Capítulo 1

Pruebas no Paramétricas

En este capítulo abordaremos las pruebas no parámetricas, estas pruebas tiene una ventaja muy importante sobre las pruebas tradicionales pues no tienen el supuesto de que la población de donde se obtiene la muestra sea de una familia paramétrica. Sin embargo, el precio que se paga por llevar a cabo este tipo de pruebas es que muchas veces los métodos no tendrán el nivel de significancia deseado y lo más importante, serán menos potentes que su versión paramétrica.

En este curso veremos 2 tipos de pruebas:

- Pruebas basadas en la distribución Binomial.
- Pruebas basadas en Rango.

1.1. Pruebas basadas en la distribución Binomial

Estas pruebas son llamadas Binomiales porque la distribución del estadístico de prueba que se utiliza para contrastar la hipótesis sigue una distribución Binomial completamente conocida bajo H_0.

1.1.1. Prueba para proporciones

Imaginemos que tenemos $X_1 \ldots, X_n$ m.a. de un fenómeno aleatorio que sólo admite dos posibilidades $X_i \in C_1$ con probabilidad p o bien $X_i \in C_2$ con probabilidad $1 - p$, donde p es el parámetro que nos indica con qué proporción (o probabilidad) observaremos $X_i \in C_1$.

Prueba de dos colas

Se plantea entonces la siguiente hipótesis:

$$H_0 : p = p^* \quad vs \quad H_1 : p \neq p^*$$

Se define el estadístico de prueba:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \in C_1)} = \#\ \textbf{de observaciones en}\ C_1$$

Entonces bajo H_0 se sabe que al ser T suma de v.a. Bernoulli, se tiene que $T \sim Bin\,(n, p^*)$, entonces si H_0 es cierta se espera que T tome valores en la parte densa de la densidad binomial, luego entonces sabemos que debemos rechazar H_0 tanto si T toma valores muy pequeño como muy grandes, es decir la regla que se plantea es **Rechazar** H_0 a un nivel de significancia α si:

$$T \leq w_{\alpha_1} \quad o \quad T > w_{1-\alpha_2}$$

Donde $\alpha_1 + \alpha_2 = \alpha$. En este caso debido a que la distribución es discreta, difícilmente lograremos que la prueba tenga exactamente significancia igual a α, es por eso que la prueba se ajusta a encontrar los cuantiles tales que $\alpha_1 + \alpha_2 = \alpha^* \leq \alpha$ donde α^* es la probabilidad de cometer el error tipo 1 que más se acerque **por abajo** de α.

Observe que en este caso no se especifica cómo encontrar α_1 y α_2, en caso de que la distribucón T bajo H_0 sea simétrica ($p^* = 1/2$) es fácil definir $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, sin embargo esta idea puede no ser la ideal cuando la distribución es muy asimétrica. Para solucionar esto se propone métodos que encuentren α_1 y α_2 tal que el intervalo formado por $(w_{\alpha_1}, w_{1-\alpha_2})$ sea de longitud minima sujeto a $\alpha_1 + \alpha_2 = \alpha$.

Consideremos el siguiente ejemplo:

Supongamos que tenemos un $X_1, \ldots, X_{10}$ m.a. de un fenómeno aleatorio que sólo admite dos valores tal que $\mathbb{P}\,(X_i \in C_1) = p$.

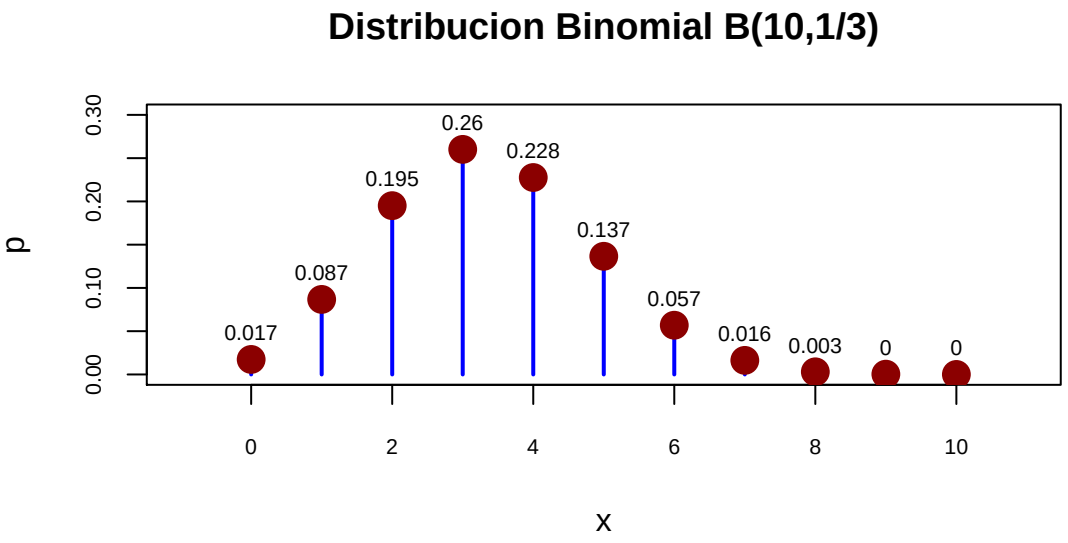
Se plantea la hipótesis

$$H_0 : p = \frac{1}{3} \quad vs \quad H_1 : p \neq \frac{1}{3}$$

Entonces bajo H_0 se tiene que T sigue una distribución Binomial de parámetros $n = 10$ y $p = \frac{1}{3}$.

```
n=10
p=1/3
x=0:10
p=dbinom(x,size=n,prob=p)
```

```r
plot(x,p,type="h",xlim=c(-1,11),ylim=c(0,0.3),lwd=2,col="blue",ylab="p",
     main="Distribucion Binomial B(10,1/3)",cex.axis=0.7)
points(x,p,pch=16,cex=2,col="dark red")
text(x,p,round(p,3),pos=3,cex=0.7)
```



Supongamos que nos piden rechazar un nivel de significancia α, entonces encontraremos α_1 y α_2 de la siguiente forma.

La moda la distribución se obtiene con $T = 3$, al acumular 0.26 de probabilidad, luego el siguiente más grande quitando a $T = 3$, es $T = 4$ al contar con una probabilidad puntual de 0.228 lo cual acumula, junto con el paso anterior, una probabilidad de $0.26+0.228 = 0.488$. Continuamos este proceso hasta que la probabilidad acumulada sobrepase por primera vez a la probabilidad $1 - \alpha$, en nuestro ejemplo como $\alpha = 0.05$, el algoritmo se detiene hasta que acumulemos *por primera vez* más de 0.95. En este caso el algoritmo se detiene cuando $T \in \{1, \ldots, 6\}$, en este caso

$$\mathbb{P}(T \in \{1, \ldots, 6\}) = 0.964$$

Por lo tanto se escoge $\alpha_1 = \mathbb{P}(T \in \{0\}) = 0.017$ y $\alpha_2 = \mathbb{P}(T \in \{7, 8, 9, 10\}) = 0.019$. Entonces $w_{\alpha_1} = 0$ y $w_{1-\alpha_2} = 6$ Por lo tanto rechazamos H_0 si

$$T \leq 0 \quad o \quad T > 6$$

En este caso la prueba tendrá una significancia de $\alpha_1 + \alpha_2 = 0.017 + 0.019 = 0.36$. En caso de que se requiera una prueba exactamente al 5 % existe una forma de llevar a cabo dicho contraste por medio de una **prueba aleatorizada** (no se ve en este curso).

Intervalo de Confianza para la proporción

Una de las ventajas de esta prueba es que es posible encontrar intervalos de confianza para la proporción p, para ello recordemos que hay una relación entre un intervalo de confianza y una prueba de dos colas, en efecto, si por ejemplo hacemos la prueba paramétrica para la media de una Normal:

$$H_0 : \mu = \mu_0 \quad vs \quad \mu \neq \mu_0$$

Entonces una forma de encontrar la región de rechazo es construir un intervalo de confianza para μ y luego verificar si μ_0 se encuentra en dicho intervalo. Visto de forma inversa, ahora se plantea encontrar un intervalo a partir de la regla de rechazo que genera el contraste de hipótesis.

En nuestro caso, debemos preguntarnos, ¿Para qué valores de p^* **no se rechaza** la hipótesis $p = p^*$?. Para encontrar dichos valores se propone ir **barriendo** los distintos valores de p^* (discretizando el intervalo (0,1)) e ir verificando para cuales valores no se rechaza H_0, los p^* que tengan esta propiedad formarán un intervalo de confianza. Observe que en este caso, una vez observada la muestra T es una valor fijo y lo que va variando es p^* lo que a su vez va modificando la distribución asociada.

Veamos un ejemplo, supongamos nuevamente que $n = 10$ y que observamos una muestra tal que $T = 3$, en este caso haremos uso de la función *binom.test* del paquete R, dicha función hace la prueba extacta basada en la distribución binomial y construye el intervalo de confianza correspondiente, en el siguiente código se hace la prueba:

$$H_0 : p = \frac{1}{3} \quad vs \quad H_1 : p \neq \frac{1}{3}$$

```
alpha <- 0.05
binom.test(3,10,1/3,alternative=c("two.sided"),conf.level=1-alpha)
```

```
##
##  Exact binomial test
##
## data:  3 and 10
## number of successes = 3, number of trials = 10, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.3333333
## 95 percent confidence interval:
##  0.06673951 0.65245285
## sample estimates:
## probability of success
##                    0.3
```

En este caso el intervalo al 95 % para la proporción p que construye la función es el siguiente:

$$(0.06673951, 0.65245285)$$

Prueba de una cola

Supongamos ahora que sólo estamos interesados en pruebas de una cola, es decir nos interesa probar:

$$H_0 : p = p^* \qquad vs \qquad H_1 : p > p^*$$
$$H_0 : p \leq p^* \qquad vs \qquad H_1 : p > p^*$$

En este caso ahora nos interesa ver si tenemos evidencia como para afirmar que la verdadera proporción p es más grande que la que proponemos bajo H_0, resulta natural entonces que debemos rechazar H_0 si en la muestra observamos muchas observaciones de la clase 1, es decir, ahora estaremos rechazando si:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \in C_1)} > w_{1-\alpha}$$

Donde $w_{1-\alpha}$ es el cuantíl $1-\alpha$ de una distribución $Binomial(n, p^*)$

Supongamos el mismo ejemplo anterior pero llevado al caso de una cola, es decir nos interesa verificar la siguiente prueba de hipótesis:

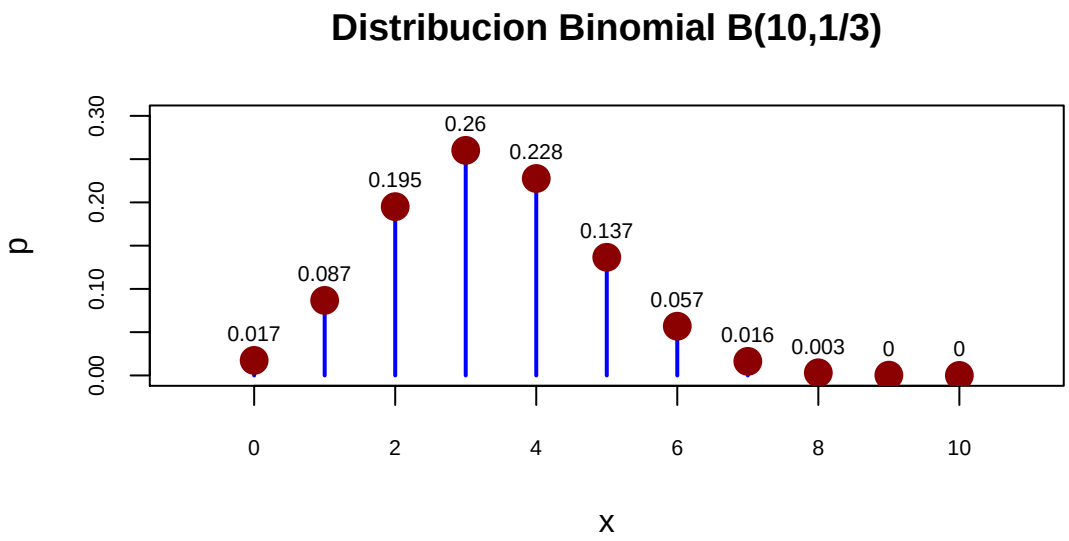
$$H_0 : p = \frac{1}{3} \qquad vs \qquad H_1 : p > \frac{1}{3}$$
$$H_0 : p \leq \frac{1}{3} \qquad vs \qquad H_1 : p > \frac{1}{3}$$

En este caso, dado que nos interesa que la cola derecha acumule α de probabilidad, ahora iremos acumulando probabilidades de derecha a izquierda hasta acumular la probabilidad deseada, como sabemos no necesariamente seremos capaces de acumular exactamente α, en cuyo caso debemos detener el proceso de acumulación hasta que sobrepasemos el α deseado y regresar al paso anterior.

Revisando nuevamente la distribución bajo H_0 del estadístico de prueba tenemos:

```
n=10
p=1/3
x=0:10
p=dbinom(x,size=n,prob=p)
plot(x,p,type="h",xlim=c(-1,11),ylim=c(0,0.3),lwd=2,col="blue",ylab="p",
     main="Distribucion Binomial B(10,1/3)",cex.axis=0.7)
```

```
points(x,p,pch=16,cex=2,col="dark red")
text(x,p,round(p,3),pos=3,cex=0.7)
```



Visto en una tabla

```
a<-as.data.frame(cbind(x,round(p,4)))
colnames(a)<-c("T","Pr")
a
```

```
##      T     Pr
## 1    0 0.0173
## 2    1 0.0867
## 3    2 0.1951
## 4    3 0.2601
## 5    4 0.2276
## 6    5 0.1366
## 7    6 0.0569
## 8    7 0.0163
## 9    8 0.0030
## 10   9 0.0003
## 11  10 0.0000
```

En este caso, acumulando de derecha a izquierda obtenemos que cuando llegamos a $T = 6$ la probabilidad acumulada es:

$$\mathbb{P}(T = 10) + \mathbb{P}(T = 9) + \mathbb{P}(T = 8) + \mathbb{P}(T = 7) + \mathbb{P}(T = 6) = 0.0765635$$

Por lo que nos hemos pasado del α deseado, en este caso si sólo acumulamos hasta $T = 7$ obtenemos:

$$\mathbb{P}(T = 10) + \mathbb{P}(T = 9) + \mathbb{P}(T = 8) + \mathbb{P}(T = 7) = 0.0196616$$

En este caso, tenemos que a un $\alpha = 5\,\%$ la prueba más adeacuada es rechazar H_0 si

$$T > 6$$

La prueba en este caso tendría un nivel de significancia del $\alpha = 0.0196616$, por lo que el cuantil $w_{1-\alpha}$ asociado es 6 y por tanto

$$T > w_{1-\alpha} = 6$$

Finalmente, si ahora nos planteamos la hipótesis de la otra cola:

$$\begin{aligned} H_0 : p = p^* \qquad & vs \qquad H_1 : p < p^* \\ H_0 : p \geq p^* \qquad & vs \qquad H_1 : p < p^* \end{aligned}$$

Resulta de manera natural repetir el proceso anterior pero ahora acumulando probabilidades de la cola izquierda y por tanto la idea será rechazar H_0 si

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \in C_1)} \leq w_\alpha$$

Una alternativa adicional a este problema es aproximar la prueba por medio de la distribución normal y asumir que el estadístico bajo H_0 tiene la siguiente propiedad:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \in C_1)} \overset{aprox}{\sim} N\left(p^*, np^*(1-p^*)\right)$$

Y luego utilizar los cuantiles de la distribución normal correpondiente, sin embargo esta aproximación puede no ser muy buena si p^* es cercano a 0 o 1, en cuyo caso se requerirá mucho tamaño de muestra para tener una buena aproximación.

1.1.2. Prueba del cuantil

En este tipo de pruebas estaremos interesados en hacer inferencia para un cuantil especifico de la distribución, la prueba se planeteará para distribuciones continuas sin embargo se puede llevar a cabo para el caso discreto.

Empecemos recordando lo que entendemos por cuantil de una distribución:

Definición 1.1.1 (Cuantil)**.** *Sea X una v.a. continua con función de distribución $F_X(x)$, decimos que x_q es el cuantil q de la v.a. X si:*

$$F_X(x_q) = \mathbb{P}\left(X \leq x_q\right) = q$$

Es decir x_q es el punto en el cual la variable aleatoria X acumula exactamente q de probabilidad.

Prueba de dos colas

En este tipo de pruebas estaremos interesados en verificar si el cuantil q de la distribución de donde proviene la muestra es cierto valor x_q^* conocido, en este caso nos interesa plantear la prueba:

$$H_0 : x_q = x_q^* \quad vs \quad H_1 : x_q \neq x_q^*$$

Nuevamente supondremos que tenemos como entrada $X_1, \ldots, X_n$ una m.a. de la distribución $F_X(x)$ y la idea consiste en proponer un estadístico de prueba que nos ayude a verificar la veracidad de la hipótesis nula. En este caso propondemos el siguiente estadístico de prueba:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x_q^*)}$$

Observe que en este caso T modela el número de observaciones en muestra que son menores o iguales al cuantil propuesto bajo la hipótesis nula. En este caso bajo H_0 tenemos que:

$$\mathbf{1}_{(X_i \leq x_q^*)} \sim Bernoulli(\mathbb{P}\left(X_i \leq x_q^*\right)) = Bernoulli(\mathbb{P}\left(X_i \leq x_q\right)) = Bernoulli(q)$$

Por lo tanto bajo H_0 el estadístico de prueba tiene una distribución completamente conocida:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x_q^*)} \overset{H_0}{\sim} Binomial\left(n, q\right)$$

Enseguida, lo que resulta natural es rechazar H_0 si T toma valores atípicos bajo la distribución bajo H_0 y por tanto rechazaremos H_0 si:

$$T \leq w_{\alpha_1} \quad o \quad T > w_{1-\alpha_2}$$

Donde $\alpha_1 + \alpha_2 = \alpha$. En este caso nuevamente debemos tener las precauciones necesarias debido al problema de discretización de la distribución del estadístico de prueba bajo H_0 para tener la prueba adeacuada para el α deseado.

Intervalo de Confianza

Nuevamente podemos llevar a cabo un proceso para encontrar intervalos de confianza para el cuantil q de la distribución. En este caso el proceso será nuevamente ir variando el valor x_q^* en la hipótesis nula e ir verificando para qué valores no se rechaza H_0, dichos valores formarán el intervalo de confianza correspondiente, en este caso debe observarse que conforme se mueve x_q^* lo que varia es el valor que toma T y no la distribución. Una parte interesante es que x_q^* sólo mueve el valor de T cuando x_q^* es un valor que está en muestra, esto último facilita mucho la búsqueda pues sólo tendremos que estar realizando la pruaba de hipótesis para valores x_q^* que estén en muestra.

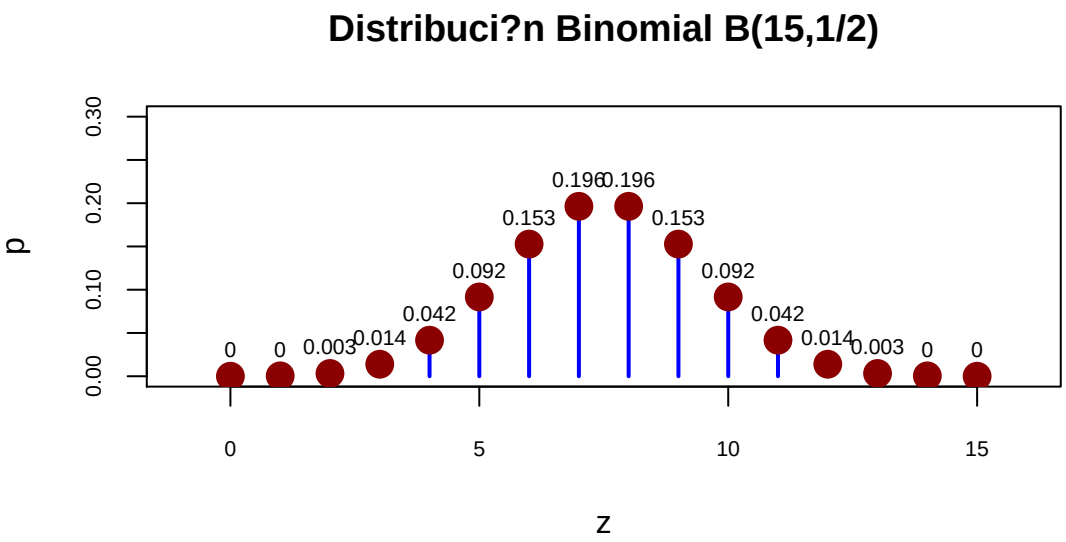
Veamos un ejemplo:

Supongamos que recibimos la siguiente muestra de tamaño 15

$$189, 233, 195, 160, 212, 176, 231, 185, 199, 213, 202, 193, 174, 166, 248$$

Estamos interesados en construir un intervalo para la mediana, entonces lo primeroque procede es ordenar la muestra, obtener la distribución binomial asociada al cuantil 0.5, luego obtener la región de rechazo más cercana al α deseado y luego extraer los estadísticos de orden donde la hipótesis no sea rechaza. Veamos el siguiente código:

```
x<-c(189,233,195,160,212,176,231,185,199,213,202,193,174,166,248)
#ordenamos muestra
x<-sort(x)
#tamao de la muestra
n<-length(x)
#En este caso el estad?stico para probar la mediana es
#T \sim Binom(n,0.5)
 z<-0:15
 p=dbinom(z,size=n,prob=1/2)
 plot(z,p,type="h",xlim=c(-1,16),ylim=c(0,0.3),lwd=2,col="blue",ylab="p",
      main="Distribuci?n Binomial B(15,1/2)",cex.axis=0.7)
 points(z,p,pch=16,cex=2,col="dark red")
 text(z,p,round(p,3),pos=3,cex=0.7)
```



Analizando el gráfico observamos que la zona de rechazo es:

$$T \leq 3 \quad o \quad T > 11$$

Del gráfico también obtemos que el nivel de significancia alcanzado por esta región de rechazo es: α= 0.0351563

Ahora basados en la metodología que planteamos, habrá que ver par qué valores x_q se rechaza y no se rechaza la hipótesis nula, afortunadamente si la muestra fue ordenada, entonces se observa que cuando $x^*_{0.05} = x_{(i)}$ entonces T es definido como el número de observaciones menores o iguales a $x_{(i)}$ tomará el valor de i. Por lo tanto el intervalo de confianza se obtiene fácilmente por medio de los estadísticos de orden que hacen cierta la hipótesis nula, es decir:

$$\left(x_{(4)}, x_{(11)}\right) = (176, 212)$$

Con una confianza del 0.9648438

Prueba de una cola

En este caso debemos tener mucho cuidado sobre la cola que tenemos que analizar del estadístico de prueba.

