

The background features a complex geometric design with overlapping triangles in shades of blue, grey, and yellow. A central white diamond shape contains the main title. In the upper right, a rectangular inset shows a blurred image of server racks with glowing blue and red lights, overlaid with a network diagram and binary code (0s and 1s).

Análisis Inteligente de Datos

**Maestría en Exploración de Datos y
Descubrimiento del Conocimiento**

Profesora: Mónica Cantoni

**Ayudantes: Cecilia Oliva - Fabiana Rossi - Pamela
Pairo**

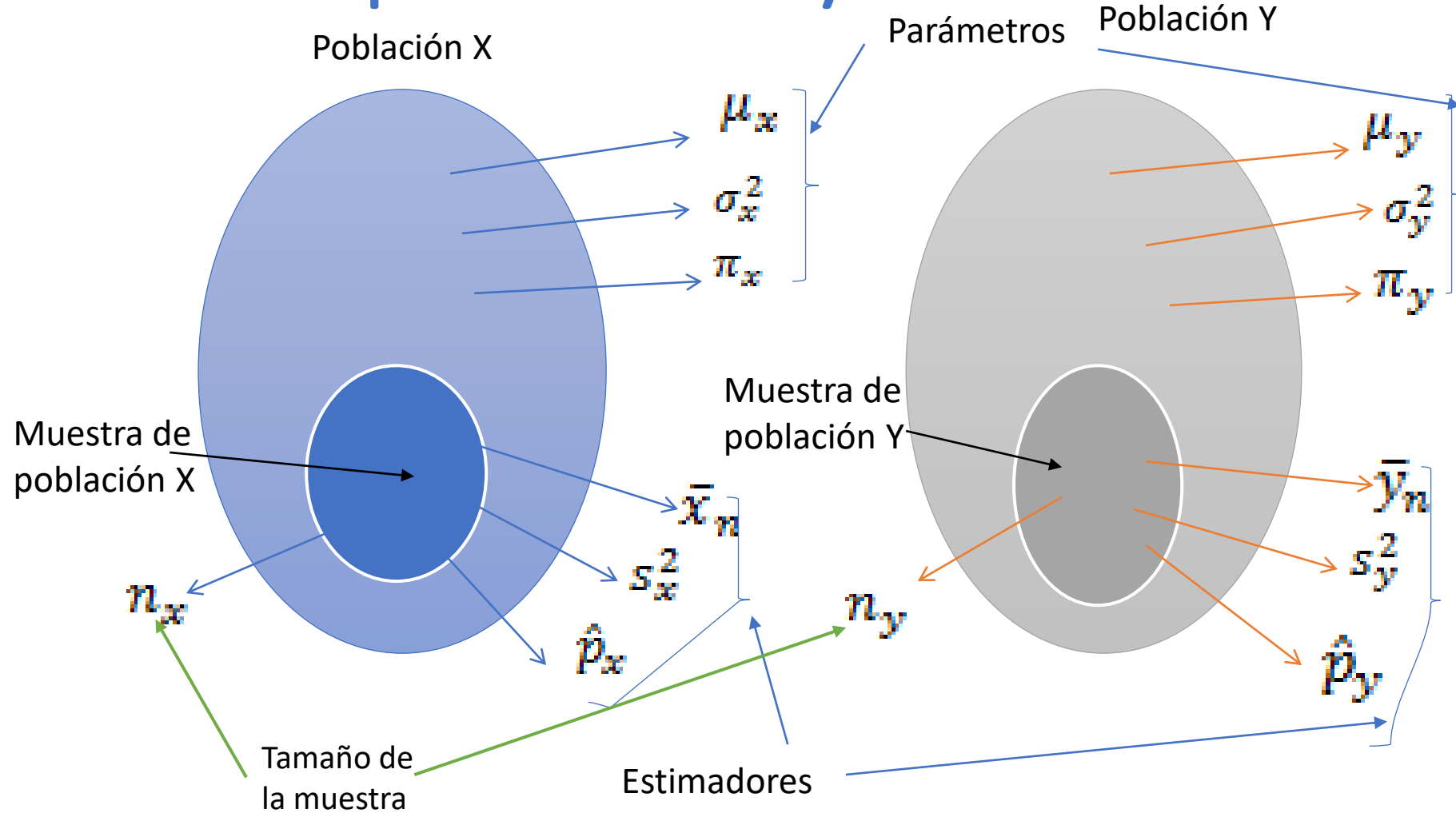
Clase 04 – Comparación de poblaciones

Agenda

- **Introducción**
- Comparación de dos poblaciones
- Métodos paramétricos
- Métodos no paramétricos
- Comparación de medias de más de dos poblaciones normales

Comparación de dos poblaciones.

Notación de parámetros y estimadores



Diferencia de medias de poblaciones normales para dos muestras independientes

Parámetro $(\mu_x - \mu_y)$ Estimador $(\bar{x} - \bar{y})$

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n_x}}\right) \quad \bar{y} \sim N\left(\mu_y, \frac{\sigma_y}{\sqrt{n_y}}\right)$$

$$\bar{x} - \bar{y} \sim N\left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$$



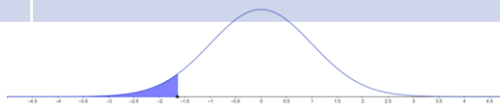
Intervalos de Confianza. Muestras independientes tomadas de poblaciones normales con varianzas poblacionales conocidas. Se usa el estadístico Z. Nivel de Confianza: $1 - \alpha$

$$P \left[(\bar{x} - \bar{y}) - Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right] = (1 - \alpha)$$

