# Taller #3

Sergio Andrés Díaz Vera  Samuel Ruíz Martínez  Hernan Supelano Vega
Daniel Felipe Cendales G.

## Intervalos de confianza

1. Sean $Y_1, \ldots, Y_n$ con $Y_i \overset{iid}{\sim} \mathcal{N}\left(\mu_0, \sigma^2\right)$ para todos $i = 1, \ldots, n$

    a. Tomemos $\mu_0 = 2, \sigma = \sqrt{2}$

```
# Asignación de parámetros
mu_0 <- 2; sigma <- sqrt(2)

# Fijamos la semilla
set.seed(31415)
```

    b. Simulaciones

```
## Metaparámetros
N <- 1000           # Número de simulaciones
n <- 100            # Tamaño de cada muestra
alpha <- 0.05

## Simulaciones
intervalos <- sapply(1:N, function(k){
    x <- rnorm(n = n, mean = mu_0, sd = sigma)
    x_barra <- mean(x)
    lims <- x_barra + c(-1, 1)*qnorm(1 - alpha/2)*sigma/sqrt(n)
    c(lims, x_barra)
})

## Trasponemos los resultados
intervalos <- t(intervalos)

## Función que verifica si se está entre un intervalo
entre <- function(x, valor)
    ifelse(x[1] <= mu_0 & mu_0 <= x[2], 1, 0)

## Contamos la cantidad de intervalos que contienen a mu_0
cantidad <- apply(intervalos, MARGIN = 1,
                  FUN = entre, valor = mu_0)
```

    c. Cantidad de intervalos que contienen a $mu_0$

```
## Intervalos que contienen a mu_0
sum(cantidad)
```

```
## [1] 958
```

Que era de esperarse, ya que aproximadamente $1000 * (1 - \alpha)$ de los intervalos deben contener a la media.

2. Sean $Y_1, \ldots, Y_n$ con $Y_i \overset{iid}{\sim} \mathcal{N}\left(\mu_0, \sigma^2\right)$ para todos $i = 1, \ldots, n$

    a. Tomemos $\mu_0 = 4, \sigma = 3$

```
# Asignación de parámetros
mu_0 <- 4; sigma <- 3

# Fijamos la semilla
set.seed(27182)
```

    b. Simulaciones

```
## Metaparámetros
N <- 1000           # Número de simulaciones
n <- 100            # Tamaño de cada muestra
alpha <- 0.05

## Simulaciones
muestras <- sapply(1:N, function(k){
    rnorm(n = n, mean = mu_0, sd = sigma)
})

## Función que calcula intervalos y valores P
cAlculos <- function(y){
    prueba_t <- t.test(y, alternative = "two.sided",
                       conf.level = 1 - alpha,
                       mu = mu_0)
    c(prueba_t$conf.int, prueba_t$p.value)
}

## Aplicamos la función a cada una de las muestras
resultados <- apply(muestras, MARGIN = 2, FUN = cAlculos)

## Cálculo de la cantidad de intervalos que contienen a mu_0
cantidad <- apply(resultados, 2, FUN = entre, valor = mu_0)
```

- Cantidad de intervalos que contienen a $mu_0$:

```
## Intervalos que contienen a mu_0
sum(cantidad)
```

```
## [1] 944
```

Lo cual concuerda con la teoría.