Supongamos entonces que ahora se plantea la hipótesis:

$$H_0 : x_q = x_q^* \qquad vs \qquad H_1 : x_q > x_q^*$$
$$H_0 : x_q \leq x_q^* \qquad vs \qquad H_1 : x_q > x_q^*$$

Observemos que estamos interesados en ver si la muestra nos da la suficiente evidencia como para rechazar H_0 y decir que $x_q > x_q^*$. La pregunta es entonces, ¿Qué tipo de muestra nos hace pensar que $x_q > x_q^*$?. Nuevamente la idea será contar el número de observaciones menores a x_q^*, si H_1 fuera cierta entonces x_q^* es un punto en donde la distribucion aún no acumula q de probabilidad, es decir se esperaría que el núumero de observaciones menores a x_q^* divido entre n fuera mucho más pequeño que q y por lo tanto esto se traduce en pedir que el estadístico de prueba tenga pocas observaciones, es decir, la idea ahora es rechazar H_0 si:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x_q^*)} \leq w_\alpha$$

Donde w_α es el cuantil α de la distribución $Binomial(n, q)$

De forma análoga, en la prueba para la otra cola es:

$$H_0 : x_q = x_q^* \qquad vs \qquad H_1 : x_q < x_q^*$$
$$H_0 : x_q \geq x_q^* \qquad vs \qquad H_1 : x_q < x_q^*$$

Rechazaremos H_0 si

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x_q^*)} > w_{1-\alpha}$$

Donde, nuevamante $w_{1-\alpha}$ es el cuantil $1 - \alpha$ de la distribución $Binomial(n, q)$.

Estas pruebas deben de tomar las precauciones debidas para encontrar la región más adecuada en función al nivel de significancia deseado.

Finalmente, en caso de que se tenga una muestra suficientemente grande podremos aplicar la aproximación a la normal:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x_q^*)} \overset{aprox}{\sim} N\left(q, nq(1-q)\right)$$

Y por tanto definir la región de rechazo en términos de los cuantiles de la normal apropiada.

1.1.3. Prueba del Signo

Esta prueba pretende comparar la mediana de dos poblaciones. El supuesto principal es que muestreamos de ellas de forma simulatanea, es decir que al momento de obtener la muestra extraemos un vector formado por las variables (X_i, Y_i) donde X_i es el valor de la variable en la primera población, mientras que Y_i es el valor de la variable en la segunda población.

Un ejemplo de esta situación de muestreo puede darse en un experimento donde se aplica mediciones antes y después de un tratamiento especifico a un mismo objeto.

Imaginemos que existe un nuevo método de afinación de un motor y se pretende evaluar si el tratamiento es efectivo, para ello se toman 10 automóviles, primero se les mide su nivel de contaminación (previo a la afinación) y guardamos dichos datos en la variable X, posteriomente se lleva a cabo el tratamiento (afinación) y al mismo coche se le hace la misma prueba y guardamos su nivel de contaminación en la variable Y, entonces al final obtendriamos 10 parejas de observaciones formadas por las mediciones de contaminacón de los autos. En este problema estaríamos interesados en probar si estadísticamente el tratamiento es efectivo, para ello podríamos suponer que X y Y son poblaciones con distribuciones no necesariamente iguales en las que nos interesa probar si Y tiende a tomar valores más pequeños que X, esto lo podríamos plantear enterminos de una medida de tendencia central como lo es la mediana:

$$H_0 : Med(X) = Med(Y) \quad vs \quad H_1 : Med(X) \neq Med(Y)$$

O bien

$$\begin{aligned} H_0 : Med(X) \leq Med(Y) \qquad vs \qquad H_1 : Med(X) > Med(Y) \\ H_0 : Med(X) \geq Med(Y) \qquad vs \qquad H_1 : Med(X) < Med(Y) \end{aligned}$$

En la primera prueba hipótesis estamos interesados en verificar si hay efecto (Positivo o Negativo) del tratamiento en la población mientras que las otras hipótesis sólo nos interesa verificar el efecto únicamiente positivo o negativo de la prueba, en nuestro ejemplo de autos, nos interesa medir un efecto negativo es decir que realmente disminuye contaminación del automóvil, en ese caso se tiene interés en la segunda prueba de hipótesis.

Un **supuesto** adicional que asume la prueba es que la diferencia entre las medianas de X y Y es igual a la mediana de la diferencia esto es, suponiendo que $Z = Y - X$, entonces:

$$Med(Z) = Med(Y - X) = Med(Y) - Med(X)$$

Este último supuesto es escencial para definir el estadístico de prueba para el problema que se plantea.(Ver *The Difference Between the Median of a Difference and the difference of the Medians* de Nigel F. Nettheim)

Pruebas de dos colas

Suponga que se plantea lo siguiente:

$$H_0 : Med(X) = Med(Y) \quad vs \quad H_1 : Med(X) \neq Med(Y)$$

Se supone entonces que recibimos una muestra **bivariada** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, luego contruyamos la v.a Z en función de X y Y como:

$$Z = Y - X$$

Entonces la muestra bivariada es tranformada en una muestra univariada $Z_1 = Y_1 - X_1, \ldots, Z_n = Y_n - X_n$, luego bajo H_0 y los supuestos se tiene que:

$$Med(Z) = med(Y - X) = Med(Y) - Med(X) \overset{H_0}{=} 0$$

Entonces por lo anterior, la prueba de hipótesis se transforma en:

$$H_0 : Med(Z) = 0 \quad vs \quad H_1 : Med(Z) \neq 0$$

Este último problema ya fue resuelto pues no es más que la prueba del cuantil para $q = 0.5$. El estadístico utilizado en esta prueba vimos que es:

$$\begin{aligned} T &= \sum_{i=1}^{n} \mathbf{1}_{(Z_i \leq 0)} = \textbf{\# de observaciones menores o iguales a 0} \\ &= \textbf{\# de signos negativos en la resta } Y_i - X_i \end{aligned}$$

La última igualdad se debe a que suponemos que por continuidad de X y Y se tiene que $\mathbb{P}(Z = 0) = 0$. Sin emabrgo, en la práctica se estila utilizar otro estadístico de prueba similar:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(Z_i > 0)} = \textbf{\# de signos positivos en la resta } Y_i - X_i$$

En este caso nuevamente se tendría que bajo H_0 se tiene que $T \sim Binom(n, 0.5)$ y por tanto resulta de manera natural rechazar H_0 a un nivel de significancia α si:

$$T \leq w_{\alpha_1} \quad o \quad T > w_{1-\alpha_2}$$

Donde $\alpha_1 + \alpha_2 = \alpha$. Afortunadamente en este caso tenemos que el estadístico de prueba tiene una distribución simétrica lo que nos permite hacer $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$. No obstante al prueba se debe ajustar al α adecuado debido a la discretización de T.

Pruebas de una cola

En caso de que se esté interesado en probar una sola cola se plantea entonces la hipótesis:

$$H_0 : Med(X) = Med(Y) \qquad vs \qquad H_1 : Med(X) < Med(Y)$$
$$H_0 : Med(X) \geq Med(Y) \qquad vs \qquad H_1 : Med(X) < Med(Y)$$

En este caso se está interesado en verificar si X tiende a tomar valores más pequeños que Y, en este caso entonces si en la muestra observamos que el signo de la diferencia $Z = Y - X$ tiende a tomar postivos, es evidencia para inclinarse por H_1, tener muchos positivos en Z implica entonces tener pocos negativos por lo tanto se propone rechazar H_0 si:

$$T = \sum_{i=1}^{n} \mathbf{1}_{(Z_i > 0)} = \textbf{\# de signos positivos} > w_{1-\alpha}$$

Donde $w_{1-\alpha}$ es el cuantil α de la distribución $Binomial(n, 0.5)$

Por otro lado, ahora estamos interesados en la otra cola entonces la prueba es:

$$H_0 : Med(X) = Med(Y) \qquad vs \qquad H_1 : Med(X) > Med(Y)$$
$$H_0 : Med(X) \leq Med(Y) \qquad vs \qquad H_1 : Med(X) > Med(Y)$$

En cuyo caso, ahora la muestra indica que se rechace H_0 si observa muchos signos negativos en

la variable Z, luego entonces se tiene que analizar la cola izquierda de la distribución.

$$T = \sum_{i=1}^{n} \mathbf{1}_{(Z_i>0)} = \#\ \textbf{de signos positivos} \leq w_\alpha$$

Donde w_α es el cuantil α de la distribución $Binomial(n, 0.5)$. Donde nuevamente se tiene que tomar las consideraciones necesarias para tener la prueba de significancia más cercano a α.

Caso Discreto

La prueba del signo puede ser adapatada al caso en que las variables X y Y son discretas, sin embargo ahora se tiene que tomar en cuenta los posibles empates pues debido a la discretización se puede dar que $\mathbb{P}(X_i = Y_i) > 0$.

La forma en como se adapta la prueba es sencilla, simplemente se propone eliminar todos los empates que hayan aparecido en la muestra y se lleva a cabo la prueba como en el caso continuo, es decir, se procede a contar el número de signos positivos y luego comparar ese resultado con cuantiles de la binomial respectiva donde ahora n es un tamaño de muestra reducido tras eliminar los empates encontrados.

Muchos autores plantean que la solución de eliminar los empates no es **justo** ya que las observaciones con empate en realidad son a favor de la hipótesis nula. Una posible solución que se ha planteado en la prueba de dos colas es **cambiar** los empates por simulaciones de signos generados de una Bernoulli con probabilidad de exito igual a 0.5 y luego llevar a cabo la prueba tradicional, la idea del cambio es favorecer entonces a la hipótesis nula con observaciones que son de esperarse bajo H_0.

1.1.4. Prueba de McNemar

Esta prueba es un caso especial del test del signo trabajado en la sección anterior, la diferencia radica en que esta prueba supone que tanto X como Y son dicotómicas, es decir, que sólo pueden tomar dos posibles valores digamos 0 y 1, debido a este supuesto las observaciones serán entonces parejas de la forma $(0,0), (0,1), (1,0), (1,1)$ y podrán ser tabuladas de la siguiente forma:

X/Y	0	1
0	a=# número de (0,0)	b=# número de (0,1)
1	c=# número de (1,0)	d=# número de (1,1)

Un ejemplo clásico donde se puede aplicar este caso es en el contexto de la política, supongamos que existen 2 candidatos políticos y definamos a X la v.a. que modela el voto hacia alguno

de los candidatos previo a un debate público (tratamiento), definamos entonces que $X = 0$ si se vota por el candidato A y $X = 1$ si se vota por el candidato B. Despúes del debate (tratamiento), se lleva a cabo nuevamente la medición en las personas y ahora la v.a. Y modela el voto despúes de dicho debate. Una pregunta interesante es entonces resolver si el debate logró cambiar de opinion de los votantes.