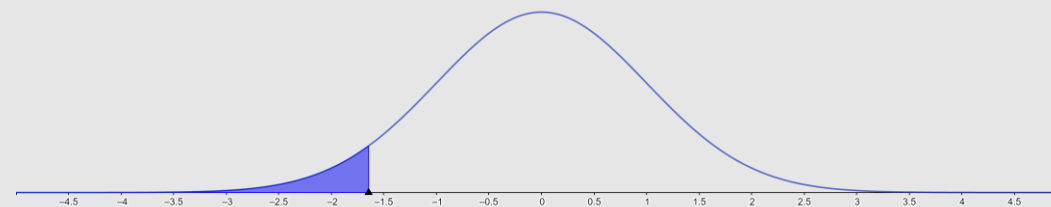
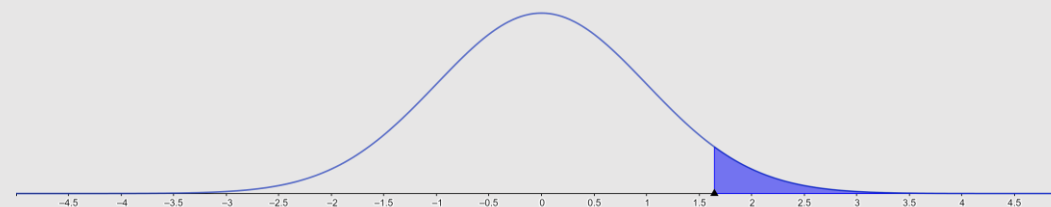
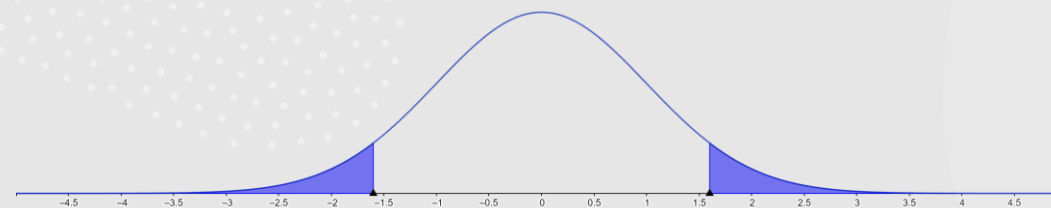
$$IC_{\mu_x - \mu_y} \left[(\bar{x} - \bar{y}) - Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}; (\bar{x} - \bar{y}) + Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right] = (1 - \alpha)\%$$

$$IC_{\mu_x - \mu_y} \left[(\bar{x} - \bar{y}) \mp Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right] = (1 - \alpha)\%$$

Test de Hipótesis. Muestras independientes tomadas de poblaciones normales con varianzas poblacionales conocidas. Se usa el estadístico Z. Nivel de Significación: α

Hipótesis	Estadístico	Criterio de rechazo	
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y \neq 0 \end{cases}$	$Z_e = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)}}$ $(\mu_x - \mu_y) = 0$	Rechazar H_0 si $Z_e < -Z_{1-\frac{\alpha}{2}}$ O Rechazar H_0 si $Z_e > Z_{1-\frac{\alpha}{2}}$	
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y > 0 \end{cases}$		Rechazar H_0 si $Z_e > Z_{1-\alpha}$	
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y < 0 \end{cases}$		Rechazar H_0 si $Z_e < -Z_{1-\alpha}$	

Zonas de rechazo usando Normal



Muestras independientes y grandes, varianzas poblacionales desconocidas

Parámetro $(\mu_x - \mu_y)$ Estimador $(\bar{x} - \bar{y})$

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim N(0,1)$$

Intervalos de Confianza muestras independientes y grandes, varianzas poblacionales desconocidas

$$IC_{\mu_x - \mu_y} \left[(\bar{x} - \bar{y}) - Z_{(1-\alpha/2)} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}; (\bar{x} - \bar{y}) + Z_{(1-\alpha/2)} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \right] = (1 - \alpha)\%$$

Muestras independientes y grandes, varianzas poblacionales desconocidas

Población Normal o desconocida, muestras independientes y grandes ($n_x > 30$ y $n_y > 30$) con varianzas poblacionales desconocidas. Se usa el estadístico Z con un nivel de significación α .

Hipótesis	Estadístico	Criterio de rechazo
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y \neq 0 \end{cases}$	$Z_e = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}}$ $(\mu_x - \mu_y) = 0$	Rechazar H_0 si $Z_e < -Z_{1-\frac{\alpha}{2}}$ o Rechazar H_0 si $Z_e > Z_{1-\frac{\alpha}{2}}$
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y > 0 \end{cases}$		Rechazar H_0 si $Z_e > Z_{1-\alpha}$
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y < 0 \end{cases}$		Rechazar H_0 si $Z_e < -Z_{1-\alpha}$

Muestras normales independientes con varianzas desconocidas pero se suponen que son iguales.

Parámetro $(\mu_x - \mu_y)$ Estimador $(\bar{x} - \bar{y})$

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim T(n_x + n_y - 2)$$

Intervalo de Confianza Muestras normales independientes con varianzas desconocidas pero se suponen que son iguales.

$$P \left[(\bar{x} - \bar{y}) - T_{\left(1-\frac{\alpha}{2}, n_x+n_y-2\right)} \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} < (\mu_x - \mu_y) < (\bar{x} - \bar{y}) + T_{\left(1-\frac{\alpha}{2}, n_x+n_y-2\right)} \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right] = 1 - \alpha$$

$$IC_{(\mu_x - \mu_y)} \left[(\bar{x} - \bar{y}) - T_{\left(1-\frac{\alpha}{2}, n_x+n_y-2\right)} \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}; (\bar{x} - \bar{y}) + T_{\left(1-\frac{\alpha}{2}, n_x+n_y-2\right)} \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right] = (1 - \alpha)\%$$

