Anotaciones Examen Parcial

Jesus Mauricio Huayhua Flores

2024-05-12

Librerias

```
library(DescTools)
library(ggplot2)
library(MASS)
```

Lectura de archivos

```
# Para leer archivos xlsx
data <- read_xlsx(path="directory/file.xlsx")
head(data)
# Para leer archivos csv
data <- read.csv("directory/file.csv")
```

Pruebas de bondad de ajuste

Distribución multinominal

- Repetir n veces, de forma independiente
- $k (k \geqslant 2)$ resultados o categorias
- $C_1, C_2, \ldots, C_k$ con $p_1, p_2, \ldots, p_k$
- $\sum_{i=1}^{k} p_i = 1$
- Notación: $(X_1, X_2, \ldots, X_k) \sim Mult(n, p_1, p_2, \ldots, p_k)$

Función de probabilidad conjunta

Si $(X_1, X_2, \ldots, X_k) \sim Mult(n, p_1, p_2, \ldots, p_k)$

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

- $X_1, X_2, \ldots, X_k$ **NO** son independientes - $X_1 \sim B(n, p_i), \forall i = 1, \ldots, k$ - $E(X_i) = np_i$ - $V(X_i) = np_i(1 - p_i)$ - $Cov(X_i, X_j) = -np_i p_j, (si)\ i \neq j$

Pruebas Chi-cuadrado

Utilizado para contrastar hipótesis acerca de los parámetros, de 1 o varias distribuciones multinominales.

$$W = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$$

Teorema 1

Si **n** es grande, $\forall i = 1, \ldots, k$ se cumple que $E_i = np_i \geqslant 5$

$$W = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

Tiene distribución Chi-cuadrado con (K - 1) grados de libertad $W \sim X^2_{(k-1)}$

Prueba Chi-cuadrado para bondad de ajuste

Es la prueba para determinar si la variable que observamos se ajusta o no a una distribución teórica.

$$\begin{cases} H_0 : & F_Y = F_o \\ H_1 : & F_Y \neq F_o \end{cases}$$

Experimento multinominal de k categorías excluyentes.

- $C_1, \ldots, C_k$ con probablidades $p_1, \ldots, p_k$
- Frecuencias observadas $x_1, \ldots, x_2$, donde $\sum_{i=1}^{k} x_i = n$
- X_i números de veces en que ocurre la categoría C_i
- Contrastar si las frecuencias observadas de cada categoría difieren significativamente, de las frecuencias esperadas $e_i = np_i$

$$\begin{cases} H_0 : & \forall i : p_i = p_i^o \\ H_1 : & \exists i | p_i \neq p_i^o \end{cases}$$

Si $e_i = np_i^0 \geqslant 5$, la estadística de prueba:

$$W = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k} \frac{(O_i - e_i^0)^2}{ei}$$

Punto crítico, se distribuye Chi-cuadrado con $k - 1$ grados de libertad: $W \sim X^2_{k-1}$.

Pasos para la prueba de Chi-cuadrado(x^2) de bondad de ajuste

1. Establecer las hipótesis sobre la función de distribución desconocida

$$\begin{cases} H_0 & : F_Y = F_0 \\ H_1 & : F_Y \neq F_0 \end{cases}, con\ F_0\ conocida$$

2. Obtener una tabla aleatoria de tamaño **n** de Y, con una tabla de distribución de frecuencias.

| Intervalos | marca de clase | frecuencia observada | frecuencia esperada |

3. Se calculan las frecuencias esperadas E_i^0
4. Se calcula Estadístico de prueba

$$U_0 = \sum_{i=1}^{k} \frac{(O_i - E_i^0)^2}{E_i^0} \sim X_(k - m - 1)^2$$

5. Se recha si $U_0 > X^2_{1-\alpha,k-m-1}$, donde r es la cantidad de parametros desconocidos.

Comandos utilices en R

Opción1:

```
normalityTest(horas ~ pastillas, test="shapiro.test", data=Dataset)
```

