USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

El test de la T de Student, ¿sólo en poblaciones normales?

Jorge Rodríguez Crespo

2020 / 2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

El test de la T de Student, ¿sólo en poblaciones normales?

Jorge Rodríguez Crespo

07/2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: El test de la T de Student, ¿sólo en poblaciones normales?
Breve descripción del contenido
El test de la T de Student es el procedimiento más habitual para realizar un contraste de hipótesis sobre la media. Tiene la ventaja de que la distribución del estadístico de contraste es una T de Student cuando los datos proceden de una población con distribución normal. Sin embargo, cuando los datos no proceden de una distribución normal, la distribución del estadístico no es una T de Student ni resulta fácil de determinar de manera exacta, lo cual puede repercutir en las propiedades del contraste. En este trabajo estudiaremos las propiedades del test de la T de Student, cuando los datos no proceden de la normal, y lo compararemos con otros tests, paramétricos y no paramétricos, que permiten resolver el problema del contraste de manera alternativa. Lo haremos mediante simulaciones, y lo ilustraremos también con datos reales.

Índice general

Resumen viii

Introducción xi

1. Test de la T de Student **1**

1.1. La T de Student 1

1.1.1. Caracterización y propiedades 3

1.2. El test de la T de Student en poblaciones normales 3

1.2.1. Intervalos de confianza y contrastes de hipótesis para la media sobre una población normal 4

1.2.2. Comparación de dos medias en muestras emparejadas 5

1.2.3. Comparación de dos medias en muestras independientes 5

1.3. El test de la T de Student en poblaciones no normales 7

2. Contrastes paramétricos sobre la media **9**

2.1. Conceptos generales de contrastes de hipótesis 9

2.2. Optimalidad del test T en poblaciones normales 11

2.2.1. Hipótesis nula simple contra alternativa simple 11

2.2.2. Contrastes unilaterales y bilaterales 14

2.3. Optimalidad en poblaciones no normales 16

2.3.1. Caso exponencial 16

2.3.2. Caso uniforme 21

3. Contrastes sobre medidas de posición central **27**

3.1. Test de la Mediana 28

3.1.1. Presentación del test 28

3.1.2. Extensión a otros cuantiles 28

3.2. Test de la mediana vs. test T en poblaciones simétricas 31

3.2.1. La distribución de Laplace 31
3.2.2. Caso normal 33
3.3. Contraste de Wilcoxon de rangos con signos (RSW) 34
3.4. Test RSW vs. test T bajo distribuciones simétricas 37
3.4.1. Caso doble exponencial (Laplace) 38
3.4.2. Caso Normal 39

4. Comparación de dos poblaciones **41**
4.1. Situaciones paramétricas. El caso normal. 41
4.2. Test de los signos 42
4.3. Test de Wilcoxon-Mann-Whitney 44
4.3.1. Función de potencia para la prueba basada en el estadístico W . . . 47
4.3.2. Test T vs. Wilcoxon-Mann-Whitney 49

Código de R **51**

Bibliografía **67**

Resumen

El test de la T de Student sirve para estimar la media de una población normalmente distribuida cuando el tamaño muestral es pequeño y la varianza es desconocida. Cuando los datos no son normales la potencia del test de la T de Student puede ser inferior a otros tests, ya sean paramétricos o no paramétricos. En este trabajo estudiaremos en qué contrastes, tanto en una población como en dos, la prueba T se ve superada o no por otros tests alternativos. Para ello, nos apoyaremos en ejemplos, donde se aportarán las representaciones de las funciones de potencia para dichos tests. En algunos casos será un problema complejo saber la distribución que seguirá el estadístico, pudiendo ser éste también difícil de calcular. Por este motivo, para realizar buenas aproximaciones de las funciones de potencia nos ayudaremos con simulaciones realizadas en R.

