}{\sigma}$	Z	- La varianza es conocida - El tamaño de muestra es mayor que 30 o los datos se distribuyen normalmente
Media Poblacional	$\mu = \mu_0$	$\sqrt{n}\frac{(\bar{X} - \mu_0)}{s}$	$t - student$ n-1 grados de libertad	- La varianza es desconocida - El tamaño de muestra es menor que 30
Proporción Poblacional	$p = p_0$	$\sqrt{n}\frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}$	Z	$n\hat{p}\ y\ n(1-\hat{p}) \geq 10$

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z	- Las varianzas son conocidas - Los tamaños de muestra son mayores que 30
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$t - student$ Con $\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ grados de libertad	- Las varianzas son desconocidas y diferentes - Los tamaños de muestra son menores que 30
Diferencia de dos medias	$\mu_1 - \mu_2 = d$	$\frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 1}}}$	$t - student$ Con n-2 grados de libertad	- Las varianzas son desconocidas e iguales - Los tamaños de muestra son menores que 30
Diferencia de medias pareadas	$\mu_d = d$	$\sqrt{n}\frac{(\bar{D} - d)}{s_d}$	$t - student$ Con n-1 grados de libertad	- Muestras pareadas

Otras Pruebas de Hipótesis

Parámetro – Caso	Hipótesis Nula	Comando R	¿Cuándo se usa?
Coeficiente de correlación de Pearson	$\rho = 0$	`cor.test(x, y, method = “pearson")`	- Las dos variables son continuas
Coeficiente de correlación de Spearman	$\rho = 0$	`cor.test(x, y, method = "spearman")`	- Una variable continua y una categorica
Independencia de variables categóricas	$\rho = 0$	`chisq.test(table(x,y))`	- Las dos variables son cuantitativas

Parámetro – Caso	Hipótesis Nula	Estadístico de Prueba	Distribución	¿Cuándo se usa?
Diferencia de dos proporciones	$p_1 - p_2 = d$	$\frac{(\hat{p}_1 - \hat{p}_2) - d}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}}$	Z	$n\hat{p}\ y\ n(1-\hat{p}) \geq 10$
Varianza poblacional	$\sigma = \sigma_0$	$\frac{(n-1)s^2}{\sigma_0^2}$	$Chi\ cuadrado\ (\chi^2)$ n-1 grados de libertad	- Las observaciones se distribuyen de forma normal
Diferencia de varianzas	$\sigma_1^2 = \sigma_2^2$	$\frac{s_1^2}{s_2^2}$	$f - fisher$ n-1 grados de libertad en el numerador y n-1 grados de libertad en el denominador	- Las dos poblaciones son normales e independientes

Importancia de la normalidad de los datos

Aunque la media de un conjunto de datos, para tamaños de muestra grande siempre sigue una distribución normal, es necesario probar en los datos si se comportan acorde a una distribución normal.

Esto con el fin de conocer si las variables medidas en la muestra pueden ser descritas con parámetros de tendencia central y dispersión alrededor de dichos parámetros

Pruebas de normalidad de una muestra

1. Hacer un histograma y un boxplot de los datos
2. Aplicar una prueba formal como las de *Shapiro-Wilk* y de *Kolmogorov-Smirnov*

Ho : Los datos siguen una distribución normal

Ha : Los datos NO siguen una distribución normal

3. Hacer el grafico de la distribución acumulada empírica de los datos vs los de una distribución normal (qplot)

Pruebas de normalidad de una muestra

```
shapiro.test(datos)

ks.test(datos)

qqnorm(datos)

qqline(datos)
```

¿Cómo comparar muestras no normales?

Pruebas no paramétricas

Test de Kruskal-Wallis: Se trata de un test que emplea rangos para contrastar la hipótesis de que k muestras han sido obtenidas de una misma población.

Condiciones

- No es necesario que las muestras que se comparan provengan de una distribución normal.

- Homocedasticidad: dado que la hipótesis nula asume que todos los grupos pertenecen a una misma población y que por lo tanto tienen las mismas medianas, es requisito necesario que todos los grupos tengan la misma varianza. Se puede comprobar con representaciones gráficas o con los test de Levenne o Barttlet.

- Misma distribución para todos los grupos: la distribución de los grupos no tiene que ser normal pero ha de ser igual en todos (por ejemplo que todos muestren asimetría hacia la derecha).

En R

```
# Para probar homogeneidad
leveneTest(variable ~ grupos, data = datos,
center = "median")

# Para probar diferencias
kruskal.test(variable ~ grupos, data = datos)
```