Tenemos entonces la necesidad de verificar si existe una diferencia entre las medianas de X y Y, sin embargo debido su naturaleza de las variables en realidad lo que interesa es verificar si **después** de la aplicación de un tratamiento hace que la v.a. X cambie su proporción de $1's$. Ahora observemos que debido a la dicotomía de las variables con las que trabajamos, el hecho de que X no cambie su proporción de $1's$ implica que se espera que $\mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 1, Y = 0)$, es decir, la probabilidad de que un votante cambie de A a B es la misma de que cambie de B a A, este supuesto hace que la proporción de votantes no cambie despúes del tratamiento.

En términos de hipótesis planteamos entonces lo siguiente:

$$H_0 : \mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 1, Y = 0) \quad vs \quad H_1 : \mathbb{P}(X = 0, Y = 1) \neq \mathbb{P}(X = 1, Y = 0)$$

En nuestra notación entonces los casos a=# número de (0,0) y d=# número de (1,1) serán considerados empates y por tanto serán eliminados de la prueba, entonces la decisión debe de recaer en los valores observados en b=# número de (0,1) y c=# número de (1,0), observe que en este caso b modela el número de votantes que cambiaron de opinion del candidato A al candidato B mientras que c modela el caso en donde el voto cambió del candidato B al candiato A.

En el contexto de la prueba del signo la pareja (0,1) tiene un signo positivo y por tanto se propondrá utilizar como estadístico de prueba a b

$$T = b = \textbf{ número de observaciones de la forma (0,1)}$$

Dado que los empates ya fueron eliminados, entonces el tamaño de muestra es $n = b + c$ y luego si suponemos H_0 cierta entonces:

$$T \overset{H_0}{\sim} Binomial\left(b + c, \frac{1}{2}\right)$$

y por tanto se rechazará H_0 si T toma valores muy pequeños (Debate a favor del candidato A) o si T toma valores grandes (Debate a favor del candidato B) donde para la regla de decisión se tomarán en cuenta los cuantiles de la distribución Binomial respectiva.

Algunos autores suponen muestras grandes en estos estudios y por tanto no utilizan la dis-

tribucón binomial sino que llevan a cabo la aproximación normal es decir:

$$T \overset{H_0}{\sim} Binomial\left(b+c, \frac{1}{2}\right) \overset{aprox}{\sim} N\left(np, np(1-p)\right) \overset{aprox}{\sim} N\left((b+c)\frac{1}{2}, \frac{b+c}{4}\right)$$

y por tanto rechazar H_0 basado en los cuantiles de la normal asociada. Otros autores deciden estandarizar la Normal y luego elevarla al cuadrado para obtener la distribución χ^2, es decir, se propone el estadístico de prueba:

$$T_1 = \left(\frac{T-(b+c)\frac{1}{2}}{\sqrt{\frac{b+c}{4}}}\right)^2 = \left(\frac{b-(b+c)\frac{1}{2}}{\sqrt{\frac{b+c}{4}}}\right)^2 \overset{aprox}{\underset{H_0}{\sim}} \chi^2_{(1)}$$

Simplificando el estadístico toma la forma:

$$T_1 = \frac{(b-c)^2}{b+c} \overset{aprox}{\underset{H_0}{\sim}} \chi^2_{(1)}$$

Y por tanto se propone rechazar H_0 si T_1 toma un valor más grande que el cuantil $\chi^{2(1-\alpha)}_{(1)}$.

1.1.5. Prueba Cox and Stuart

La prueba Cox and Stuart es utilizada para verificar si los valores que obtenemos en la muestra siguen alguna tendencia conforme se van observando, para ello la prueba supone que tenenmos $X_1, \ldots, X_n$ variables aleatorias independientes pero no necesariamente identicamente distribuidas, de hecho la idea de la prueba es verificar si las variables tienen alguna tendencia o bien son identicamente distribuidas con la misma media.

El método que proponen los autores es simple, con la muestra recibida $X_1, \ldots, X_n$, se define $c = \frac{n}{2}$ si n es par y $c = \frac{n+1}{2}$ si n es impar, luego generamos las parejas:

Si n es par	**Si n es impar**
(X_1, X_{c+1})	(X_1, X_{c+1})
(X_2, X_{c+2})	(X_2, X_{c+2})
$\vdots$	$\vdots$
(X_c, X_n)	(X_{c-1}, X_n)

Lo que se hace entonces es divir la muestra en dos partes y generar las parejas correspondientes donde en caso de ser n impar se pierde una observación, en este caso X_c

Si los datos tienen tendencia positiva entonces se espera observar signos postivos en la pareja (X_i, X_{c+i}), es decir, $X_{c+i} - X_i > 0$, mientras que si no hay tendencia se observaría un número aleatorio de signos positivos y negativos. Por otro lado si los datos tienen tendencia negativa entonces se espera observar que $X_{c+i} - X_i < 0$ (muchos signos negativos).

Prueba de dos Colas

Dada $X_1, \ldots, X_n$ se pretende contrastar:

$$\begin{aligned} H_0 &: \textbf{Los datos no tienen tendencia} \\ H_1 &: \textbf{Los datos tienen tendencia (positiva o negativa)} \end{aligned}$$

Con la muestra generamos las parejas (X_i, X_{i+c}), definimos la estadística:

$$T = \#\ \textbf{de signos positivos en la diferencia}\ (X_{i+c} - X_i)$$

Bajo H_0 se espera ver signos positivos y negativos de forma aleatoria por lo que se tendría:

$$\begin{aligned} T &\overset{H_0}{\sim} Binomial\left(c, \frac{1}{2}\right) \quad \textbf{n par} \\ T &\overset{H_0}{\sim} Binomial\left(c-1, \frac{1}{2}\right) \quad \textbf{n impar} \end{aligned}$$

Nota: En caso de tener empates en las parejas formadas, estas deben de eliminarse y ajustar el parámetro de la binomial correspondiente.

Se propone rechazar H_0 si T toma valores muy pequeños o muy grandes en función de su distribución teórica. Es decir, rechazar H_0 si:

$$T \leq w_{\alpha_1} \quad o \quad T > w_{1-\alpha_2}$$

Donde $\alpha_1 + \alpha_2 = \alpha$ y w_{α_1}, $w_{1-\alpha_2}$ los cuantiles α_1 y $1 - \alpha_2$ correspondientes de la distribución binomial, algo interesante de esta prueba es que en este caso bajo H_0 se tiene una distribución binomial simétrica y por tanto en este caso $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$. Nuevamente se debe enfatizar que se debe de ajustar la significancia de la prueba debido a la discretización del estadístico de prueba.

Prueba de una cola

Dada $X_1, \ldots, X_n$ se pretende contrastar:

$$\begin{aligned} H_0 &: \textbf{Los datos no tienen tendencia positiva} \\ H_1 &: \textbf{Los datos tienen tendencia positiva} \end{aligned}$$

Al igual que en la prueba de dos colas, se generan las parejas (X_i, X_{i+c}) y definimos la estadística:

$$T = \#\ \textbf{de signos positivos en la diferencia}\ (X_{i+c} - X_i)$$

Bajo H_0 se espera ver signos positivos y negativos de forma aleatoria por lo que se tendría que:

$$\begin{aligned} T &\overset{H_0}{\sim} Binomial\left(c, \frac{1}{2}\right) \quad \textbf{n par} \\ T &\overset{H_0}{\sim} Binomial\left(c-1, \frac{1}{2}\right) \quad \textbf{n impar} \end{aligned}$$

Nota: En caso de haber empates en las parejas, estas deben de eliminarse en cuyo caso se debe de ajustar el parámetro c de la Binomial.

Se propone rechazar H_0 si T toma valores muy grandes pues eso implica que hubo muchos signos positivos lo que es a favor de H_1, por lo tanto rechazamos H_0 si:

$$T > w_{1-\alpha}$$

Donde $w_{1-\alpha}$ es el cuantil correspondiente de la distribución binomial.

De igual forma si se pretende probar:

$$\begin{aligned} H_0 &: \textbf{Los datos no tienen tendencia negativa} \\ H_1 &: \textbf{Los datos tienen tendencia negativa} \end{aligned}$$

Ahora se rechaza H_0 si T toma valores muy pequeños. (Muchos signos negativos)

$$T \leq w_{\alpha}$$

1.1.6. Prueba Cox and Stuart para correlación

Existe una modificación natural de la prueba Cox and Stuart para probar correlación entre dos variables.

Supongamos que tenemos una muestra bivariada de variables aleatorias continuas

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

Se desea verificar si existe algún tipo de correlación entre las variables X y Y.

La propuesta es la siguiente, con la muestra recibida, se ordenan las parejas respecto la variable Y generando una muestra ordenada de la forma:

$$\begin{gathered} \left(X_{i_1}, Y_{(1)}\right) \\ \left(X_{i_2}, Y_{(2)}\right) \\ \vdots \\ \left(X_{i_n}, Y_{(n)}\right) \end{gathered}$$

De la muestra bivariada extraemos la muestra $X_{i_1}, \ldots, X_{i_n}$ y **aplicamos la prueba de tendencia a estos datos**, si existe una tendencia positiva eso quiere decir que la muestra $X_{i_1}, \ldots, X_{i_n}$ crece respecto al orden de aparición, sin embargo por construcción la muestra fue ordenada en función de los valores de Y, eso se traduce en que la muestra está correlacionada positivamente. Por el contrario si se observa una tendencia negativa entonces se concluye que hay una correlación negativa entre X y Y. Finalmente, si la prueba de tendencia no es rechazada, eso es equivalente a no encontrar correlación entre las variables.

La prueba Cox and Stuart para correlación no es más que una aplicación de la prueba de tendencia, sin embargo en la literatura existe más pruebas para detectar correlación que son más potentes como por ejemplo la prueba τ de Kendall o ρ de Spearman. (Ver Conover Capitulo 5 Sección 5)

1.2. Pruebas basadas Rangos

Las pruebas basadas en rangos como su nombre lo indica se basan fundamentalmente en asignar Rangos a la muestra recibida, definamos entonces lo que entendemos por rangos de una muestra.

Definición 1.2.1 (Asignación de Rangos a una Muestra)**.** *Supongamos que recibimos $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución $F_X(x)$. Ordenemos la muestra de menor a mayor es decir:*

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

Donde $X_{(1)} = min\{X_1, \ldots, X_n\}$ y $X_{(n)} = max\{X_1, \ldots, X_n\}$. Supongamos que en la muestra no hay empates, es decir siempre ocurre que $X_{(i)} < X_{(i+1)}$ para toda i. Entonces definimos el rango de la muestra ordenada como

$$R(X_{(i)}) = i$$

Cuando existan empates en la muestra ordenada, por ejemplo $X_{(i)} = X_{(i+1)} = \ldots = X_{(i+k)}$ para alguna i, en ese caso el rango asociado a todas estas obervaciones será igual al promedio de los rangos que se les hubiera asignado suponiendo que no habia empate, es decir:

$$R(X_{(i+q)}) = \frac{i + (i+1) + \ldots + (i+k)}{k+1}; \quad q \in \{0, \ldots, k\}$$

En resumen, entonces podemos decir que $R(X_i)$ es el rango asociado a la observación i, y no es más que la posición que tiene X_i en la muestra ordenada.