Opción2:

```
# Histograma
auxhist <- hist(X,xlab="Largo (mm)", main= "Histograma del Largo")   #  Al crear el histograma se guardan las características de la distribución.
# Prueba de bondad de ajuste
cortes_histo <- auxhist$breaks
probAcum_DistNormal <- pnorm(cortes_histo, mean = mean(X), sd = sd(X))
n <- length(probAcum_DistNormal)
intervalos <- data.frame(probAcum_DistNormal[-n],probAcum_DistNormal[-1])
prop.esperada <- intervalos[,2]-intervalos[,1]
freq.esperada <- (prop.esperada/sum(prop.esperada))*length(X)
freq.observada <- auxhist$counts
chisq.test(freq.observada, p=prop.esperada, rescale.p=TRUE, simulate.p.value = T)
```

Pruebas no paramétricas

Los investigadores están más familiarizados con las pruebas paramétricas, estas tienen supuestos usualmente acerca del tipo de variables y la distribución de la variable, los cuales tiene que ser verificados y en caso no cumplas los supuestos, utilizar pruebas no paramétricas.

Prueba Wilcoxon

Utilidad

Evaluación de la medida de posición de una muestra.

No se requiere de ningún supuesto acerca de la forma de la distribución de la población.

- $H_0 : Me = Me_0$

Procedimiento

Diferencia entre cada valor observado y el valor hipotético de la mediana.

$$d = (X - med_0)$$

Calculamos al diferencia sin tomar el signo de las mismas.

En caso de empate, se asigna un rango promedio de todas las diferencias empatadas.

La suma de rangos positivos S^+ es el estadístico de prueba, el cual es comparado con un valor de la talba de Wilcoxon

Forma práctica, se enumar de menor a mayor diferencia, y posteriormente se verifica en cuantos se encuentra empate en la diferencia para posteriormente realizar el promedio de las diferencias de dichos valores

Otros estadísticos de prueba

Si la muestra es grande se puede usar el siguiente estadístico de prueba:

$$Z_c = \frac{S^+ - \mu_{S^+}}{\sigma_{S^+}} \sim N(0,1)$$

$$\mu_{S^+} = \frac{n(n+1)}{4} \qquad \sigma_{S^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

El par de hipótesis de esta prueba pueden ser estos:

Unilateral izquierda	Bilateral	Unilateral derecha
H_0: $Me_D \geq Me_{D0}$	H_0: $Me_D = Me_{D0}$	H_0: $Me_D \leq Me_{D0}$
H_1: $Me_D < Me_{D0}$	H_1: $Me_D \neq Me_{D0}$	H_1: $Me_D > Me_{D0}$

En R

```
wilcox.test(data, mu = , alternative = "greater",conf.level = 1 - alpha, correct = F)
```

Prueba de Hipótesis para 2 parámetros

Diferencia de Medias

- Contrastar la diferencia de 2 medias $H_0 : \mu_x - \mu_y = \delta_o$
- Puede ser tanto muestras independientes o muestras relacionadas.

Para parámetros de muestras independientes

Hipótesis:

$$\begin{cases} H_0 & : \sigma_x^2/\sigma_y^2 = 1 \\ H_1 & : \sigma_x^2/\sigma_y^2 \neq 1 \end{cases}$$