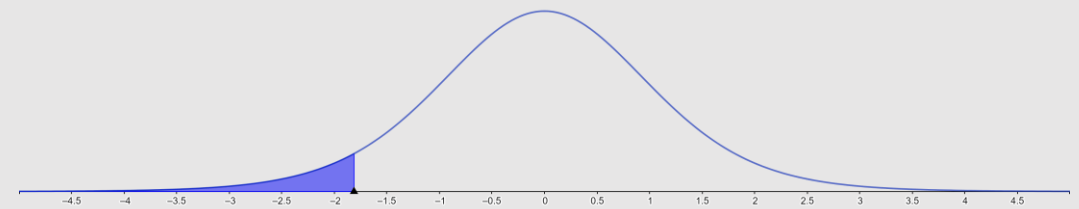
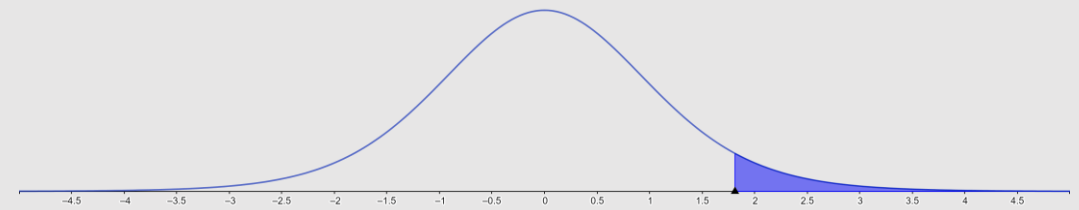
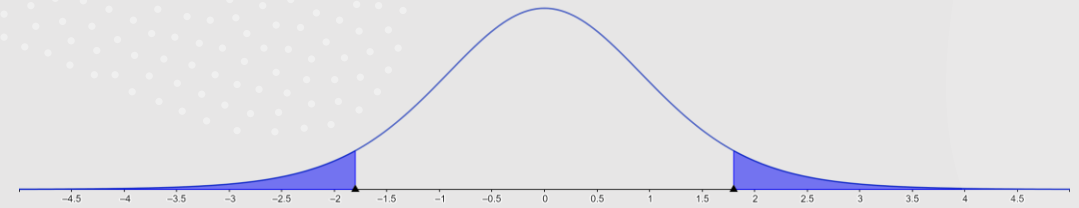
$$IC_{(\mu_x - \mu_y)} \left[(\bar{x} - \bar{y}) \mp T_{\left(1-\frac{\alpha}{2}, n_x+n_y-2\right)} \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right] = (1 - \alpha)\%$$

Muestras normales independientes con varianzas desconocidas pero se suponen que son iguales.

Población Normal, varianzas poblacionales desconocidas pero iguales, muestras independientes y pequeñas (n_x y/o $n_y \leq 30$). Se usa el estadístico t con $n_x + n_y - 2$ grados de libertad y un nivel de significación α .

Hipótesis	Estadístico	Criterio de rechazo
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y \neq 0 \end{cases}$	$t_e = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\left(\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}\right)} \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$	<p>Rechazar H_0 si $t_e < -t_{(1-\frac{\alpha}{2}; n_x + n_y - 2)}$</p> <p>O Rechazar H_0 si $t_e > t_{(1-\frac{\alpha}{2}; n_x + n_y - 2)}$</p>
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y > 0 \end{cases}$		<p>Rechazar H_0</p> <p>si $t_e > t_{(1-\alpha; n_x + n_y - 2)}$</p>
$\begin{cases} H_0: \mu_x - \mu_y = 0 \\ H_1: \mu_x - \mu_y < 0 \end{cases}$		<p>Rechazar H_0</p> <p>si $t_e < -t_{(1-\alpha; n_x + n_y - 2)}$</p>

Zonas de rechazo usando t de Student



Caso 4. Intervalo de Confianza para la diferencia de medias, si las muestras están relacionadas – (muestras dependientes o datos pareados o emparejadas). Se usa T de Student con $n-1$ grados de libertad.

- Consideramos que las muestras son dependientes si en los valores de una de las muestras influyen los de la otra. Los miembros de la muestra se eligen por pares, uno de cada población.
- Supongamos que tenemos una muestra de n pares de observaciones enlazadas procedentes de distribuciones normales.
- Al considerarse que las variables son dependientes, se genera una nueva variable denominada diferencia (d). Se calcula la diferencia para cada par de los valores de la variable x e y , entonces $d = x - y$. Debido a que por propiedades de la varianza de dos variables, no puede aplicarse la varianza de la suma de dos variables es igual a la suma de sus varianzas ya que son dependientes y entraría en juego la covarianza de x e y , se realiza este cambio de variable.
- Las diferencias obtenidas se consideran con el signo obtenido, sean $(-, +)$.
- Se supone que la distribución de las diferencias es normal.
- Se utiliza el estadístico T de Student con $n-1$ grados de libertad, siendo n la cantidad de pares.

Muestras pareadas

Se genera una nueva variable $d = x - y$

Se calcula la media y la desviación estándar de la variable d .

Se obtienen los estimadores: \bar{d} y S_D de la siguiente manera o utilizando la calculadora.

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

$$S_D = \sqrt{\frac{\sum_{i=1}^n (d - \bar{d})^2}{n - 1}}$$

$$t = \frac{\bar{d} - \mu_D}{\sqrt{\frac{\sum_{i=1}^n (d - \bar{d})^2}{n - 1}} \frac{1}{\sqrt{n}}} = \frac{\bar{d} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

Diagrama de anotaciones para la fórmula de t :

- Estimador: apunta a \bar{d}
- Parámetro: apunta a μ_D
- Error estándar: apunta a $\frac{S_D}{\sqrt{n}}$ (circulado)

$$P\left(\bar{d} - t_{\left(1-\frac{\alpha}{2}; n-1\right)} \frac{S_D}{\sqrt{n}} \leq \mu_D \leq \bar{d} + t_{\left(1-\frac{\alpha}{2}; n-1\right)} \frac{S_D}{\sqrt{n}}\right) = 1 - \alpha$$

$$IC(\mu_D) \left[\bar{d} - t_{\left(1-\frac{\alpha}{2}; n-1\right)} \frac{S_D}{\sqrt{n}}; \bar{d} + t_{\left(1-\frac{\alpha}{2}; n-1\right)} \frac{S_D}{\sqrt{n}} \right] = (1 - \alpha)\%$$