Veamos un ejemplo, supongamos que observamos la siguiente muestra de tamaño 5:

$$x_1 = 3, \quad x_2 = 6, \quad x_3 = 1, \quad x_4 = 7, \quad x_5 = 9$$

Ordenando la muestra:

$$x_{(1)} = x_3 = 1, \quad x_{(2)} = x_1 = 3, \quad x_{(3)} = x_2 = 6, \quad x_{(4)} = x_4 = 7, \quad x_{(5)} = x_5 = 9$$

Notemos que no hay empates por lo tanto la asignación de los rangos es la siguiente:

$$\begin{aligned} R(x_{(1)}) &= R(x_3) = 1 \\ R(x_{(2)}) &= R(x_1) = 2 \\ R(x_{(3)}) &= R(x_2) = 3 \\ R(x_{(4)}) &= R(x_4) = 4 \\ R(x_{(5)}) &= R(x_5) = 5 \end{aligned}$$

Supongamos ahora que tenemos una muestra con empates:

$$x_1 = 3, \quad x_2 = 6, \quad x_3 = 1, \quad x_4 = 6, \quad x_5 = 6$$

Ordenando la muestra:

$$x_{(1)} = x_3 = 1, \quad x_{(2)} = x_1 = 3, \quad x_{(3)} = x_2 = 6, \quad x_{(4)} = x_4 = 6, \quad x_{(5)} = x_5 = 6$$

Los rangos que se asgnarían bajo el supuesto de que no hay empates es:

$$\begin{aligned} R(x_{(1)}) &= R(x_3) = 1 \\ R(x_{(2)}) &= R(x_1) = 2 \\ R(x_{(3)}) &= R(x_2) = 3 \\ R(x_{(4)}) &= R(x_4) = 4 \\ R(x_{(5)}) &= R(x_5) = 5 \end{aligned}$$

Como tenemos empates en las observaciones $x_{(3)} = x_{(4)} = x_{(5)}$ entonces los rangos para estos casos se calcula como el promedio de las rangos que les fueron asignados, es decir:

$$R(x_{(3)}) = R(x_{(4)}) = R(x_{(5)}) = \frac{3+4+5}{3} = 4$$

Finalmente los rangos asignados a esta muestra son:

$$\begin{aligned} R(x_{(1)}) &= R(x_3) = 1 \\ R(x_{(2)}) &= R(x_1) = 2 \\ R(x_{(3)}) &= R(x_2) = 4 \\ R(x_{(4)}) &= R(x_4) = 4 \\ R(x_{(5)}) &= R(x_5) = 4 \end{aligned}$$

Las pruebas que veremos a continuación tiene como principal proceso asignar rangos a las observaciones de una muestra aleatoria recibida. Surgirá entonces la necesidad de saber como se distribuye la v.a. $R(X_i)$.

Primero notemos que bajo el supuesto de no haber empates se tiene que $R(X_i)$ es una v.a.

discreta que toma valores en el conjunto $\{1, \ldots, n\}$. Surge ahora la pregunta de saber con qué probabilidad tomará cada uno de estos valores. Para ello recordemos que si la muestra recibida es aleatoria de una sola población, entonces los n rangos que se asociarán deberan aparecer también de forma aleatoria de tal manera que se formen $n!$ posibles secuencias todas con la misma probabilidad, por ejemplo, suponiendo que tenemos 3 observaciones, los $3! = 6$ rangos posibles que podríamos obtener son:

x_1	x_2	x_3
1	2	3
1	3	2
2	1	3
2	3	1
3	1	2
3	2	1

Suponemos entonces que todos estos posibles rangos ocurren con la misma probabilidad es decir 1/6.

Ahora, bajo este supuesto surge entonces la pregunta de conocer $\mathbb{P}(R(X_i) = k)$. Para resolver este caso simplemente contemos los casos favorables y dividamos entre los casos totales. Ya sabemos que el total de casos es $n!$, de estos casos, los que son favorables con el evento $R(X_i) = k$ son $(n-1)!$ posibles secuencias pues estamos dejando fija que en la columna i siempre aparezca k. Por lo tanto:

$$\mathbb{P}(R(X_i) = k) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

Es decir, resulta que $R(X_i)$ bajo el supuesto de que tenemos m.a. de **una sola población**, tiene una distribución uniforme discreta en el espacio $\{1, \ldots, n\}$

A continuación presentamos las 4 principales pruebas basadas en Rangos.

1.2.1. Prueba Mann-Whitney

Esta prueba tiene por objetivo verificar si existe diferencias entre las medias (medianas) de dos poblaciones. Los insumos de la prueba entonces serán dos muestras aleatorias de 2 poblaciones de forma independiente, es decir, supondremos que tenemos $X_1, \ldots, X_{n_1}$ m.a. de $F_X(x)$ y $Y_1, \ldots, Y_{n_2}$ m.a. de $G_Y(y)$

Los supuestos que tiene la prueba son los siguientes:

- Las dos muestras recibidas son aleatorias de sus respectivas distribuciones

- Hay independencia entre ambas muestras, es decir $F(\underline{x}, \underline{y}) = F(\underline{x})F(\underline{y})$
- Supondremos que muestreamos de distribuciones continuas, sin embargo la prueba se puede correr para el caso discreto, no obstante en caso de haber muchos empates en la asignación de rangos para las muestras recibidas, ocasionará que la prueba pierda validez.
- Si existe una diferencia entre las distribuciones $F_X(x)$ y $G_Y(y)$ es sólo de localización y no de forma es decir $F_X(x) = G_Y(x + c)$ para toda x y alguna c. Algunos autores asumen que la prueba sigue siendo valida incluso si la forma es distinta.

Prueba de dos colas

La prueba Mann-Whitney pretende entonces contrastar la siguientes hipótesis:

$$H_0 : F_X(x) = G_Y(x) \quad vs \quad H_1 : F_X(x) \neq G_Y(x)$$

Sin embargo, la prueba comunmente se presenta en términos de medidas de tendencia central y se propone:

$$\begin{aligned} H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \qquad & vs \qquad H_1 : \mathbb{E}(X) \neq \mathbb{E}(Y) \\ H_0 : Med(X) = Med(Y) \qquad & vs \qquad H_1 : Med(X) \neq Med(Y) \end{aligned}$$

El método propuesto en la prueba es el siguiente:

- Unir ambas muestras recibidas generando una muestra de tamaño $n = n_1 + n_2$
- Asignar Rangos a la muestra unida, sin olvidar de qué población viene cada observación.
- De los Rangos obtenidos, sólo nos quedamos con los Rangos de la problación de la v.a. X
- Calcular el estadístico de prueba:

$$T = S - \frac{n_1(n_1 + 1)}{2}$$

Donde $S = \sum_{i=1}^{n_1} R(X_i)$, la suma de los rangos asociados a las observaciones de la distribución $F_X(x)$.

La idea que hay detras de este estadístico es la siguiente, si la población con distribución $F_X(x)$ tiende a tomar valores más pequeños que la población $G_Y(x)$ entonces los rangos asociados a las observaciones de X serán pequeños, de hecho el peor caso es cuando **todas** las observaciones

cayeron por debajo de las observaciones de Y, en ese caso los rangos que se asocian a la población X son los primeros n_1 naturales y por tanto:

$$S = \sum_{i=1}^{n_1} R(X_i) = 1 + 2 + \ldots + n_1 = \frac{n_1(n_1+1)}{2} \Rightarrow T = 0$$

Es decir, valores de T pequeños son compatibles con la hipótesis de que $\mathbb{E}(X) < \mathbb{E}(Y)$. El caso opuesto a esto es cuando las observaciones de X estan **todas** por arriba de las observaciones de Y, en ese caso los rangos asociados a las observaciones de X son los naturales

$$n_2 + 1, n_2 + 2, \ldots, n_2 + n_1 = n$$

y por tanto:

$$S = \sum_{i=1}^{n_1} R(X_i) = (n_2+1) + (n_2+2) + \ldots + (n_2+n_1) = n_1 n_2 + \frac{n_1(n_1+1)}{2}$$

Y por tanto, en este caso, el valor de T es:

$$T = S - \frac{n_1(n_1+1)}{2} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \frac{n_1(n_1+1)}{2} = n_1 n_2$$

Luego entonces valores grandes de T (cercanos a $n_1 n_2$) son a favor de la hipótesis $Med(X) > Med(Y)$. Todo indica entonces que hay evidencia para rechazar H_0 tanto si T es pequeño o grande, para tomar la decisión es necesario conocer la distribución de T bajo H_0 para encontrar los cuantiles correspondientes.

La distribución de T bajo H_0 no es fácil de obtener de forma análitica debido a las distintas combinaciones que se pueden dar en los rangos de la muestra unida, sin embargo se puede simular o bien consultar en tablas. (Tabla 8 de Conover). Una ventaja que tiene esta distribución es que es simétrica y por tanto los cuantiles para el rechazo de la hipótesis se obtienen de forma más fácil.

La regla entonces es, **Rechazar** H_0 a un nivel de significancia α si:

$$T \leq w_{\frac{\alpha}{2}} \quad o \quad T > w_{1-\frac{\alpha}{2}}$$

Donde w_α y $w_{1-\frac{\alpha}{2}}$ son los cuantiles de la distribución de T bajo H_0 (Se obtienen por simulación o en tablas)

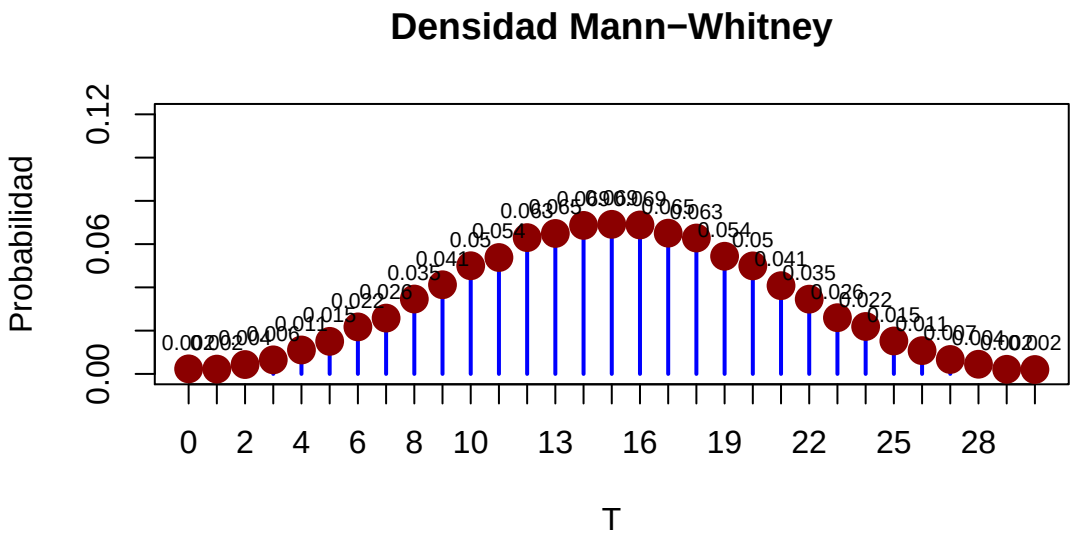
A continuación se presenta un código para simular la distribución Mann-Whitney

```r
#############################################################
#Funcion que calcula la distribucion Mann Whitney Bajo H_0 #
#############################################################
#Tamao de muestra de la primer poblacin
n=5
#Tamao de muestra de la segunda poblacin
m=6

#Tamao de muestra combinada
N=n+m

#Numero de simulaciones
nSim=500000
#Generamos un arreglo de 10,0000 simulaciones
T=rep(0,nSim)

for (i in 1:nSim){
  T[i]=sum(sample(1:N,n))-n*(n+1)/2
}
plot(table(T)/nSim,type="h",lwd=2,col="blue",ylab="Probabilidad",
     main="Densidad Mann-Whitney", xlab="T",ylim=c(0,0.12))
points(0:(n*m), table(T)/nSim,pch=16,cex=2,col="dark red")
text(0:(n*m), table(T)/nSim, round(table(T)/nSim,3), pos=3, cex=0.7)
```