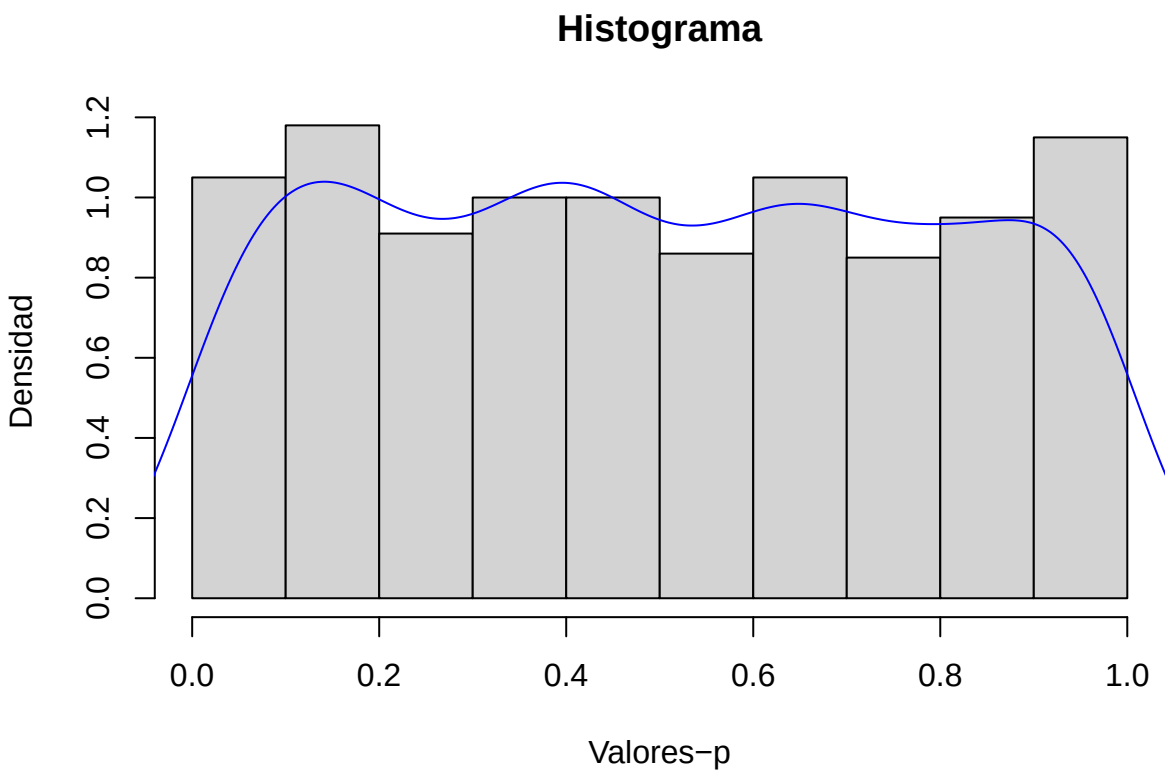
- Conteo de *valores-p* menores a 0.05

```
## Número total de rechazos
sum(resultados[3, ] < 0.05)
```

```
## [1] 56
```

- Distribución empírica (y teórica) del *valor-p*

```
## Histograma y densidad ajustada
hist(resultados[3, ], main = "Histograma",
     xlab = "Valores-p", ylab = "Densidad",
     freq = FALSE)
# Curva suavizada
lines(density(resultados[3, ]), col = "blue")
```




Que era de esperarse, ya que aproximadamente $1000 * (1 - \alpha)$ de los intervalos deben contener a la media.

## Pruebas No Paramétricas

3. Cuadro de resumen
4. Ejercicios libro *Hollander y Wolfe*
    a. Test de Wilcoxon: Ejercicios **9** y **11**

(9.) Supongamos que $n = 5$ y hemos observado $Z_1 = -1.3, Z_2 = 2.4, Z_3 = 1.3, Z_4 = 1.3$ y $Z_5 = 2.4$

Para calcular todos los posibles valores que puede tomar el estadístico, podemos generar las $2^n$ tuplas de 1's y 0's de longitud $n$, hacer el producto punto con los rangos y contruir una tabla de frecuencias.

Evitamos poner el código que nos genera la distribución del estadístico.

```
## Datos
z  <- c(-1.3, 2.4, 1.3, 1.3, 2.4)     # Observaciones
r1 <- rank(abs(z))                    # Rangos con empates
r2 <- order(abs(z))                   # Rangos sin empates
phi <- ifelse(z > 0, yes = 1, 0)

## Cálculo del estadístico con empates
t_1 <- sum(r1 * phi)
t_2 <- sum(r2 * phi)

## Dsitribución
distrib_wilc(r1)
```

```
## # A tibble: 12 x 3
##        t `P[T>=t]` `p-val`
##    <dbl> <chr>       <dbl>
```

```
##  1  15   1/32       0.0312
##  2  13   4/32       0.125
##  3  11   7/32       0.219
##  4  10.5 9/32       0.281
##  5   9   10/32      0.312
##  6   8.5 16/32      0.5
##  7   6.5 22/32      0.688
##  8   6   23/32      0.719
##  9   4.5 25/32      0.781
## 10   4   28/32      0.875
## 11   2   31/32      0.969
## 12   0   32/32      1
```