Intervalos de Confianza para Muestras apareadas

$$P \left[\bar{d} - T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{d} + T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}} \right] = 1 - \alpha$$

$$IC_{\mu_D} \left[\bar{d} - T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}}; \bar{d} + T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}} \right] = (1 - \alpha)\%$$

$$IC_{\mu_D} \left[\bar{d} \mp T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}}; \bar{d} + T_{\left(1-\frac{\alpha}{2}, n-1\right)} \frac{s_D}{\sqrt{n}} \right] = (1 - \alpha)\%$$

Muestras apareadas

Si las muestras están relacionadas. Muestras pareadas (de a pares o emparejadas). Se usa el estadístico t con n-1 grados de libertad y un nivel de significación α . (n es la cantidad de pares de datos)

$$d_i = X_i - Y_i, \\ i=1, \dots, n$$

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

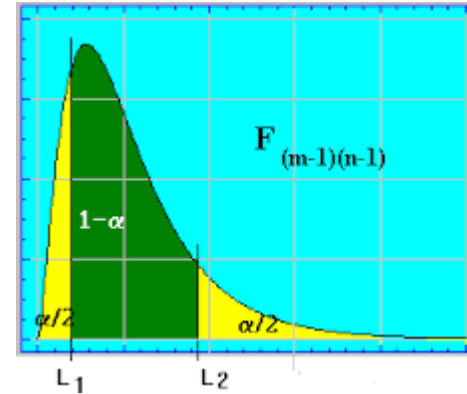
$$s_D = \sqrt{\frac{(d_i - \bar{d})^2}{n-1}}$$

Hipótesis	Estadístico	Criterio de rechazo
$\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D \neq 0 \end{cases}$	$t_e = \frac{(\bar{d}) - (\mu_D)}{\frac{s_D}{\sqrt{n}}}$ $(\mu_D) = 0$	<p>Rechazar H_0 si $t_e < -t_{(1-\frac{\alpha}{2}; n-1)}$</p> <p>O Rechazar H_0 si $t_e > t_{(1-\frac{\alpha}{2}; n-1)}$</p>
$\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D > 0 \end{cases}$		Rechazar H_0 si $t_e > t_{(1-\alpha; n-1)}$
$\begin{cases} H_0: \mu_D = 0 \\ H_1: \mu_D < 0 \end{cases}$		Rechazar H_0 si $t_e < -t_{(1-\alpha; n-1)}$

Comparación de varianzas

Intervalos de confianza para el cociente entre varianzas

Si S_x^2 y S_y^2 son los valores de las varianzas de muestras aleatorias independientes de tamaño n_x y n_y de poblaciones normales, entonces el intervalo de confianza para $\frac{\sigma_x^2}{\sigma_y^2}$ es



$$P\left(\frac{S_x^2}{S_y^2} f_{\left(\frac{\alpha}{2}, n_x-1, n_y-1\right)} < \frac{\sigma_x^2}{\sigma_y^2} < \frac{S_x^2}{S_y^2} f_{\left(1-\frac{\alpha}{2}, n_x-1, n_y-1\right)}\right) = (1-\alpha)\%$$

$$IC\left(\frac{\sigma_x^2}{\sigma_y^2}\right) \left[\frac{S_x^2}{S_y^2} f_{\left(\frac{\alpha}{2}, n_x-1, n_y-1\right)}; \frac{S_x^2}{S_y^2} f_{\left(1-\frac{\alpha}{2}, n_x-1, n_y-1\right)} \right] : (1-\alpha)\%$$

Test de Hipótesis para el cociente entre varianzas.

Hipótesis	Estadístico	Criterio de rechazo
$\begin{cases} H_0: \frac{\sigma_x^2}{\sigma_y^2} = 1 \\ H_1: \frac{\sigma_x^2}{\sigma_y^2} \neq 1 \end{cases}$	$F_e = \frac{s_x^2}{s_y^2}$	<p>Rechazar H_0</p> <p>si $F_e \leq f_{\frac{\alpha}{2}; n_x-1; n_y-1}$</p> <p>O Rechazar H_0</p> <p>si $F_e \geq f_{1-\frac{\alpha}{2}; n_x-1; n_y-1}$</p>
$\begin{cases} H_0: \frac{\sigma_x^2}{\sigma_y^2} = 1 \\ H_1: \frac{\sigma_x^2}{\sigma_y^2} > 1 \end{cases}$		<p>Rechazar H_0</p> <p>si $F_e \geq f_{1-\alpha; n_x-1; n_y-1}$</p>

Test de Hipótesis para el cociente entre varianzas.

Hipótesis	Estadístico	Criterio de rechazo
$\begin{cases} H_0: \frac{\sigma_y^2}{\sigma_x^2} = 1 \\ H_1: \frac{\sigma_y^2}{\sigma_x^2} > 1 \end{cases}$	$F_e = \frac{s_y^2}{s_x^2}$	Rechazar H_0 si $F_e \geq f_{1-\alpha; n_y-1; n_x-1}$

Pruebas no paramétricas para dos muestras independientes

- Pruebas alternativas no paramétricas, o de libre distribución para los casos donde no satisface el supuesto de normalidad.
- Se basan en rangos o *scores*.
- ¿En qué casos se debe preferir un test de rangos?
- Si los datos no son numéricos y corresponden a categorías ordinales, los rangos contienen la misma información que los datos.
- Los datos tienen que ser independientes.
- Si la variable es numérica, su distribución no es Normal y la muestra es pequeña, no valen los tests que hemos presentado anteriormente.