Abstract

The Student's T test is used to estimate the mean of a normally distributed population when the sample size is small and the variance is unknown. When the data are not normal, the power of the Student's T test may be lower than in other tests, whether parametric or non-parametric. In this project we will study in which contrasts, whether in one or two populations, the T test is surpassed or not by other alternative tests. To do this, we will rely on examples, where the representations of the power functions of these tests will be provided. Occasionally it will be a complex problem to know which distribution the statistic will follow, and it may also be difficult to calculate. Therefore, we will simulate in R to achieve an approximate representation of the power functions.

Introducción

A lo largo de la historia de la estadística un problema muy frecuente ha sido el de estudiar las carácterísticas que posee una determinada población, para lo que es necesario hacer inferencia sobre ciertos parámetros de dicha población. Como ejemplo, una de estas características puede ser la media, que ofrece una gran cantidad de información además de poseer ciertas propiedades muy útiles.

El problema de hacer inferencia sobre la media cuando los datos proceden de una distribución normal con variabilidad conocida es sencillo, pero, ¿qué ocurre si dicha varianza es desconocida? Pues bien, en 1908 William Sealy Gosset dio la solución a este problema al introducir la distribución T de Student. Esta distribución servirá para estimar la media de una población normalmente distribuída cuando el tamaño muestral es pequeño, ya que para tamaños grandes la aproximación por la normal ofrece practicamente el mismo resultado.

Ahora bien, en problemas aplicables en el mundo real se van a dar pocas situaciones en las que los datos de la muestra provengan de una distribución normal y además la varianza sea conocida. Entonces, podrían darse dos casos: que los datos provengan de una distribución conocida sin ser normales (caso paramétrico), o bien que no se sepa la distribución de partida (caso no paramétrico).

En este trabajo se abordará el problema de estimar la media en los casos descritos anteriormente, tanto en una como en dos poblaciones. Además, estudiaremos en qué situaciones el test T es realmente eficiente, comparándolo con otros tests, paramétricos y no paramétricos, lo que puede ser un problema bastante difícil ya que a veces no podremos saber que distribución seguirá el estadístico ni resultará fácil de calcular.

Por lo tanto, para realizar estas comparaciones entre los tests sugeridos introduciremos algunos conceptos, como los tipos de errores que se pueden cometer o la potencia de un test. Por otro lado, se expondrá el Lema de Neyman-Pearson y algún que otro resultado muy útiles para saber cuál es el test uniformemente más potente y en qué situaciones existirá dicho test.

De todos modos, como ya hemos comentado anteriormente, calcular el estadístico o saber qué distribución sigue no será nada fácil. Por ello, en determinados casos recurriremos

a las simulaciones mediante el progrma informático R, especialmente para simular las funciones de potencia de los tests estudiados. Estas representaciones, debido al elevado número de simulaciones, nos darán una idea bastante aproximada a la realidad, siendo posible ver a simple vista cuándo la potencia de un test supera a la de otro.

La estructura de este trabajo está constituída por cuatro capítulos: en el primero se expondrá el test de la T de Student; en el segundo se probará que el test descrito en el capítulo primero es el más potente, y además, estudiaremos la optimalidad mediante el Lema de Neyman-Pearson en los casos particulares de una distribución exponencial y otra uniforme; en el tercero particularizaremos el contraste para situaciones en las que la distribución sea simétrica, como la doble exponencial (Laplace), dado que en estos casos el contraste sobre la media será equivalente a realizar un contraste sobre la mediana; y por último, en el cuarto capítulo se hará frente al problema de comparación de dos poblaciones. Tanto en el tercer capítulo como en el cuarto se describirán dos tests no paramétricos, y estudiaremos como se comportan estos tests frente al test de la T de Student.

Asimismo, cabe mencionar que en este trabajo también se aportarán ejemplos de problemas que pueden darse en el mundo real, donde se plantearán contrastes sobre hipótesis (caso unilateral y bilateral) y se verá qué test es el óptimo y más recomendado, aportando la representación de la función de potencia de dicho test y la de otros tests alternativos.