A continuación podemos el *valor-p* asociado a cada test usando (y evitando) los empates.

```
## Valor P asociado
subset(distrib_wilc(r1), t == t_1)
```

```
## # A tibble: 1 x 3
##       t `P[T>=t]` `p-val`
##   <dbl> <chr>       <dbl>
## 1    13 4/32        0.125
```

```
## Valor P asociado
subset(distrib_wilc(r2), t == t_2)
```

```
## # A tibble: 1 x 3
##       t `P[T>=t]` `p-val`
##   <dbl> <chr>       <dbl>
## 1    14 2/32       0.0625
```

Podemos ver que, usando la distribución erronea, el *valor-p* asociado es de 0.0625. Sin embargo, usando el test y distribución apropiados, el *valor-p* es de 0.125. Lo que implica que se hubiese tomado decisiones diferentes si el nivel de significancia del test hubiese sido del 1%.

(11.) Supongamos que tenemos $n$ observaciones y queremos juzgar el par de hipótesis $H_0 : \theta = 0 \;\; vs. \;\; H_a : \theta \neq 0$

Supongamos que la región de confianza viene dada por el conjunto de valores $\left\{0, 1, \frac{n(n+1)}{2} - 1, \frac{n(n+1)}{2}\right\}$.

Notemos que $T^+$ puede reescribirse como un producto punto entre 2 vectores:

$$T^+ = \begin{bmatrix} 1 & 2 & \cdots & n \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = [1:n]\boldsymbol{\psi}$$

Cada posible vector $\psi$ tiene una probabilidad de $\frac{1}{2^n}$. Para que $T^+$ tome el valor 0, todos los componentes de *psi* deben ser 0 y para que tome el valor 1 el primer componente de $\psi$ debe ser 1. De igual forma, para que tome el valor $\frac{n(n+1)}{2}$ todos los componentes de $\psi$ deben ser 1 y para que tome el valor $\frac{n(n+1)}{2} - 1$, el primer componente de $\psi$ debe ser 0 y el resto 1.

Por ende $\alpha = P\left[T \leq 1 \;\; o \;\; T \geq \frac{n(n+1)}{2} - 1\right] = P\left[T \in \left\{0, 1, \frac{n(n+1)}{2} - 1, \frac{n(n+1)}{2}\right\}\right] = \frac{4}{2^n} = \frac{1}{2^{n-2}}$

b. Test del Signo: Ejercicios **47** y **51**

(47.) Supongamos que $F_1 = \cdots = F_{20} = F$. Además se tiene que $F(0) = 0.3$. Queremos comparar el par de hipótesis $H_0 : \theta = 0 \;\; vs. \;\; H_a : \theta > 0$

Bajo $H_0$ se tiene que $B \sim binom(n = 20, p = 0.5)$, luego el valor que nos a la región de confianza viene dado por:

```
## Ajuste de hiperparámetros
alpha <- 0.0577
n <- 20
p0 <- 0.5
pa <- 1 - 0.3

## Cálculo del cuantil
t_c <- qbinom(prob = p0, size = n, p = alpha,
              lower.tail = FALSE) + 1
t_c
```

```
## [1] 14
```

Como $F(0) = 0.3$ entonces, bajo $H_a$, $B \sim pbinom(n = 20, p = 1 - 0.3)$ y por ende el cálculo de la potencia viene dado por $P_a\left[B \geq 14\right]$

```
pbinom(q = t_c - 1, prob = pa, size = n, lower.tail = FALSE)
```

```
## [1] 0.6080098
```

(51.) Hacemos las mismas suposiciones que en el punto anterior. Basándonos en la aproximación, tenemos que el cuantil que nos da la región de rechazo viene dado por:

```
## Cuantíl que nos da la región de rechazo
( z_q <- qnorm(alpha, lower.tail = FALSE) )
```

```
## [1] 1.574378
```