Test de Mann-Whitney-Wilcoxon

Para Dos Muestras Independientes

También llamada Prueba de la suma de rangos de Mann-Whitney

- Esta prueba tiene por objeto verificar si dos muestras independientes fueron extraídas de la misma población (o de distintas poblaciones con igual distribución)
- Es aplicable para:
 - Detectar diferencias entre dos poblaciones sobre la base de sendas muestras seleccionadas de cada población.
 - Detectar diferencias entre dos muestras provenientes de la misma población a las que se somete a tratamientos distintos.
- No es necesario el supuesto de continuidad de la/s población/es que origina/n la/s muestra/s. Por tanto la prueba es válida para cualquier tipo de población: continua, discreta o mezcla de ambas.
- Si las distribuciones poblacionales son iguales, también lo son sus medianas, sus medias, etc.
- Para que dos muestra sean independientes, la independencia se debe cumplir dentro de los elementos de cada muestra y entre todos y cada uno de los elementos de ambas muestras.
- **H0: las dos muestras fueron extraídas de la misma población o de poblaciones idénticas . Las diferencias no son significativas.**
- **H1: las dos muestras no fueron extraídas de la misma población o de poblaciones idénticas. Las diferencias son significativas.**

Estadístico - Test de Mann-Whitney-Wilcoxon

n_1 = número de elementos en la muestra 1

n_2 = número de elementos en la muestra 2

R_1 = suma de los rangos de los elementos en la muestra 1

R_2 = suma de los rangos de los elementos en la muestra 2

Estadístico U $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

Fórmula alternativa para el estadístico U $U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

$$\mu_d = \frac{n_1 n_2}{2} \quad \sigma_d = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$Z_e = \frac{U - \mu_d}{\sigma_d}$$

$$Z_c = \pm Z_{(1-\frac{\alpha}{2})}$$

Se rechaza la hipótesis nula si $z_e > Z_{(1-\frac{\alpha}{2})}$ o si $z_e < -Z_{(1-\frac{\alpha}{2})}$

Ejemplo - Test de Mann-Whitney-Wilcoxon

Una compañía dedicada a la venta de material musical y literario tiene abiertos dos establecimientos, en Pilar y San Isidro. Para tratar de analizar las ventas diarias, se eligieron al azar seis días de un determinado período, obteniéndose los volúmenes de ventas, en miles de pesos:

Utilizando un 5% de significación, ¿señalan estos datos la existencia de diferencias significativas en cuanto a los volúmenes de ventas de dichos establecimientos? Realice el contraste de Mann-Whitney.

Datos - Test de Mann-Whitney-Wilcoxon

Pilar	San Isidro
65	48
84	30
65	98
75	87
60	55
54	96

Test de la mediana

- Este test posee las siguientes características:
- Puede generalizarse a más de dos grupos y resulta ser una alternativa al test de Mann-Whitney-Wilcoxon cuando interesa un test para el parámetro de posición.
- Puede aplicarse sin que se cumpla el supuesto de igualdad distribucional de las dos poblaciones.
- Puede ser usado con datos numéricos u ordinales.
- x_1, x_2, \dots, x_{n_x} observaciones independientes con mediana θ_x
- y_1, y_2, \dots, y_{n_y} observaciones independientes con mediana θ_y

$$H_0: \theta_x = \theta_y$$

$$H_1: \theta_x \neq \theta_y$$

Test de la mediana

Este test, tal como lo calculan la mayoría de los paquetes estadísticos, no acepta hipótesis alternativas unilaterales.

Para definir el estadístico, se ordenan los $n_x + n_y$ datos y se calcula la mediana general de los datos agrupados de ambas muestras, digamos θ . Luego, se cuenta el número de observaciones menores o iguales que la mediana y el número de observaciones mayores que la mediana de cada una de las muestras. Estos datos se vuelcan a una tabla de doble entrada:

	Muestra X	Muestra Y
$<$	m_x	m_y
\geq	M_x	M_y
Total	n_x	n_y

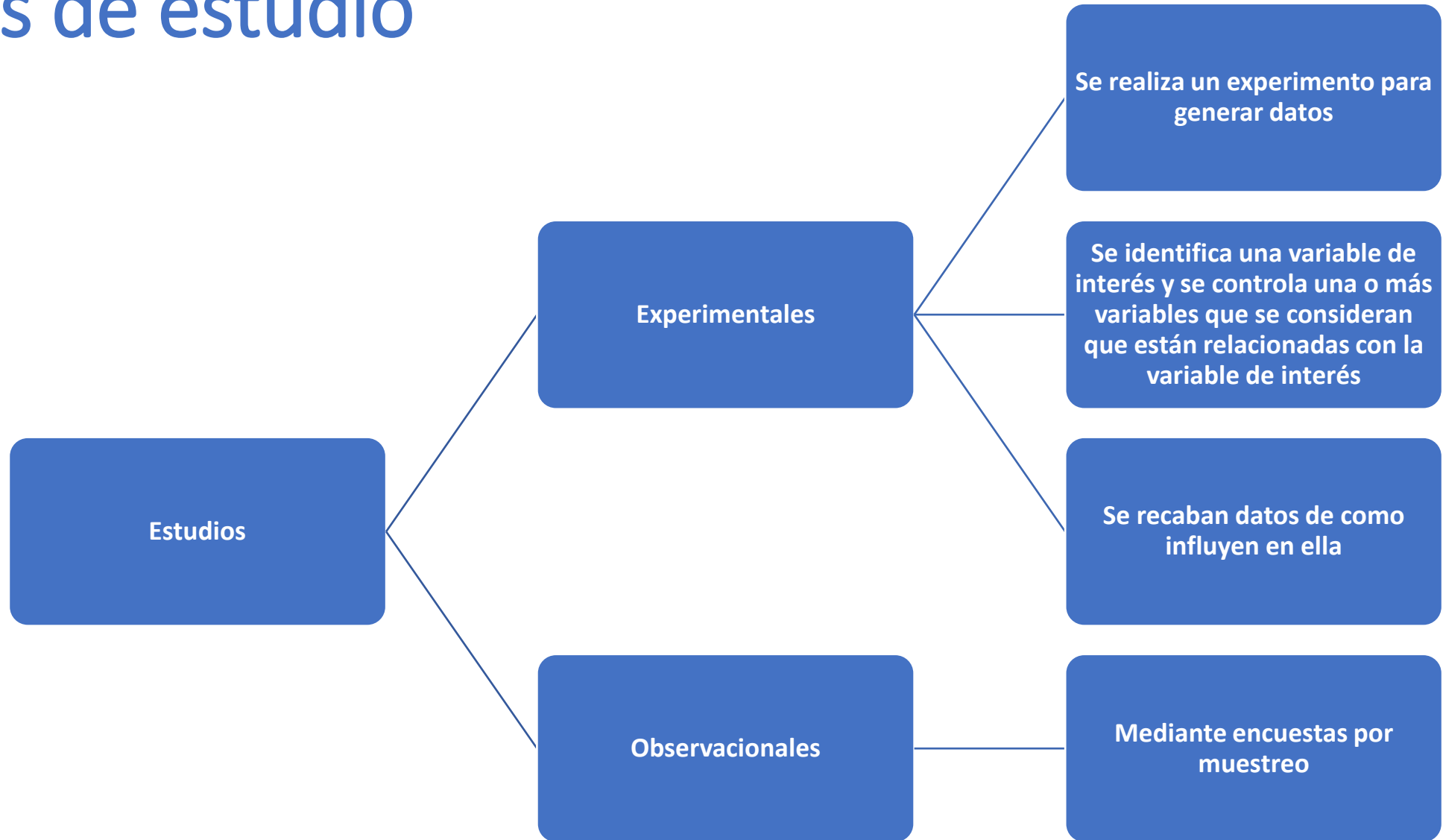
Test de la mediana

Si H_0 fuera verdadera, las proporciones entre los datos menores que la mediana y los mayores que la mediana, deberían ser similares en las dos muestras; es decir, se esperaría que:

$$\frac{m_x}{n_x} \cong \frac{m_y}{n_y} \cong \frac{M_x}{n_x} \cong \frac{M_y}{n_y}$$

El estadístico del test mide la distancia entre lo observado y lo esperado cuando H_0 es verdadera. Si las muestras son relativamente grandes, el estadístico tiene distribución aproximada Chi cuadrado con 1 grado de libertad, X^2

Tipos de estudio

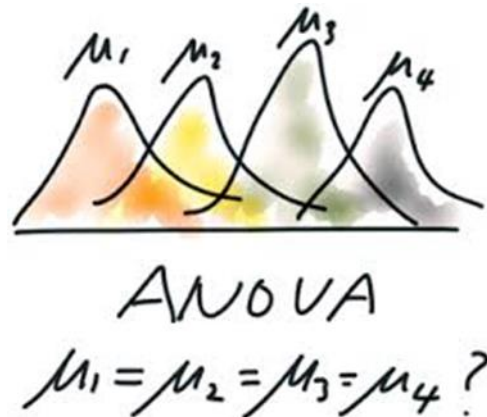


Diseño de experimentos

- La estrategia de diseño de experimentos supone que los factores (tratamientos) se asignan al azar a las unidades experimentales de modo de eliminar el sesgo y simular mejor las condiciones presentes en el modelo. ANOVA a un criterio de clasificación.
- Variable cualitativa es el tratamiento y el dato obtenido es una variable cuantitativa.
- En ocasiones, se introduce una variable de bloqueo (bloques) para reducir el error experimental. Los bloques son completamente aleatorios. Este procedimiento se denomina ANOVA a dos criterios de clasificación.

Tres o más grupos: análisis de la varianza de un factor (ANOVA)

- El Análisis de la Varianza (ANOVA) es un procedimiento para probar medias poblacionales de más de dos poblaciones normales.
- Hay situaciones que requieren de comparación de procesos en más de dos niveles.
- También se utiliza en regresión con el fin de analizar la calidad de la ecuación del modelo.



ANOVA a un criterio de clasificación

Se utiliza cuando se quiere analizar una respuesta cuantitativa (variable dependiente) medida bajo ciertas condiciones experimentales identificadas por una o varias variables categóricas (variables independientes) denominada **TRATAMIENTO**.

Hipótesis a plantear sobre igualdad de medias:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$
$$H_1: \text{al menos alguna de las medias es diferente}$$

K es la cantidad de tratamiento

Se especifica el nivel de significación α

Se utiliza el estadístico F de Snedecor

Requisitos

Independencia

- En relación al procedimiento de selección de la muestra

Normalidad

- Métodos gráficos
 - QQ-plot o gráfico de cuantil cuantil
- Métodos analíticos
 - Shapiro-Wilk
 - Anderson-Darling