Capítulo 1

Test de la T de Student

1.1. La T de Student

Actualmente se conoce a la distribución T de Student como la distribución usada para hacer inferencia sobre la media de una población normalmente distribuida cuando el tamaño muestral es pequeño. Esta distribución fue formalmente introducida por el químico William Sealy Gosset (1876-1937) en el año 1908. Éste trabajaba para la famosa marca de cerveza Guinnes (de Dublín) con el fin de lograr avances para el proceso industrial y la comercialización de la marca. El artículo donde se introducía al mundo esta distribución fue publicado en la revista inglesa Biometrika en el año 1908. Gosset firmó con un pseudónimo, Student, para que así se mantuvieran en secreto los procesos y avances industriales de la cervecera.

Antes de introducir el concepto de distribución de la T de student, comenzaremos hablando sobre los grados de libertad. Los **grados de libertad** son el grado de información dada por el conjunto de observaciones de la muestra que se emplean en estimar los valores de parámetros desconocidos o en estimar la variabilidad de los tests. Se determinan dependiendo del número de parámetros del modelo y del número de observaciones.

Por un lado, es obvio que si aumenta el número de observaciones, aumenta la cantidad de información, lo que se traduce como un incremento del número de los grados de libertad.

Por otra parte, si añadimos parámetros al modelo estamos "gastando" información, lo que reducirá el número de los grados de libertad.

Se calculan mediante la fórmula $(n - r)$ donde n es el tamaño de la muestra y r el número parámetros que es necesario estimar.

Definición 1.1. Decimos que una variable aleatoria X continua se distribuye según el modelo de probabilidad t **o** T **de Student con** k **grados de libertad** (k entero positivo)

si tiene como función de densidad:

$$f_X(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}(1+\frac{x^2}{k})^{-\frac{k+1}{2}}$$

para $x \in \mathbb{R}$, donde $\Gamma(p) = \int_0^\infty e^{-x}x^{p-1}\,dx$ es la función gamma. Escribiremos $X \sim t_k$ o $X \sim t(k)$.

La gráfica de la función de densidad es simétrica respecto del eje de ordenadas (independientemente del valor de k), y se asemeja bastante a la de una normal estándar:

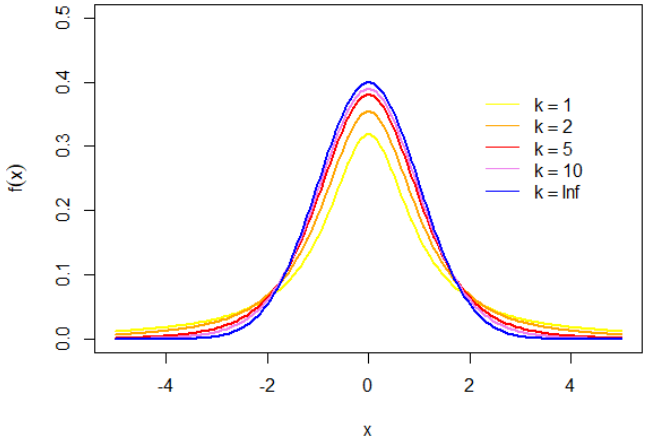

Figura 1.1: Función de densidad de la T de student con k grados de libertad

Además, la distribución T de student con k grados de libertad se puede definir como la distribución de la variable aleatoria

$$T := \frac{Z}{\sqrt{\frac{X}{k}}} \sim t_k$$

donde

- Z es una variable aleatoria que sigue una distribución normal estándar, es decir, $Z \sim N(0,1)$.
- $X \sim \chi^2_k$. Esto es, X es una variable aleatoria que tiene distribución Chi-cuadrado con k grados de libertad.
- X y Z son variables aleatorias independientes.