Denotemos por $z_q$ al cuantil anterior y por $p_a$ el valor de la hipótesis alterna, es decir $p_a = P[X - \theta > 0]$. Recordemos que el estadístico construido (bajo $H_0$) viene dado por:

$$Z = \frac{B - np_0}{\sqrt{np_0(1 - p_0)}}$$

Por ende, la potencia, que es rechazar la hipótesis nula dado que es falsa viene dada por:

$$\begin{aligned}
1 - \beta &= P\left[Z > z_q\right] \\
&= P\left[\frac{B - np_0}{\sqrt{np_0(1 - p_0)}} > z_q\right] \\
&= P\left[B > z_q\sqrt{np_0(1 - p_0)} + np_0\right] \\
&= P\left[B - np_a > z_q\sqrt{np_0(1 - p_0)} + n(p_0 - p_a)\right] \\
&= P\left[\frac{B - np_a}{\sqrt{np_a(1 - p_a)}} > \frac{z_q\sqrt{np_0(1 - p_0)} + n(p_0 - p_a)}{\sqrt{np_a(1 - p_a)}}\right] \\
&= 1 - \Phi\left(\frac{z_q\sqrt{np_0(1 - p_0)} + n(p_0 - p_a)}{\sqrt{np_a(1 - p_a)}}\right)
\end{aligned}$$

Numéricamente:

```
1 - pnorm((z_q*sqrt(n*p0*(1 - p0)) + n*(p0 - pa)) /
            sqrt(n*pa*(1 - pa)))
```

```
## [1] 0.5925124
```

Al comparar, vemos que las potencias en ambos casos son muy parecidas, mas o menos del 60%.

c. Test del signo, muestras pareadas:

d. Test de Wilcoxon dos muestras pareadas: **8** y **13**

(8.) Observamos $X_1 = 2.1$, $X_2 = 1.9$, $X_3 = 2.6$, $X_4 = 3.3$, $Y_1 = 1.9$, $Y_2 = 2.6$ y $Y_3 = 3.7$.

A continuación mostramos la distribución de $W$, es decir $P[W \geq w]$

```r
# Toma de los datos
x <- c(2.1, 1.9, 2.6, 3.3)
y <- c(1.9, 2.6, 3.7)

# Unión de las muestras
juntas <- c(x, y)
r <- rank(juntas)

# Cálculo del estadístico
( w_c <- sum(rep(c(0, 1), c(4, 3)) * r) )
```

```
## [1] 13
```

```r
# Cálculo de todos los valores posibles
W <- table(combn(x = r, m = 3, FUN = sum))
N <- length(W)
P_w <- sapply(1:N, function(k) sum(W[k:N])) / choose(7, 3)
names(P_w) <- names(W)

# P[W >= w]
round(P_w, 3)
```

```
##     6   7.5     9    10  10.5  11.5    12    13  13.5  14.5    15    16  17.5
## 1.000 0.971 0.914 0.771 0.743 0.629 0.571 0.429 0.314 0.257 0.143 0.114 0.057
```

Podemos ver que $P[W \geq 13] = 0.4285714$

(13.) Supongamos que rechazamos $H_0$ si $w_c = \frac{n(2m+n+1)}{2}$ o si $w_c = \frac{n(n+1)}{2}$. Notemos que $w_c = \frac{n(n+1)}{2}$ si en la muestra combinada obtenemos los primeros $n$ números, con probabilidad $\binom{m+n}{n}^{-1}$ o $w_c = \frac{n(2m+n+1)}{2}$ si obtenemos los $n$ números más grandes, i.e. obtenemos $m+1, m+2, \ldots, m+n$.

Por ende

$$w_c = \sum_{k=1}^{n}(m+k) = mn + \frac{n(n+1)}{2} = \frac{2mn + n(n+1)}{2} = \frac{n(2m+n+1)}{2}$$

con probabilidad $\binom{m+n}{n}^{-1}$. Es decir, bajo $H_0$

$$\begin{aligned}
\alpha &= P\left[\frac{n(n+1)}{2} \leq W \text{ ó } W \geq \frac{n(2m+n+1)}{2}\right] \\
&= \frac{1}{\binom{m+n}{n}} + \frac{1}{\binom{m+n}{n}} \\
&= \frac{2}{\binom{n+m}{n}} \\
&= \frac{2}{\frac{(n+m)!}{n!n!}} \\
&= \frac{2n!n!}{(n+m)!}
\end{aligned}$$

e. Test de Kruskal-Wallis: ejercicios **6** y **7**

(6.) Supongamos que $k = 4$ y $n_1 = n_2 = n_3 = 1$ y $n_4 = 2$. Veamos un caso particular de la expresión $\sum_{i=1}^{4} \frac{R_j^2}{n_j}$.