Homocedasticidad

- Bartlett
- Levene

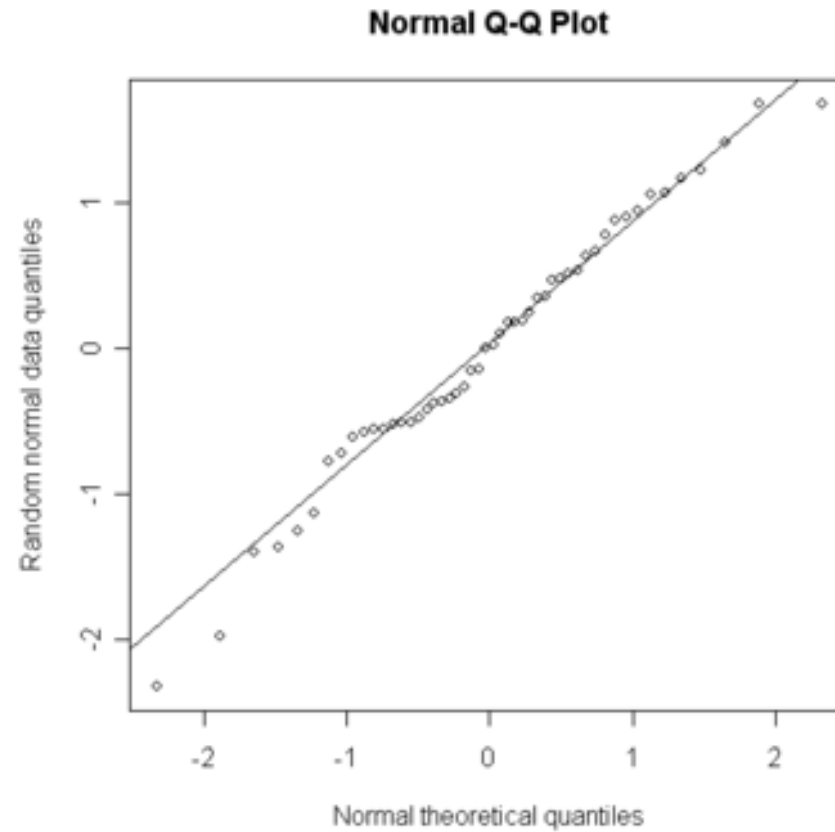
Test de normalidad

- Dentro de las herramientas conocidas, se dispone de distintos tests de normalidad así como de un gráfico que compara los cuantiles empíricos con los esperados, en el caso de que el supuesto se verifica. Este gráfico se denomina QQ-plot o gráfico de cuantil cuantil.
- El programa R tiene implementada una batería de tests de normalidad incluidos en la librería nortest. Dos de los más conocidos y potentes son el test de Shapiro-Wilk y el test de Anderson-Darling

Normalidad - Planteo de Hipótesis

- *Normalidad*
- H_0 : *Los datos siguen una distribución normal*
- H_1 : *los datos no siguen una distribución normal*
- Criterio de rechazo de H_0 cuando $p\text{-valor} < \alpha$
- Se fija un $\alpha = 0,05$

Normalidad - Gráficos de cuantil-cuantil



Normalidad - Test de Shapiro-Wilk

- Se usa para contrastar la normalidad de un conjunto de datos.
- El test de *Shapiro-Wilks* plantea la hipótesis nula que una muestra proviene de una distribución normal. Se elige un nivel de significanza, por ejemplo 0,05, y la hipótesis alternativa sostiene que la distribución no es normal.

H_0 : La distribución es normal

H_1 : La distribución no es normal,

o más formalmente aún:

$H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$

$H_1 : X \not\sim \mathcal{N}(\mu, \sigma^2).$

Normalidad - Test de Anderson-Darling

- El estadístico Anderson-Darling mide qué tan bien siguen los datos una distribución específica.
- Para un conjunto de datos y distribución en particular, mientras mejor se ajuste la distribución a los datos, menor será este estadístico.
- Por ejemplo, se puede utilizar el estadístico de Anderson-Darling para determinar si los datos cumplen el supuesto de normalidad.
- Las hipótesis para la prueba de Anderson-Darling son:
- H_0 : Los datos siguen una distribución especificada
- H_1 : Los datos no siguen una distribución especificada

Homocedasticidad - Planteo de Hipótesis

- *Homocedasticidad*

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

H_1 : al menos alguna de las varianzas es diferente

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1: \exists(i, j) \sigma_i^2 \neq \sigma_j^2$$

- Criterio de rechazo de H_0 cuando p-valor $< \alpha$
- Se fija un $\alpha = 0,05$

Homocedasticidad - Test de Bartlett

- Si el test de Bartlett no rechaza la hipótesis de nulidad, significa que, no hay evidencia estadística significativa de que la varianza de alguno de los subgrupos difiera de las otras.
- El problema de este test es su sensibilidad a la falta de normalidad. Esto implica que puede ocurrir que el mismo rechace la hipótesis nula por no cumplirse el supuesto de normalidad en lugar de rechazarla por no cumplirse el supuesto de homocedasticidad.
- Una alternativa más robusta, lo que significa que no es sensible a la falta de normalidad o a la presencia de algún valor atípico, la brinda el test de Levene.

Homocedasticidad - Test de Levene

- El test de Levene realiza un nuevo análisis de la varianza para los valores absolutos de los residuos de las observaciones respecto de la mediana, o la media, de su grupo.
- Cuando el p-valor del test es mayor a 0,05 significa que el test de Levene no rechaza la hipótesis nula de homocedasticidad, lo que brinda la misma conclusión que el test de Bartlett. Por lo tanto, podemos suponer que se cumple la hipótesis de homocedasticidad.
- Faltaría analizar el cumplimiento del supuesto de normalidad de la distribución de los residuos, que es equivalente a analizar el supuesto de normalidad de la distribución de la variable original.

Descomposición de la suma de cuadrados totales. Demostrar que $SCT = SCE + SCD$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}_{..})^2$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..})]^2$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}_{..}) + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k (n_i - 1)S_{i.}^2 + \sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

$$\sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2 = SCE$$

$$\sum_{i=1}^k (n_i - 1)S_{i.}^2 = SCD$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = SCT$$

Tabla ANOVA

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio	Estadístico F
Variación entre grupos SCE	$\sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2$	$k - 1$	$SCE / k - 1$	$F = \frac{SCE / k - 1}{SCD / n - k}$
Variación dentro de los grupos SCD	$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$ $= \sum_{i=1}^k (n_i - 1) s_i^2$	$n - k$	$SCD / n - k$	
Variación total SCT	$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2$	$n - 1$		Región crítica $\left\{ \frac{SCE / k - 1}{SCD / n - k} > f_{k-1, n-k, 1-\alpha} \right\}$

ANOVA – Ejemplo

- Cuatro atletas de triple salto saltan las siguientes distancias, en metros, en diferentes reuniones de atletismo. ¿Puede suponerse que los cuatro saltan la misma distancia? Realice el análisis de la varianza utilizando un nivel de significación de 0,05.

Atletas	1	8,25	6,97	7,65	7,86	
	2	7,31	8,26	6,5	9,42	10,4
	3	7	7,25	8		
	4	6,69	7,32			

Tabla ANOVA

ANÁLISIS DE VARIANZA						
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	Fe	P-valor	Fc
SCE	3.42321405	3	1.14107135	0.99586683	0.43401097	3.70826482
SCD	11.4580717	10	1.14580717			
SCT	14.8812857	13				

Test de Kruskal-Wallis no paramétrico para k muestras independientes

- El objetivo de esta prueba es verificar si k muestras independientes provienen de la misma población o de poblaciones idénticas. Es una extensión de la prueba de rangos de Mann Whitney para más de dos poblaciones.
- **H_0 : las k muestras provienen de la misma población o de poblaciones idénticas.**
- **H_1 : algunas de las k muestras no provienen de la misma población o de poblaciones idénticas.**
- Si la hipótesis alternativa es cierta, al menos un par de grupos no provienen de la misma población o de poblaciones idénticas.