1.1.1. Caracterización y propiedades

Si X es una variable aleatoria tal que $X \sim t_k$ se satisface:

- $E[X] = \mu = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}(1 + \frac{x^2}{k})^{-\frac{k+1}{2}}\, dx$=0 para $k = 2, 3, ...$ y para $k = 1$, la variable X no tiene media pues $E\left[|X|\right] = +\infty$.
- La mediana y la moda toman el valor 0 para cualquier grado de libertad.
- Para $k > 2$, la varianza toma el valor:

$$Var[X] = \sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx$$
$$= \int_{-\infty}^{\infty} x^2 \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}(1 + \frac{x^2}{k})^{-\frac{k+1}{2}}\, dx = \frac{k}{k-2}$$

  mientras que para $k = 1, 2, \quad E[X^2] = \infty$

Es obvio que las gráficas de la distribución T de student y la normal estándar son muy parecidas: ambas alcanzan el valor máximo en cero, son unimodales, simétricas y con forma de campana. La diferencia notoria está en que la T tiene las colas más amplias que la normal. De todos modos, a medida que k aumenta, la dispersión de la curva de la distribución T disminuye, y esto da lugar a que, cuando $k \to \infty$, el límite de la distribución T coincide con la distribución normal estándar.

1.2. El test de la T de Student en poblaciones normales

El estimar o hacer inferencia sobre la media de una población siempre ha sido objeto de estudio de la Estadística. En la práctica, si partimos de una población normal, lo más seguro es que se desconozca cuál es su varianza. Por ello, para realizar contrastes de hipótesis y hallar intervalos de confianza será necesario el uso de otra prueba que no sea la normal. A esta prueba se le llamará test de la T de Student, donde usaremos la cuasivarianza, que es un estimador insesgado de la varianza.

Definición 1.2. Dada una muestra aleatoria simple $X_1, ..., X_n \in N(\mu, \sigma^2)$, se define la **cuasivarianza** muestral como:

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Xi - \bar{X})^2$$

En esta sección nuestro objetivo será ver que, partiendo de poblaciones normales, a partir del teorema de Fisher seremos capaces de hallar intervalos de confianza y de realizar

contrastes de hipótesis relativas a la media cuando la varianza es desconocida. A partir de ahora y en lo que resta de sección supondremos que las poblaciones son normales.

Teorema 1.3 (Teorema de Fisher)**.** *Sean $X_1, ..., X_n \in N(\mu, \sigma^2)$ independientes. Consideremos la media muestral, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, y la varianza y cuasi-varianza muestrales, $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ y $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ respectivamente. Entonces se verifica:*

- $\bar{X} \in N(\mu, \frac{\sigma^2}{n})$
- $\frac{nS^2}{\sigma^2} = \frac{(n-1)S_c^2}{\sigma^2} \in \chi^2_{n-1}$
- *$\bar{X}$ y S^2 (o S_c^2) son independientes*

1.2.1. Intervalos de confianza y contrastes de hipótesis para la media sobre una población normal

Para hallar un intervalo de confianza para la media (con varianza desconocida) sustituiremos la varianza poblacional por la cuasivarianza muestral. Surge entonces el problema de que el pivote $(\bar{X} - \mu)/(S_c/\sqrt{n})$ deja de seguir una distribución normal estándar. Con ayuda del Teorema de Fisher podremos ver que este pivote sigue una distribución T de student con $(n-1)$ grados de libertad.

En efecto, el Teorema de Fisher prueba que:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in N(0,1) \qquad \text{y} \qquad \frac{(n-1)S_c^2}{\sigma^2} \in \chi^2_{n-1}$$

y son independientes entre sí. Por lo tanto

$$\frac{\bar{X} - \mu}{S_c/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\sqrt{((n-1)S_c^2/\sigma^2)/(n-1)}} \in T_{n-1} \tag{1.1}$$

Se puede deducir, a partir del pivote, el intervalo de confianza para la media cuando la varianza es desconocida:

$$\left(\bar{X} - t_{\alpha/2}\frac{S_c}{\sqrt{n}},\ \ \bar{X} + t_{\alpha/2}\frac{S_c}{\sqrt{n}}\right) \tag{1.2}$$