Prueba de una cola

En este caso se propone contrastar alguna de estas hipótesis

Para la cola izquierda:

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \qquad vs \qquad H_1 : \mathbb{E}(X) < \mathbb{E}(Y)$$
$$H_0 : \mathbb{E}(X) \geq \mathbb{E}(Y) \qquad vs \qquad H_1 : \mathbb{E}(X) < \mathbb{E}(Y)$$

En este caso sólo nos interesa ver la cola izquierda de T y por tanto se propone **Rechazar** H_0 a un nivel de significancia α si:

$$T \leq w_\alpha$$

Para la cola derecha:

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \qquad vs \qquad H_1 : \mathbb{E}(X) > \mathbb{E}(Y)$$
$$H_0 : \mathbb{E}(X) \leq \mathbb{E}(Y) \qquad vs \qquad H_1 : \mathbb{E}(X) > \mathbb{E}(Y)$$

En este caso sólo nos interesa ver la cola derecha de T y por tanto se propone **Rechazar** H_0 a un nivel de significancia α si:

$$T > w_{1-\alpha}$$

Aproximación hacia la Normalidad

Debido a la simetría del estadístico de prueba T, existe una aproximación hacia la distribución normal, para ello se debe calcular primero la esperanza y varianza del estadístico de prueba **(TAREA)**:

$$\mathbb{E}(T) = \mathbb{E}\left(S - \frac{n_1(n_1+1)}{2}\right) = \frac{n_1 n_2}{2}$$

$$\mathrm{Var}(T) = \mathrm{Var}(S) = \frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12}$$

Luego entonces asumiendo que:

$$T \overset{aprox}{\sim} N\left(\mathbb{E}(T), \mathrm{Var}(T)\right) = N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12}\right)$$

Por lo tanto si la muestra es suficientemente grande podríamos definir el estadístico:

$$Z = \frac{T - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \overset{aprox}{\sim} N(0,1)$$

Y por lo tanto tomar la decisión del rechazo de H_0 en función de los cuantiles de la normal estándar.

1.2.2. Prueba Kruskal-Wallis

La prueba Kruskal-Wallis no es más que la generalización de la prueba Mann-Whitney para el caso de k poblaciones. En este caso el test pretende verificar si todas las distribuciones son iguales o si existe al menos una población que tiende a tomar valores distintos a los demás.

En este caso supondremos que recibimos k muestras aleatorias independientes de k poblaciones distintas es decir:

Sample 1	Sample 2	...	Sample k
X_{11}	X_{21}	...	X_{k1}
X_{12}	X_{22}	...	X_{k2}
$\vdots$	$\vdots$	...	X_{k3}
X_{1n_1}	X_{2n_2}	...	X_{kn_k}

Donde suponemos entonces que la muestra $\underline{X}_i = (X_{i1}, \ldots, X_{i,n_i})$ proviene de la distribución $F_i(x)$, además observemos que en este caso, el tamaño de cada muestra es n_i por lo que el tamaño de muetra total es $n = \sum_{i=1}^{k} n_i$.

Los supuestos que tiene la prueba son los siguientes:

- Las k muestras recibidas son aleatorias de sus respectivas distribuciones
- Hay independencia entre las k muestras, es decir

$$F(\underline{x}_1, \ldots, \underline{x}_k) = F_1(\underline{x}_1) \ldots F_k(\underline{x}_k)$$

- Supondremos que muestreamos de distribuciones continuas aunque la prueba se puede correr para el caso discreto pero en caso de haber muchos empates en las muestras recibidas ocasiona que la prueba pierda validez

- Si existe una diferencia entre las distribuciones $F_1(x), \ldots, F_k(x)$ es sólo de localización y no de forma es decir para cualesquiera 2 pares de distribuciones $F_i(x), F_j(x)$ se tiene que existe c tal que $F_i(x) = F_j(x+c)$ para toda x. Algunos autores asumen que la prueba sigue siendo valida incluso si la forma es distinta entra las distribuciones.

La prueba Kruskal-Wallis pretende contrastar la hipótesis:

$$H_0 : F_1(x) = \ldots = F_k(x) \quad vs \quad H_1 : F_i(x) \neq F_j(x) \quad p.a. \quad i \neq j$$

Aunque muchos autores proponen mejor utilizar la versión de medias (asumiendo que las medias existen)

$$H_0 : \mathbb{E}(X_1) = \ldots = \mathbb{E}(X_k) \quad vs \quad H_1 : \mathbb{E}(X_i) \neq \mathbb{E}(X_j) \quad p.a. \quad i \neq j$$

El método de la prueba consiste nuevamente en mezclar todas la muestras y formar una sola secuencia de observaciones de tamaño $n = \sum_{i=1}^{k} n_i$, a dicha secuencia le asignamos rangos y entonces la muestra es transformada obteniendo una tabla como sigue:

Sample 1	Sample 2	...	Sample k
$R(X_{11})$	$R(X_{21})$	...	$R(X_{k1})$
$R(X_{12})$	$R(X_{22})$	...	$R(X_{k2})$
$\vdots$	$\vdots$	...	$R(X_{k3})$
$R(X_{1n_1})$	$R(X_{2n_2})$	...	$R(X_{kn_k})$

Antes de proponer el estadístico de prueba, analicemos la variable aleatoria $R(X_{ji})$, primero notemos que dicha v.a. sólo puede tomar valores en el espacio formado por los primeros n naturales $\{1, \ldots, n\}$, la pregunta natural que ahora surge es conocer la probabilidad de que tome cada uno de estos números.

Bajo el supuesto de H_0, sabamos que toda la muestra viene de una sola población por lo que los rangos que se asocian deberian seguir un comportamiento aleatorio similar al proceso de seleccionar muestrar aleatorias sin reemplazo de una población de tamño n, por lo tano utilizando la teoría del muestreo aleatorio simple de una población de tamño n, se puede probar que:

$$\mathbb{P}\left(R(X_{ji}) = q\right) = \frac{1}{n} \quad j \in \{1, \ldots, k\}; \quad i \in \{1, \ldots, n_j\}; \quad q \in \{1, \ldots, n\}$$

Es decir bajo H_0 el Rango que se le asocia a la observación X_{ji} sigue una distribución uniforme sobre el espacio $\{1, \ldots, n\}$. Con lo anterior podemos encontrar la esperanza y varianza de la v.a.

$R(X_{ji})$.

$$\mathbb{E}(R(X_{ji})) = \sum_{q=1}^{n} q\mathbb{P}\left(R(X_{ji}) = q\right) = \sum_{q=1}^{n} q\frac{1}{n} = \frac{1}{n}\left(\frac{n(n+1)}{2}\right) = \frac{n+1}{2}$$

$$\mathbb{E}\big(R(X_{ji})^2\big) = \sum_{q=1}^{n} q^2\mathbb{P}\left(R(X_{ji}) = q\right) = \sum_{q=1}^{n} q^2\frac{1}{n} = \frac{1}{n}\left(\frac{n(n+1)(2n+1)}{6}\right) = \frac{(n+1)(2n+1)}{6}$$

$$\mathrm{Var}(R(X_{ji})) = \mathbb{E}\big(R(X_{ji})^2\big) - \mathbb{E}(R(X_{ji}))^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

Definamos ahora la suma de los rangos asociados a la población j como:

$$R_j = \sum_{i=1}^{n_j} R\left(X_{ji}\right)$$

Entonces si H_0 fuera cierta podríamos obtener la esperanza de la v.a. R_j

$$\mathbb{E}(R_j) = \mathbb{E}\left(\sum_{i=1}^{n_j} R\left(X_{ji}\right)\right) = \sum_{i=1}^{n_j} \mathbb{E}(R\left(X_{ji}\right)) = \sum_{i=1}^{n_j} \frac{n+1}{2} = \frac{n_j(n+1)}{2}$$

De igual forma se puede obtener la varianza de R_j salvo que hay que tener mucho cuidado pues las variables $R\left(X_{ji}\right)$ no son independientes, en ese caso:

$$\mathrm{Var}(R_j) = \mathrm{Var}\left(\sum_{i=1}^{n_j} R\left(X_{ji}\right)\right) = \sum_{i=1}^{n_j} \mathrm{Var}(R\left(X_{ji}\right)) + \sum_{\substack{q=1 \\ q\neq p}}^{n_j}\sum_{p=1}^{n_j} Cov\left(R(X_{jq}), R(X_{jp})\right)$$

Se prueba a partir de esto **(TAREA)** que:

$$\mathrm{Var}(R_j) = \frac{n_j(n+1)(n-n_j)}{12}$$

Sabemos entonces que R_j modela la suma de los rangos asociados a la población j y que se espera bajo H_0 que $\mathbb{E}(R_j) = \frac{n_j(n+1)}{2}$ y que $\mathrm{Var}(R_j) = \frac{n_j(n+1)(n-n_j)}{12}$.

Una forma de verificar la veracidad de H_0, es proponer como estadístico de prueba a la distancia al cuadrado entre el valor observado y esperado de R_j y luego sumarlos sobre todos los j, es decir

$$\sum_{j=1}^{k}(R_j - \mathbb{E}(R_j))^2 = \sum_{j=1}^{k}\left(R_j - \frac{n_j(n+1)}{2}\right)^2$$

Sin embargo la distribución de esta estadística no es facíl, no obstante si suponemos normalidad

con muestras suficientemente grandes podemos encontrar una solución alternativa:

$$\frac{R_j - \mathbb{E}(R_j)}{\sqrt{\text{Var}(R_j)}} = \frac{R_j - \frac{n_j(n+1)}{2}}{\sqrt{\frac{n_j(n+1)(n-n_j)}{12}}} \overset{aprox}{\sim} N(0,1)$$

Entonces:

$$\frac{(R_j - \mathbb{E}(R_j))^2}{\text{Var}(R_j)} = \frac{\left(R_j - \frac{n_j(n+1)}{2}\right)^2}{\frac{n_j(n+1)(n-n_j)}{12}} = \frac{12}{n+1}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j(n-n_j)} \overset{aprox}{\sim} \chi^2_{(1)}$$

Si todas las $R_j's$ fueran independientes habríamos terminado el problema pues se propondría como estadístico de prueba a la suma de las χ^2, es decir:

$$T' = \sum_{j=1}^{k} \frac{12}{n+1}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j(n-n_j)} = \frac{12}{n+1}\sum_{j=1}^{k}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j(n-n_j)}$$

Sin embargo es obvio que **no** podemos asumir que $T' \overset{aprox}{\sim} \chi^2_{(k)}$, pues sabemos que las $R_j's$ son **dependientes**, de hecho $\sum_{j=1}^{k} R_j = \frac{n(n+1)}{2}$.