1. Verificando la distribución, si:

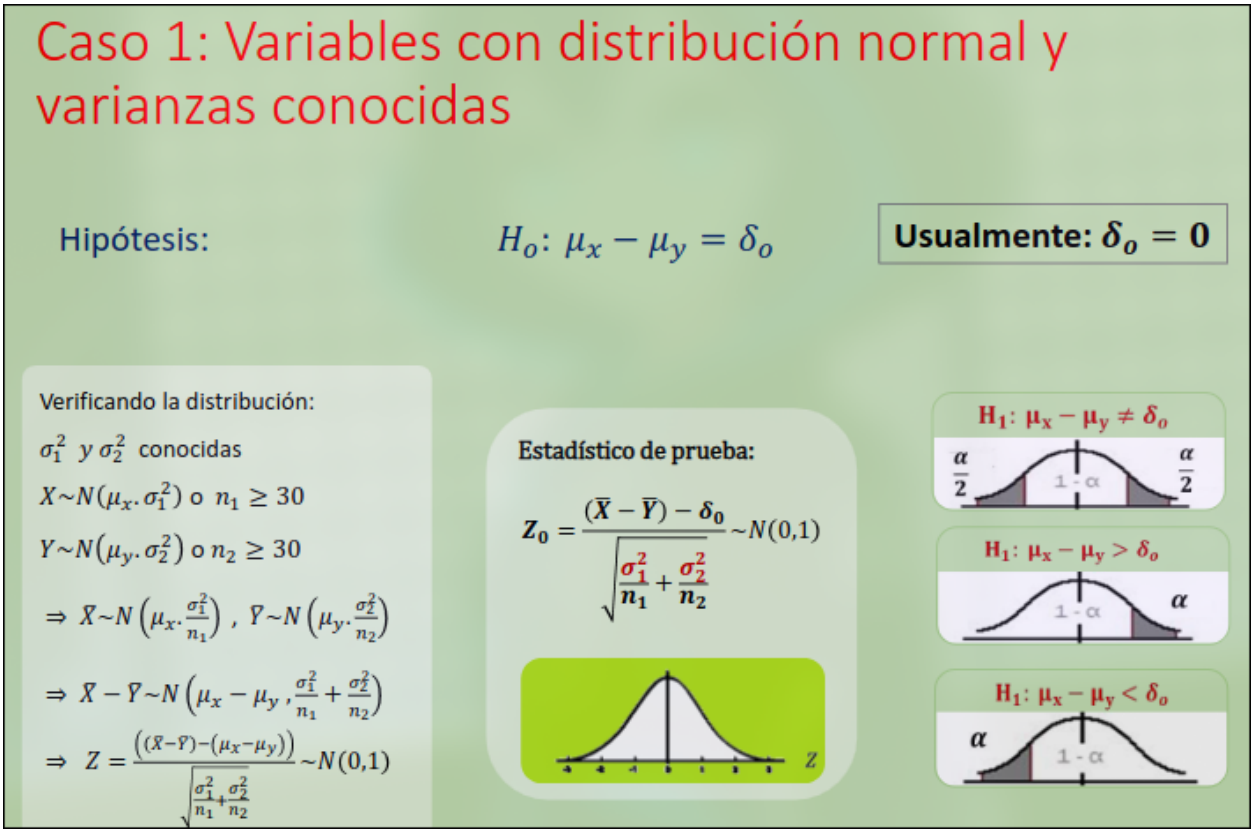
$$X \sim N(\mu_1, \sigma_1^2) \; y \; Y \sim N(\mu_2, \sigma_2^2) \Rightarrow F = \frac{S_x^2}{S_y^2} \sim Fisher(n_1 - 1, n_2 - 1)$$

Donde $n1$ y $n2$ son el tamañao de la muestra para X y Y respectivamente.

Para hallar la fisher es con el siguiente comando:

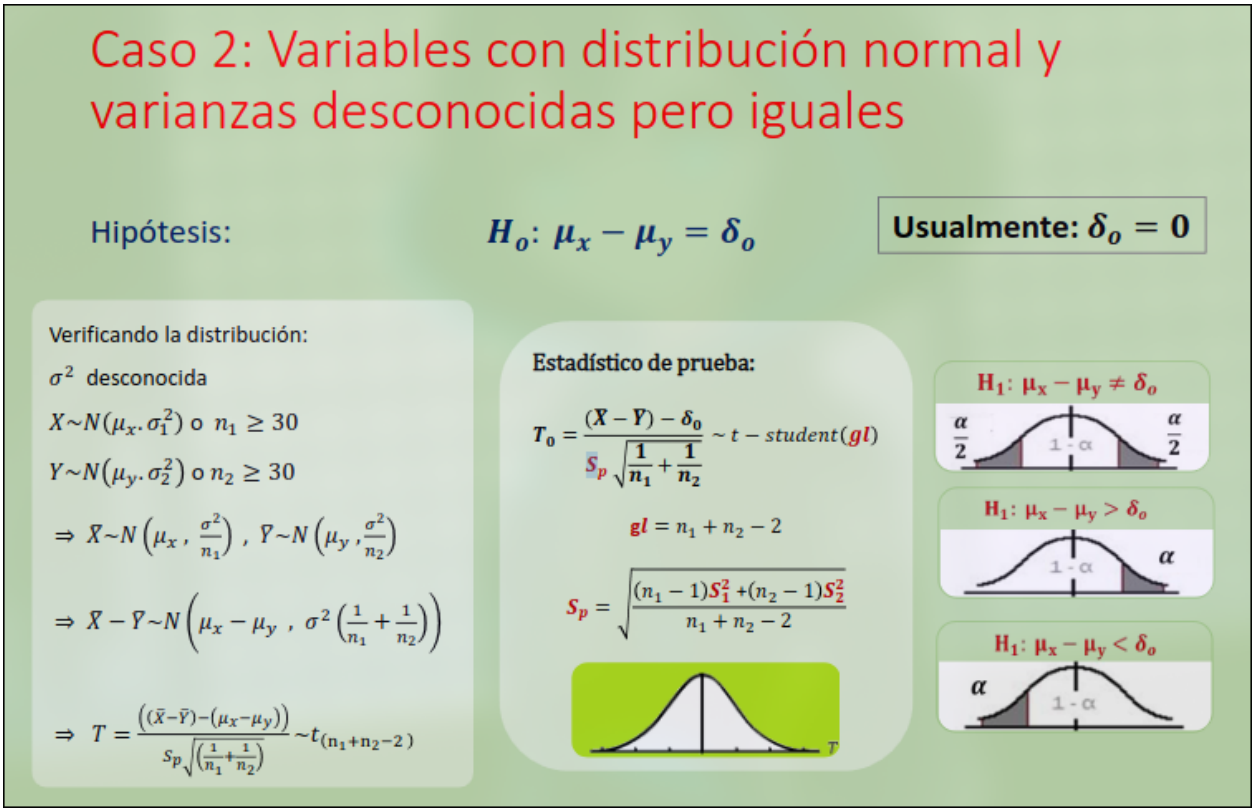
```
alpha = 0.05
qf(alpha/2,n1 - 1,n2 - 1)
qf(1-alpha/2,n1 - 1 ,n2 -1)
```

Varianzas conocidas Espacio




```
z0 = ((medianX - mediamY) - delta0)  / (sqrt())
```

Varianzas desconocidas e iguales espacio




```
n1 = 5
n2 = 5
s1 = 0.1811
s2 = 0.067
media1 = 0.417
media2 = 0.151
sp = sqrt(((n1 - 1)*(s1^2) + (n2 - 1)*(s2^2))   /   (n1 + n2 - 2) )
to = (media1 - media2) / (sp * sqrt(1/n1 + 1/n2))

qt(alpha,n1 + n2 - 2)

# to > qt rechazamos HO
```

Varianzas desconocidas y diferentes Espacio

Caso 3: Variables con distribución normal y varianzas desconocidas y diferentes

Hipótesis: $H_o: \mu_x - \mu_y = \delta_o$ **Usualmente: $\delta_o = 0$**

Verificando la distribución:

σ_1^2 y σ_2^2 desconocidas

$X \sim N(\mu_x, \sigma_1^2)$ o $n_1 \geq 30$

$Y \sim N(\mu_y, \sigma_2^2)$ o $n_2 \geq 30$

$$\Rightarrow \bar{X} \sim N\left(\mu_x, \frac{\sigma_1^2}{n_1}\right), \ \bar{Y} \sim N\left(\mu_y, \frac{\sigma_2^2}{n_2}\right)$$

$$\Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\right)$$

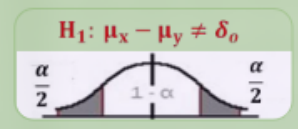
$$\Rightarrow T = \frac{\left((\bar{X}-\bar{Y}) - (\mu_x - \mu_y)\right)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \sim t_{(gl)}$$

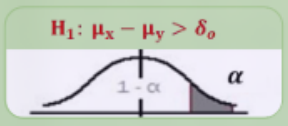
$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

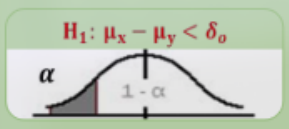
Estadístico de prueba:

$$T_0 = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}} \sim t - student(gl)$$

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$








```
gl = (S1/n1 + S2/n2) / ( ((s1/n1)^2)/(n1 - 1) +  ((s2/n2)^2)/(n2 - 1) )
qt(1-alpha, n1 + n2 - 2)
```

Prueba de hipótesis para la diferencia de medias de muestras relacionadas

Prueba de hipótesis de diferencia de medias para muestras pareadas o relacionadas

Hipótesis: $H_o: \mu_x - \mu_y = \delta_o$

Usualmente: $\delta_o = 0$

Los objetos de estudio son los mismos o están relacionados.
La diferencia de las variables sigue una distribución normal.

Verificando la distribución:

$$\bar{D} = \bar{X} - \bar{Y} = \sum D_j / n$$

$$T = \frac{(\bar{D} - \mu_D)}{S_{\bar{D}}} \sim t - Student(n-1)$$

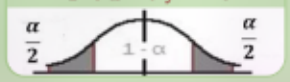
$$S_D = \sqrt{\frac{\sum(D_j - \bar{D})}{n-1}}$$

$$S_{\bar{D}} = \frac{S_D}{\sqrt{n}}$$

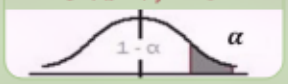
Estadístico de prueba:

$$T = \frac{(\bar{D} - \delta_o)}{S_{\bar{D}}} \sim t - Student(n-1)$$

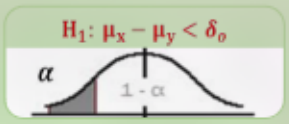
$H_1: \mu_x - \mu_y \neq \delta_o$



$H_1: \mu_x - \mu_y > \delta_o$



$H_1: \mu_x - \mu_y < \delta_o$



Prueba de hipótesis para la Diferencia de Medias $H_o: \mu_x - \mu_y = \delta_o$ de **muestras independientes** a nivel α de significancia

Caso	Estadística de Prueba	$H_o: \mu_x - \mu_y = \delta_o$	Rechazar H_0 si
Caso 1: σ_1^2 y σ_2^2 conocidas $X \sim N(\mu_x, \sigma_1^2)$, $Y \sim N(\mu_y, \sigma_2^2)$, o $n_1 \geq 30, n_2 \geq 30$	$Z_0 = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$
Caso 2: misma σ^2 desconocida $X \sim N(\mu_x, \sigma^2)$, $Y \sim N(\mu_y, \sigma^2)$, o $n_1 \geq 30, n_2 \geq 30$	$T_0 = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t - student(gl)$ $gl = n_1 + n_2 - 2 \quad S_p = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$T_0 > t_{1-\alpha}$ $T_0 < -t_{1-\alpha} = t_\alpha$ $\lvert T_0 \rvert > t_{1-\alpha/2}$
Caso 3: σ_1^2 y σ_2^2 desconocidas $X \sim N(\mu_x, \sigma_1^2)$, $Y \sim N(\mu_y, \sigma_2^2)$, o $n_1 \geq 30, n_2 \geq 30$	$T_0 = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\dfrac{S_x^2}{n_1} + \dfrac{S_y^2}{n_2}}} \sim t - student(gl)$ $gl = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$T_0 > t_{1-\alpha}$ $T_0 < -t_{1-\alpha} = t_\alpha$ $\lvert T_0 \rvert > t_{1-\alpha/2}$
Caso 4: σ_1^2 y σ_2^2 desconocidas $n_1 \geq 30, n_2 \geq 30$	$Z_0 = \dfrac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim N(0,1)$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$