Por tanto, suponiendo que $X_1, ..., X_n \in N(\mu, \sigma^2)$ independientes, si se quisiera realizar un contraste de hipótesis sobre la media con **varianza desconocida** siendo H_0 cierta, entonces el contraste unilateral será:

$$\text{Rechazar } H_0 : \mu \geq \mu_0 \qquad \text{si } \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} > t_\alpha$$

y en el caso de un contraste bilateral:

$$\text{Rechazar } H_0 : \mu = \mu_0 \qquad \text{si } \frac{|\bar{X}-\mu_0|}{S_c/\sqrt{n}} > t_{\alpha/2}$$

donde t_α y $t_{\alpha/2}$ son las abscisas que dejan a la derecha una probabilidad de α y $\alpha/2$, respectivamente, en una distribucion T_{n-1}.

1.2.2. Comparación de dos medias en muestras emparejadas

Consideremos dos variables aleatorias X e Y suponiendo que $X \in N(\mu_1, \sigma_1^2)$ e $Y \in N(\mu_2, \sigma_2^2)$ que se miden simultáneamente en cada individuo de una muestra de n individuos independientes. En este momento se podría realizar un contraste bilateral (ver si las medias son iguales) o unilateral (ver si una media es mayor que la otra). El objetivo será contrastar si la diferencia de las medias $(\mu_1 - \mu_2)$ es cero o si es mayor/menor que cero. Si $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$ con $D_i = X_i - Y_i$, $i \in 1, ..., n$, aplicando el Teorema de Fisher,

$$\frac{\bar{X}-\bar{Y}}{S_c/\sqrt{n}} \in T_{n-1}$$

por lo que siendo $t_{\alpha/2}$ la abscisa que deja una probabilidad de $\alpha/2$ a la derecha de la distribución T_{n-1}, rechazaremos $H_0 : \mu_1 = \mu_2$ a favor de $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{|\bar{X}-\bar{Y}|}{S_c/\sqrt{n}} > t_{\alpha/2}$$

1.2.3. Comparación de dos medias en muestras independientes

Suponemos dos poblaciones normales con respectivas medias y varianzas: $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$.

Se extrae de manera independiente una muestra aleatoria simple de cada población, y se obtienen los estimadores muestrales de la media y la varianza para cada una de las muestras. Los estimadores para μ_1 y μ_2 son

$$\bar{X}_1 = \frac{X_{11}+...+X_{1n_1}}{n_1} \qquad \bar{X}_2 = \frac{X_{21}+...+X_{2n_2}}{n_2}$$

y para σ_1^2 y σ_2^2 estarán dados por

$$S_{c1}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 \qquad S_{c2}^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$$

Caso de varianzas conocidas

En el caso en el que las **varianzas** fueran **conocidas**, el contraste de igualdad de medias ($H_0 : \mu_1 = \mu_2$ cierta) sería muy sencillo ya que el estadístico de contraste sería de la forma

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \in N(0,1)$$

Por lo tanto, siendo $z_{\alpha/2}$ la abscisa que deja una probabilidad de $\alpha/2$ a la derecha (en la distribución normal), rechazaremos $H_0 : \mu_1 = \mu_2$ a favor de $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$$

pero, si las varianzas son desconocidas, el problema de contrastar la igualdad de medias se complica bastante.