Supongamos que la asignación de los rangos es: $(4), (2), (5), (1, 3)$. Luego esta suma viene dada por:

$$\begin{aligned}
\sum_{i=1}^{4} \frac{R_j^2}{n_j} &= 4^2 + 2^2 + 5^2 + \frac{(1+3)^2}{2} \\
&= 2^2 + 4^2 + 5^2 + \frac{1^2 + 3^2 + 2 \cdot (1 \cdot 3)}{2} \\
&= 2^2 + 4^2 + 5^2 + \frac{1^2 + 3^2}{2} + \frac{2 \cdot (1 \cdot 3)}{2} \\
&= 2^2 + 3^2 + 5^2 + (1^2 + 3^2) - \frac{1^2 + 3^2}{2} + \frac{2 \cdot (1 \cdot 3)}{2} \\
&= \sum_{i=1}^{5} i^2 - \frac{1^2 + 3^2 - 2 \cdot (1 \cdot 3)}{2} \\
&= \sum_{i=1}^{5} i^2 - \frac{(1-3)^2}{2}
\end{aligned}$$

Y esto en general para cualquier par de valores. Luego, podemos calcular todos los posibles valores del estadístico sumando los cuadrados de los 5 primeros números naturales y restándole la diferencia al cuadrado de todos los posibles conjuntos de dos elementos. Con esto en mente, llevamos a cabo el código

```
N <- 5                                  # Cantidad de individuos
suma <- N*(N + 1)*(2*N + 1)/6           # Suma de los primeros 5 naturales cuadrados

# Conjuntos de tamaño 2
n4 <- combn(x = 1:N, m = 2,
            FUN = function(k) sum((k[1] - k[2])^2/2))

# Posibles valores
R2 <- suma - n4

# Valores del estadístico
H <- 12/(N*(N + 1))*R2 - 3*(N + 1)

# Frecuencias
f_h <- table(H)
```

```r
names(f_h)[1] <- round(as.numeric(names(f_h)[1]), 1)

# P[H >= h]
P_h <- sapply(1:length(f_h),
              function(x) sum(f_h[x:length(f_h)])) / choose(5, 2)
names(P_h) <- names(f_h)
P_h
```

```
## 0.8 2.2 3.2 3.8
## 1.0 0.9 0.7 0.4
```

(7.) Supongamos que $k = 3$, $n_1 = n_2 = n_3 = 2$ con repeticiones. Para el cálculo de la distribución, necesitamos calcular todas las posibles particiones de un conjunto de 6 elementos con subconjuntos de tamaño 2.

```r
# Datos
X <- c(2.7, 3.4, 2.7, 4.5, 4.9, 2.7)
r <- rank(X)
N <- 6

# Particiones en tamaños de conjuntos 2
( A <- partitions::setparts(c(2, 2, 2)) )
```

```
##
## [1,] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [2,] 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2
## [3,] 3 3 2 2 2 2 1 1 1 3 3 2 3 3 2
## [4,] 3 2 3 3 3 2 3 2 3 1 1 1 3 2 3
## [5,] 2 3 3 3 2 3 3 3 2 3 2 3 1 1 1
## [6,] 1 1 1 2 3 3 2 3 3 2 3 3 2 3 3
```

```r
# Cálculo:
R <- NULL

for(i in 1:ncol(A)){
    sum(r[A[, i] == 1])^2 +
    sum(r[A[, i] == 2])^2 +
    sum(r[A[, i] == 3])^2 -> R[i]
}

# H
h <- 12/(N*(N + 1))*(R/2) - 3*(N + 1)
H <- table(h)
P_h <- sapply(1:length(H),
              function(x) sum(H[x:length(H)])) / ncol(A)
names(P_h) <- round(as.numeric(names(H)), 3)

# P[H >= h]
P_h
```

```
## 0.286     2 2.571 3.714
##   1.0   0.6   0.4   0.2
```

6. Supongamos que $X \sim \mathcal{N}(\theta, \sigma^2)$. Deseamos probar $H_0 : \theta = 0$ *vs.* $H_a : \theta > 0$ y disponemos de una muestra de tamaño $n = 20$

- Sabemos que $Z = \dfrac{B - 20/2}{\sqrt{20/4}} = \dfrac{B - 10}{\sqrt{5}} \overset{\text{aprox}}{\sim} \mathcal{N}(0, 1)$. Luego el cuantil que define la región de rechazo

viene dado por $z_{1-\alpha}$. Con lo que podemos establecer que

$$\begin{aligned}\alpha &= P_0\left[Z > z_{1-\alpha}\right] \\ &= P_0\left[\frac{B-10}{\sqrt{5}} > z_{1-\alpha}\right] \\ &= P_0\left[B > z_{1-\alpha}\sqrt{5}+10\right]\end{aligned}$$