Prueba de hipótesis para la Diferencia de Medias $H_o: \mu_x - \mu_y = \delta_o$ de **muestras relacionadas** a nivel α de significancia

Caso	Estadística de Prueba	$H_o: \mu_x - \mu_y = \delta_o$	Rechazar H_0 si
Caso 1: σ_D^2 es conocida y D tiene distribución normal o n es grande	$D = X - Y \quad Z_0 = \dfrac{\bar{D} - \delta_0}{\sigma_D/\sqrt{n}} \sim N(0,1)$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$
Caso 2: σ_D^2 es desconocida y D tiene distribución normal	$D = X - Y \quad T_0 = \dfrac{\bar{D} - \delta_0}{S_D/\sqrt{n}} \sim T(n-1)$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$T_0 > t_{1-\alpha}$ $T_0 < -t_{1-\alpha} = t_\alpha$ $\lvert T_0 \rvert > t_{1-\alpha/2}$
Caso 3: σ_D^2 es desconocida y el tamaño de muestra n es suficientemente grande	$D = X - Y \quad Z_0 = \dfrac{\bar{D} - \delta_0}{S_D/\sqrt{n}} \sim N(0,1)$	$H_1: \mu_x - \mu_y > \delta_0$ $H_1: \mu_x - \mu_y < \delta_0$ $H_1: \mu_x - \mu_y \neq \delta_0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$

Prueba de hipótesis para la Diferencia de proporciones $H_0: p_1 - p_2 = 0$ a nivel α de significancia

Caso	Estadística de Prueba	$H_0: p_1 - p_2 = 0$	Rechazar H_0 si
Caso 1: poblaciones independientes tamaños de muestra n_1 y n_2 grandes	$Z_0 = \frac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} \sim N(0,1)$	$H_1: p_1 - p_2 > 0$ $H_1: p_1 - p_2 < 0$ $H_1: p_1 - p_2 \neq 0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$
Caso 2: Una población tamaño de muestra n grande	$Z_0 = \frac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1) + \bar{p}_2(1-\bar{p}_2) + 2\ \overline{p_1}\ \overline{p_2}}{n}}} \sim N(0,1)$	$H_1: p_1 - p_2 > 0$ $H_1: p_1 - p_2 < 0$ $H_1: p_1 - p_2 \neq 0$	$Z_0 > z_{1-\alpha}$ $Z_0 < -z_{1-\alpha} = z_\alpha$ $\lvert Z_0 \rvert > z_{1-\alpha/2}$

Prueba de hipótesis para la Razón de Varianzas $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ a nivel α de significancia

Caso	Estadística de Prueba	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$	Rechazar H_0 si
$X \sim N(\mu_1, \sigma_1^2)$ $Y \sim N(\mu_2, \sigma_2^2)$	$F_0 = \frac{S_x^2}{S_y^2} \sim Fisher(n_1 - 1, n_2 - 1)$	$H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} = 1$	$F_0 > F_{1-\alpha}$ $F_0 < F_\alpha$ $F_0 > F_{1-\alpha/2}$ o $F_0 < F_{\alpha/2}$

Prueba de hipótesis NO paramétricas para 2 o más parámetros

Prueba de U de Mann Whitney

- Es considerada como la alternativa no paramétrica a una prueba de diferencias de medias.
- Se basa en los rangos observados
- Permite comparar las **medianas** de dos muestras independientes son diferentes.