Estadístico - Test de Kruskal-Wallis no paramétrico para muestras independientes

$$x_{ij} = \Theta_i + \varepsilon_{ij} \quad i = 1, \dots, k \quad j = 1, \dots, n_i$$

Para facilitar el calculo se usa la siguiente expresión

k es la cantidad de muestras o grupos.

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$$

La distribución de H es libre y puede demostrarse que es asintóticamente, que bajo H_0 , es una $\chi^2_{(k-1)}$

Por tanto, La región critica es $H > \chi^2_{(1-\alpha; k-1)}$

k: es la cantidad de grupos o tratamientos

$$(1 - \alpha) = 0.95$$

La prueba de Kruskal-Wallis fue propuesta por William Henry Kruskal y W. Allen Wallis en el artículo "Use of ranks in one-criterion variance analysis" publicado en el "Journal of American Statistics Association" en 1952.

Kruskal-Wallis - Ejemplo:

- Un fabricante de juguetes desea conocer si existen diferencias en cuanto a la calidad de las tres marcas de pilas alcalinas más extendidas en el mercado con el fin de recomendarlas para su utilización en un nuevo juguete que se va a promocionar. Para comparar las tres marcas, toma muestras aleatorias de pilas de cada una de ellas y controla el tiempo que permanece funcionando el juguete en cuestión.
- Utilizando un nivel de significación del 5%, ¿puede decirse que existen diferencias significativas en las calidades de estas marcas de pilas alcalinas? En su caso, indique qué marcas presentan diferencias significativas. Realice el contraste de Kruskal-Wallis.

Kruskal-Wallis - Tabla de Datos

Marca de pilas	A	B	C
Horas de duración	125	87	55
	140	100	50
	218	85	78
	78	65	93
	98	115	60
		83	78

Kruskal-Wallis – Código en R

```
# Kruskal Wallis
library(ggplot2) # Paquete para confeccionar dibujos
install.packages("pgirmess")
library(pgirmess)
# Paquete con herramientas para lectura, escritura y transformación
de datos
Horas=c(125,140,218,78,98,87,100,85,65,115,83,55,70,78,93,60,78)
Marca=as.factor(c(rep("A",5), rep("B",6), rep("C",6)))
Calidad=data.frame (Marca, Horas)
Calidad
# Carga la base de datos
ggplot(Calidad,aes(x=Marca,y=Horas,fill=Marca)) +
  geom_boxplot() +
  xlab("") +
  scale_fill_brewer(palette="Pastel1")
# Produce boxplots

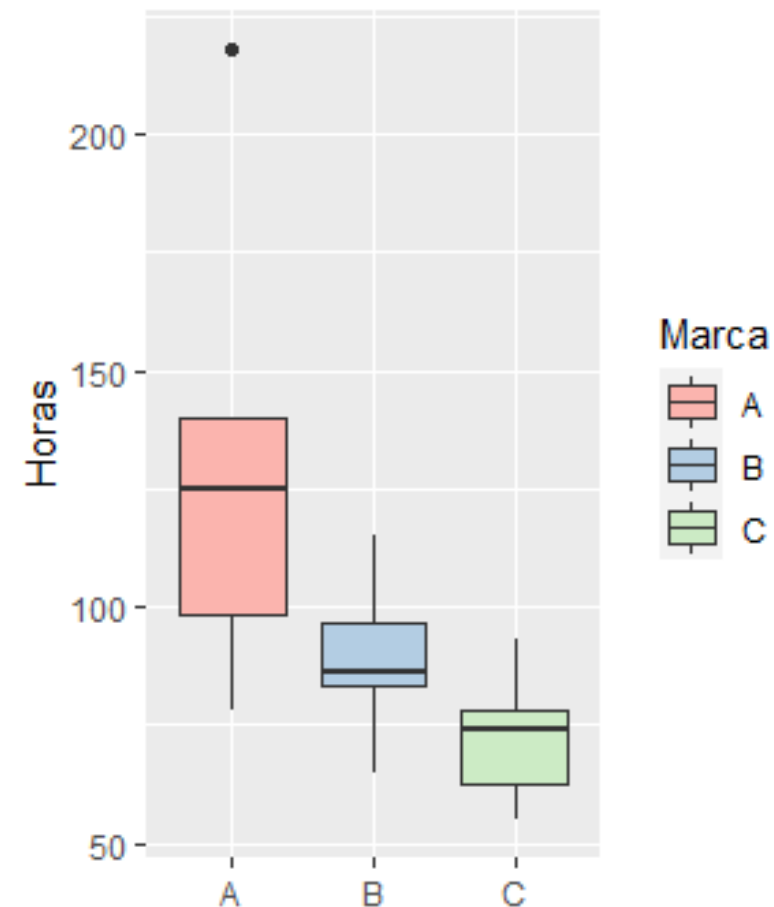
marcaA=Calidad[Calidad$Marca=="A",2]
marcaB=Calidad[Calidad$Marca=="B",2]
marcaC=Calidad[Calidad$Marca=="C",2]
shapiro.test(marcaA)
shapiro.test(marcaB)
shapiro.test(marcaC)
# Aplica el test de Shapiro–Wilk a cada grupo
kruskal.test(Horas, Marca)
# Realiza el test de Kruskal–Wallis
```

Kruskal-Wallis – Resultados en R

Kruskal-Wallis rank sum test

data: Horas and Marca

Kruskal-Wallis chi-squared = 7.3182, df = 2, p-value = 0.02576



¿Preguntas?