Y por ende $K = z_{1-\alpha}\sqrt{5}+10$

- Suponiendo que la hipótesis alterna $p = p_a$ es cierta, podemos ver que

$$\begin{aligned}1-\beta &= P_a[B > K] \\ &= P_a\left[B - 20p_a > K - 20p_a\right] \\ &= P_a\left[\frac{B-20p_a}{\sqrt{20p_a(1-p_a)}} > \frac{K-20p_a}{\sqrt{20p_a(1-p_a)}}\right] \\ &= P_a\left[Z > \frac{K-20p_a}{\sqrt{20p_a(1-p_a)}}\right] \\ &= 1 - P_a\left[Z \leq \frac{K-20p_a}{\sqrt{20p_a(1-p_a)}}\right] \\ &= 1 - \Phi\left(\frac{K-20p_a}{\sqrt{20p_a(1-p_a)}}\right)\end{aligned}$$

7. Supongamos que vamos a probar $H_0 : \theta = 0 \;\; vs. \;\; H_a : \theta > 0$ y tenemos una muestra de tamaño 20. Supongamos adicionalmente que no hay empates.

   Bajo esta configuración, se tiene que $E\left[T^+\right] = \frac{20(20+1)}{4} = 105$ y $V[T^+] = \frac{n(n+1)(2n+1)}{24} = 717.5$

- Sabemos que $Z = \dfrac{T^+ - 105}{\sqrt{717.5}} \overset{\text{aprox}}{\sim} \mathcal{N}(0,1)$. Luego el cuantil que define la región de rechazo viene dado por $z_{1-\alpha}$. Con lo que podemos establecer que

$$\begin{aligned}\alpha &= P_0\left[Z > z_{1-\alpha}\right] \\ &= P_0\left[\frac{T^+-105}{\sqrt{717.5}} > z_{1-\alpha}\right] \\ &= P_0\left[T^+ > z_{1-\alpha}\sqrt{717.5}+105\right]\end{aligned}$$

Y por ende $K = z_{1-\alpha}\sqrt{717.5}+105$

- Suponiendo que la hipótesis alterna $p = p_a$ es cierta, podemos ver que

$$
\begin{aligned}
1-\beta &= P_a[T^+ > K] \\
&= P_a\left[T^+ - 210p_a > K - 20p_a\right] \\
&= P_a\left[\frac{T^+ - 210p_a}{\sqrt{2870p_a(1-p_a)}} > \frac{K-210p_a}{\sqrt{2870p_a(1-p_a)}}\right] \\
&= P_a\left[Z > \frac{K-210p_a}{\sqrt{2870p_a(1-p_a)}}\right] \\
&= 1 - P_a\left[Z \leq \frac{K-210p_a}{\sqrt{2870p_a(1-p_a)}}\right] \\
&= 1 - \Phi\left(\frac{K-210p_a}{\sqrt{2870p_a(1-p_a)}}\right)
\end{aligned}
$$

# Maestría

15. **Demostración:** bajo la hipótesis nula, sabemos que $E[\bar{R}_j] = \frac{N+1}{2}$ para todo $j = 1, 2, \ldots, k$.

Definamos $T_j = \bar{R}_j - E[\bar{R}_j]$. Como $\sum_{j=1}^{k} R_j = N$, podemo ver que $R_k = N - \sum_{j=1}^{K-1} R_j$ y por ende vemos que solo $k-1$ de los $R_j$ son independientes. Si definimos el vector $\mathbf{T}$ como

$$
\mathbf{T} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_{k-1} \end{bmatrix}
$$

Vemos que a medida que $\min\{n_1, \ldots, n_k\} \to \infty$ el vector (con tamaño $k-1$) $\mathbf{T}$ tendrá una distribución normal multivariada con media $\mathbf{0}$ y matriz de varianzas y covarianzas $\mathbf{\Sigma}$.

Como el estadístico $H$ es una forma cuadrática del vector $\mathbf{T}$ entonces $H \sim \chi_{k-1}$ siempre que $\min\{n_1, \ldots, n_k\} \to \infty$