El gran aporte que hizo Kruskal en 1952 fue probar que si a cada sumando se le multiplica por el término $\frac{n-n_j}{n}$ entonces la suma **sí** tiene una distribución χ^2 pero pierde un grado de libertad es decir:

$$T = \frac{12}{n+1}\sum_{j=1}^{k}\frac{n-n_j}{n}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j(n-n_j)} = \frac{12}{n(n+1)}\sum_{j=1}^{k}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j} \overset{aprox}{\sim} \chi^2_{(k-1)}$$

Finalmente entonces Kruskal en 1952 propone como estadístico de prueba:

$$T = \frac{12}{n(n+1)}\sum_{j=1}^{k}\frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j}$$

Y luego entonces se rechaza H_0 a un nivel de significancia α si $T > \chi^{2(1-\alpha)}_{(k-1)}$, donde $\chi^{2(1-\alpha)}_{(k-1)}$ es el cuantil $1-\alpha$ de la distribución χ^2 con $(k-1)$ grados de libertad.

Se puede probar además que el estadístico de prueba puede transformarse en lo siguiente **(TAREA)**:

$$T = \left(\frac{12}{n(n+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right) - 3(n+1)$$

Un problema que tiene la prueba es que supone muestras grandes para poder asumir una

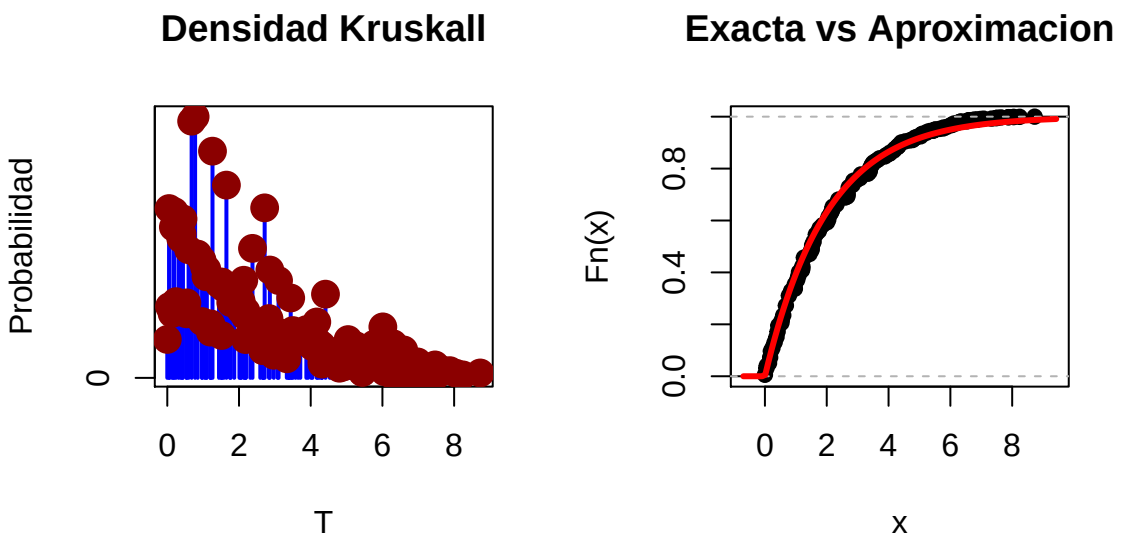
buena aproximación hacia la χ^2, es por eso que existen tablas de la prueba para el caso de que se tienen muestras pequeñas. (Ver tabla 12 del Conover y sólo ataca el caso k=3).

Lo anterior nos motiva a tener programas que nos ayuden a simular la distribución de T bajo H_0, a continuación presentamos un código en R que lleva a cabo la simulación de la distirbución haciendo uso de la funcion *sample* la cual simula precisamente la obtención de los rangos basado en un muestreo aleatorio simple:

```
#############################################################################
#Funcion que calcula los cuantiles  de la distribucion KuskalWallis Bajo H_0 #
#############################################################################
#############################################################################
#El programa esta diseniado para cuando se tiene k=3 muestras independientes #
#############################################################################
k=3
#Introduzca el tamanio de las muestras de cada poblacion
n1=5
n2=3
n3=3


#Numero de simulaciones
nSim=50000

n=c(n1,n2,n3)
N=n1+n2+n3
T<-rep(0,nSim)
x=1:N
for (i in 1:nSim){
  s=sample(x,N)
  R=c(sum(s[1:n1]),sum(s[(n1+1):(n1+n2)]),sum(s[(n1+n2+1):N])  )
  T[i]=12/(N*(N+1))*(sum(R^2/n))-3*(N+1)
}
#######################################
#Graficamos la distribucion exacta   #
#######################################
par(mfrow = c(1, 2))
plot(as.numeric(names(table(T))),table(T)/nSim,type="h",lwd=2,
     col="blue",ylab="Probabilidad",main="Densidad Kruskall", xlab="T")
points(as.numeric(names(table(T))), table(T)/nSim,pch=16,cex=2,col="dark red")
f<-ecdf(T)
plot(f,main="Exacta vs Aproximacion")
curve(pchisq(x,k-1),add=TRUE,col=2,lwd=3)
```



1.2.3. Prueba Wilcoxon

La prueba Wilcoxon es un test similar a la prueba del signo pero tiene la ventaja de ser más potente porque toma en cuenta las magnitudes de las diferencias y no sólo el signo.

Como insumos de la prueba supondremos entonces que recibimos una muestra bivariada de la forma (X_i, Y_i). La prueba del signo sabemos que ataca el problema analizando los signos de la diferencia $Y_i - X_i$, la idea que ahora propone Wilcoxon es no sólo fijarnos en el signo sino también en los rangos de las diferencia.

Prueba de dos colas

Nuevamente estamos interesados en la hipótesis:

$$H_0 : F_X(x) = G_Y(x) \quad vs \quad H_1 : F_X(x) \neq G_Y x$$

O visto en términos de medias (suponiendo que existen)

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \quad vs \quad H_1 : \mathbb{E}(X) \neq \mathbb{E}(Y)$$

Dada $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra bivariada, la metodología es la siguiente:

- Para cada pareja (X_i, Y_i), definir la diferencia $D_i = Y_i - X_i$

- En caso de haber empates ($D_i = 0$), eliminarlos de la muestra bivariada quedandonos con $n^{'}$ observaciones
- Asignar Rangos a la muestra formada por el valor absoluto de las D_i, es decir, obtener $R(|D_1|) \ldots, R(|D_{n'}|)$
- Se define el estadístico de prueba:

$$T = \sum_{i=1}^{n^{'}} R(|D_i|) \mathbf{1}_{(Y_i > X_i)}$$

Observemos que T suma únicamente los rangos de las parejas con signo positivo, es decir, se podría decir que pondera el signo obtenido con el rango del valor absoluto. La prueba entonces no sólo toma en cuenta el signo, sino también la **magnitud** de la diferencia en valor absoluto en cada pareja (X_i, Y_i).

Notemos que si todas las diferencias son negativas $Y_i - X_i < 0$ entonces $T = 0$ lo que iría en contra de la hipótesis nula, mientras que si todos los signos son positivos $Y_i - X_i > 0$ entonces T tomaría el valor de $n^{'}(n^{'} + 1)/2$ lo que contradice H_0. La estadística de prueba tomará valores entre 0 y $n^{'}(n^{'} + 1)/2$ y se debe de rechazar H_0 si:

$$T \leq w_{\alpha_1} \quad T > w_{1-\alpha_2}$$

Donde w_{α_1} y $w_{1-\alpha_2}$ son los cuantiles de la correspondientes de la distribución de T los cuales pueden ser consultados en la tabla 7 del libro de Conover. Afortundamanete dicha distribución es simétrica por lo que se puede tomar $\alpha_1 = \alpha_2 = \alpha/2$, no obstante hay que tener las consideraciones pertinentes debido a la discretización de la estadística T y encontrar el α más cercano que permita la distribución.

La pregunta es, ¿cómo se distribuye T bajo H_0?. Veamoslo con un ejemplo simple, supongamos que tenemos un tamaño de muestra bivariada igual a 3, luego bajo el supuesto de que no hay empates tendriamos 3 posibles rangos a asociar $\{1, 2, 3\}$, dichos rangos sabemos que pueden aparecer en orden aleatorio sin embargo en esta prueba no nos interesa esta parte, de hecho imaginemos que siempre ordenamos la muestra en función de los rangos, la parte importante de la estadística de prueba es que el estadístico sólo sumará Rangos que tengan una diferencia positiva, ahora bien, como suponemos H_0 cierta, se espera que los signos positivos aparezcan con probabilidad $1/2$, bajo este contexto entonces hay 2^3 posibles secuencias que pueden ocurrir (todas con probabilidad $1/2^3$) a continuacióon exihibimos todas las posibilidades con su respectivo valor de la estadística.

1	2	3	T
-	-	-	0
-	-	+	3
-	+	-	2
-	+	+	5
+	-	-	1
+	-	+	4
+	+	-	3
+	+	+	6

Entonces bajo H_0 e la distribución de T es la siguiente:

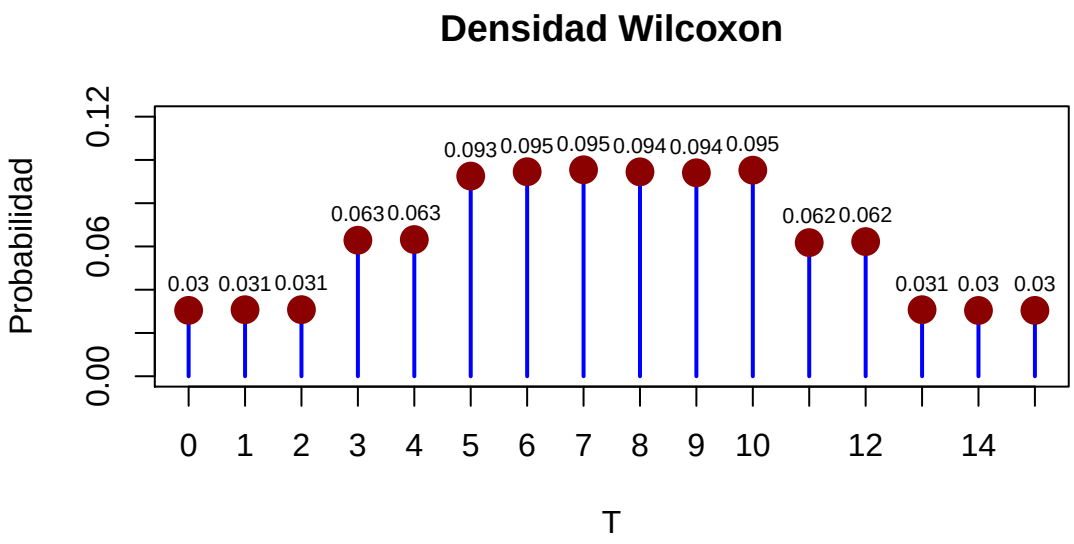
$$\mathbb{P}(T=k) = \begin{cases} \frac{1}{8} \text{ si } k \in \{0,1,2,4,5,6\} \\ \frac{2}{8} \text{ si } k \in \{3\} \\ 0 \text{ e.o.c} \end{cases}$$

Desafortundamente cuando la muestra es grande el número de combinaciones crece de forma importante lo que hace dificil encontrar la distribución exacta. Sin embargo dicha distribucón puede ser simulada. A continuación presetamos el código en R para su simulación:

```
############################################################
#Funcion que calcula   la distirbucion Wicoxon bajo H_0  #
############################################################
#Introduzca el numero de observaciones
n=5

#Numero de Simulaciones
m=100000
t<-rep(0,m)
y=1:n
for( i in  1:m){
   s=sample(c(0,1),n,replace=TRUE)
  t[i]=y%*%as.matrix(s)
}

plot(table(t)/m,type="h",lwd=2,col="blue",ylab="Probabilidad",main="Densidad Wilcoxon", xlab="T",ylim=c(0
points(0:(n*(n+1)/2), table(t)/m,pch=16,cex=2,col="dark red")
text(0:(n*(n+1)/2), table(t)/m, round(table(t)/m,3), pos=3, cex=0.7)
```



Prueba de una cola

En este caso estamos interesado en probar:

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \quad vs \quad H_1 : \mathbb{E}(X) < \mathbb{E}(Y)$$

$$H_0 : \mathbb{E}(X) \geq \mathbb{E}(Y) \quad vs \quad H_1 : \mathbb{E}(X) < \mathbb{E}(Y)$$

En la hipótesis alternativa nos interesa saber si X tiene a tomar valores más pequeños que Y, como $D_i = Y_i - X_i$ entonces ver signos positivos son a favor de H_1, lo lógico entonces es rechazar H_0 si vemos un valor muy grande de T es decir, rechazamos H_0 si:

$$T > w_{1-\alpha}$$

Por otro lado si nos interesa probar:

$$H_0 : \mathbb{E}(X) = \mathbb{E}(Y) \quad vs \quad H_1 : \mathbb{E}(X) > \mathbb{E}(Y)$$

$$H_0 : \mathbb{E}(X) \leq \mathbb{E}(Y) \quad vs \quad H_1 : \mathbb{E}(X) > \mathbb{E}(Y)$$

Ahora ver signos negativos $D_i = Y_i - X_i$ son a favor de H_1 y por tanto lo lógico será rechazar H_0 si T toma valores pequeños por tanto rechazaremos H_0 si:

$$T \leq w_{\alpha}$$

Aproximación a la normalidad

Debido a la simetría que tiene el estadístico de prueba es de esperarse que para muestras grandes se tenga una buena aproximación hacia la normalidad. Para llevar a cabo dicha aproximación es necesario obtener la esperanza y varianza de T.

Se puede probar que **(TAREA)**:

$$\mathbb{E}(T) = \frac{n(n+1)}{4} \quad \text{Var}(T) = \frac{n(n+1)(2n+1)}{24}$$

Por lo tanto la aproximación normal de T es:

$$T \overset{aprox}{\sim} N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

Estandarizando se obtiene el estadístico:

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \overset{aprox}{\sim} N(0,1)$$

Luego entonces, se calcula el estadístico Z y se rechaza en función de los cuantiles de la normal estandar.

1.2.4. Prueba de Friedman

Esta prueba es una generalización la de la prueba de Wilcoxon, el test se propone verificar igualdad de medias en una muestra multivariada, es decir supondemos que recibimos una muestra k-variada de tamaño n.

La muestra en este caso puede escribirse un una tabla de $n \times k$

$F_1(x)$	$F_2(x)$	$\ldots$	$F_k(x)$
X_{11}	X_{21}	$\ldots$	X_{k1}
X_{12}	X_{22}	$\ldots$	X_{k2}
$\vdots$	$\vdots$	$\ldots$	$\vdots$
X_{1n}	X_{2n}	$\ldots$	X_{kn}

En este caso una observación esta dada por el vector $(X_{1i}, X_{2i}, \ldots, X_{ki})$.

Estaremos interesados en verificar la hipótesis:

$$H_0 = F_1(x) = F_2(x) = \ldots = F_k(x) \quad vs \quad H_1 : F_i(x) \neq F_j(x) \quad \textbf{p.a. } i \neq j$$

En donde nuevamente si imponemos la hipótesis de que las medias existen y que dos distribuciones son distintas sólo por localizacion, entonces la hipótesis se puede plantear en terminos de media como sigue:

$$H_0 : \mathbb{E}(X_1) = \ldots = \mathbb{E}(X_n) \quad vs \quad H_1 : \mathbb{E}(X_i) \neq \mathbb{E}(X_j) \qquad \textbf{p.a. } i \neq j$$

El método propuesto por la prueba será nuevamente asignar Rangos a la muestra observada pero por cada renglón en tabla. Es decir $R(X_{ji})$ será el Rango asociado a la observación j del renglon i de tal forma que $R(X_{ji})$ es una v.a. que sólo puede tomar valores en el conjunto $\{1, 2, \ldots, k\}$. (Observe entonces que los rangos son asociados por renglon y en ningun momento se junta toda la muestra)

Transformada la muestra obtendremos una tabla generada por los rangos asociados:

$F_1(x)$	$F_2(x)$	$\ldots$	$F_k(x)$
$R(X_{11})$	$R(X_{21})$	$\ldots$	$R(X_{k1})$
$R(X_{12})$	$R(X_{22})$	$\ldots$	$R(X_{k2})$
$\vdots$	$\vdots$	$\ldots$	$\vdots$
$R(X_{1n})$	$R(X_{2n})$	$\ldots$	$R(X_{kn})$

Ahora notemos que bajo H_0 se espera que los rangos asociados por renglón sigan una distribución uniforme discreta, es decir:

$$\mathbb{P}\left(R(X_{ji}) = q\right) = \frac{1}{k} \quad j, q \in \{1, 2, \ldots, k\} \Rightarrow \mathbb{E}(R(X_{ji})) = \frac{k+1}{2}$$

Lo anterior es valido para cada renglón, es decir para $i \in \{1, 2, \ldots, n\}$. Como además suponemos m.a. del vector multivariado, entonces sabemos que los rangos asociados de renglón a renglón son independientes también, es decir:

$$R(X_{ji_1}) \quad \perp \quad R(X_{ji_2}) \quad i_1, i_2 \in \{1, 2, \ldots, n\}$$

Una vez tranformada la muestra definamos la suma de rangos por columna:

$$R_j = \sum_{i=1}^{n} R(X_{ji})$$

Observemos que si existe un j tal que la población j tienda a tomar valores más grandes que los demás, entonces se deberá observar que R_j toma valores grandes lo que iría en contra

de la hipótesis nula. Se propone entonces una estadística que mida la dicrepancia de R_j con su respectivo valor esperado.

Para definir la estadística de prueba primero calculemos la media y la varianza de R_j

$$\mathbb{E}(R_j) = \mathbb{E}\left(\sum_{i=1}^{n} R(X_{ji})\right) = \sum_{i=1}^{n} \mathbb{E}(R(X_{ji})) = \sum_{i=1}^{n} \frac{k+1}{2} = \frac{n(k+1)}{2}$$

Por otro lado la varianza es **(TAREA)**:

$$\text{Var}(R_j) = \frac{n(k+1)(k-1)}{12}$$

Entoneces suponiendo una muestra grande tenemos que:

$$\frac{R_j - \mathbb{E}(R_j)}{\sqrt{\text{Var}(R_j)}} = \frac{R_j - \frac{n(k+1)}{2}}{\sqrt{\frac{n(k+1)(k-1)}{12}}} \overset{aprox}{\sim} N(0,1)$$

Por lo tanto:

$$\left(\frac{R_j - \frac{n(k+1)}{2}}{\sqrt{\frac{n(k+1)(k-1)}{12}}}\right)^2 \overset{aprox}{\sim} \chi^2_{(1)}$$

Si las v.a. $R_1, \ldots, R_k$ fueran independientes podríamos concluir que:

$$\sum_{j=1}^{k} \frac{\left(R_j - \frac{n(k+1)}{2}\right)^2}{\frac{n(k+1)(k-1)}{12}} \overset{aprox}{\sim} \chi^2_{(k)} \tag{1.1}$$

Sin embargo, sabemos que $R_1, \ldots, R_k$ son dependientes, de hecho algo que tiene que ocurrir es que $\sum_{j=1}^{k} R_j = \frac{nk(k+1)}{2}$ por lo tanto **no podemos** argumentar independencia y asumir la distribución $\chi^2_{(k)}$.

El aporte que hizo Friedman fue probar que si se multiplica a cada sumando en (1.1) por el factor $\frac{k-1}{k}$ entonces si se obtiene una distribución χ^2 pero se pierde un grado de libertad, por lo tanto Friedman define el siguiente estadístico de prueba:

$$T = \sum_{j=1}^{k} \frac{k-1}{k} \frac{\left(R_j - \frac{n(k+1)}{2}\right)^2}{\frac{n(k+1)(k-1)}{12}} = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{n(k+1)}{2}\right)^2 \overset{aprox}{\sim} \chi^2_{(k-1)}$$

Luego entonces la idea para rechazar H_0 es fácil, se propone rechazar H_0 a un nivel de significancia

α si:

$$T > \chi^{2^{(1-\alpha)}}_{(k-1)}$$

Con $\chi^{2^{(1-\alpha)}}_{(k-1)}$ el cuantil $1-\alpha$ de una distribución χ^2 con $k-1$ grados de libertad.

Se puede probar además que una forma más comoda (computacionalmente) del es estadístico de prueba es:

$$T = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Desafortunamente la distribución es aproximada por lo que resulta necesario tener la distribución exacta para n pequeña.

A continuación se presenta el código en R que simula la distribución del estadístico de prueba, la idea es simular en cada uno de los n renglones un muestreo aleatorio sin reemplazo de una población de tamaño k.

```
#####################################################
#Programa que calcula   la distribucion Friedman  #
#####################################################
#Introduce k el numero de grupos o dimension del vector
par(mfrow = c(1, 2))
k=3
#Introduce el numero de muestras
n=5
#Introduce el numero de simulaciones
nSim=100000

M=matrix(0,n,k)
R=0

T<-rep(0,nSim)
for (i in 1:nSim){
  for (j in 1:n){
    M[j,]=sample(1:k,k,replace=FALSE)
    }
  for (l in 1:k){
    R[l]=sum(M[,l])
    }
  T[i]=12/((n*l)*(l+1))*sum(R^2)-3*n*(k+1)
}

plot(as.numeric(names(table(T))),table(T)/nSim,type="h",lwd=2,
```

```
     col="blue",ylab="Probabilidad",main="Densidad Friedman", xlab="T",ylim=c(0,0.3))
points(as.numeric(names(table(T))), table(T)/nSim,pch=16,cex=2,col="dark red")
text(as.numeric(names(table(T))), table(T)/nSim, round(table(T)/nSim,3), pos=3, cex=0.7)
f<-ecdf(T)
plot(f,main="Exacta vs Aproximacion")
curve(pchisq(x,k-1),add=TRUE,col=2,lwd=3)
```