Caso varianzas desconocidas pero iguales

Suponiendo que las varianzas son **desconocidas pero iguales** podremos obtener un estadístico de contraste que seguirá una T de Student con $(n_1 + n_2 - 2)$ grados de libertad. Como consideramos $\sigma_1^2 = \sigma_2^2$, entonces tomamos un estimador común para la varianza

$$S_p^2 = \frac{(n_1-1)S_{c1}^2 + (n_2-1)S_{c2}^2}{n_1+n_2-2}$$

siendo éste una media de las dos cuasivarianzas muestrales, ponderadas por sus grados de libertad. Ahora bien, gracias al Teorema de Fisher tenemos que

$$\frac{(n_1-1)S_{c1}^2}{\sigma^2} \in \chi^2_{n_1-1} \qquad \frac{(n_2-1)S_{c2}^2}{\sigma^2} \in \chi^2_{n_2-1}$$

y por ser las muestras independientes éstos también lo serán. Sumándolos

$$\frac{(n_1-1)S_{c1}^2}{\sigma^2} + \frac{(n_2-1)S_{c2}^2}{\sigma^2} = \frac{(n_1+n_2-2)S_p^2}{\sigma^2}$$

que seguirá una distribución $\chi^2_{n_1+n_2-2}$ donde S_p^2 es independiente de $X_1 - X_2$, lo que permite estandarizar la diferencia de medias. Por lo tanto el estadístico de contraste será

$$\frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \in T_{n_1+n_2-2}$$

y la regla de decisión consistirá en rechazar $H_0 : \mu_1 = \mu_2$ si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha/2}$$

siendo $t_{\alpha/2}$ la abscisa que deja a su derecha una probabilidad de $\alpha/2$ en una distribución $T_{n_1+n_2-2}$.

Caso de varianzas desconocidas y distintas

Cuando las varianzas se suponen desconocidas y distintas encontrar la distribución exacta del estadístico bajo H_0 es un problema difícil. Este problema se conoce como **problema de Behrens-Fisher**, y hoy en día todavía no se conoce solución óptima. De todos modos, sí hay diversas soluciones aproximadas.

Una solución comúnmente utilizada es el test de Welch, que describiremos a continuación.

Denotando γ como el entero más proximo al cálculo de

$$\frac{\left(S_{c1}^2/n_1 + S_{c2}^2/n_2\right)^2}{\frac{\left(S_{c1}^2/n_1\right)^2}{n_1-1} + \frac{\left(S_{c2}^2/n_2\right)^2}{n_2-1}}$$

entonces la regla de decisión será rechazar $H_0 : \mu_1 = \mu_2$ si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} > t_{\alpha/2}$$

siendo $t_{\alpha/2}$ la abscisa que deja a la derecha una probabilidad de $\alpha/2$ en la distribución T de Student con γ grados de libertad.

En el caso de muestras grandes se utiliza la aproximación por la normal, es decir, se rechazará $H_0 : \mu_1 = \mu_2$ si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} > z_{\alpha/2}$$

con $z_{\alpha/2}$ el cuantil que deja a su derecha una probabilidad de $\alpha/2$ en la distribución normal estándar.

1.3. El test de la T de Student en poblaciones no normales

El siguiente teorema, al que se le conoce como **Teorema Central del Límite**, refleja esta situación:

Teorema 1.4 (Lindeberg-Lèvy)**.** *Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas a X, con $\mu = E[X]$ y $\sigma^2 = Var[X]$. Sea $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ la media de las n primeras variables. Entonces, se verifica:*

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1)$$

Como consecuencia de este teorema podemos interpretar lo siguiente:

$$\bar{X}_n - \mu \xrightarrow{p} 0 \text{ (sucesión estocástica)}$$

y

$$\frac{\sigma}{\sqrt{n}} \xrightarrow{n\to+\infty} 0 \text{ (sucesión determinista)},$$

que nos indica la velocidad de convergencia a la normal estándar.

En otras palabras, el teorema central del límite (TCL) establece que, dada una muestra aleatoria lo suficientemente grande (usualmente se considera $n \geq 30$), la distribución de las medias muestrales seguirá una distribución normal. Esto se cumplirá independientemente de la forma de la distribución con la que trabajemos, con el único requisito de que tenga media y varianza. Aún así, cuando X tenga una distribución más parecida a la normal, la aproximación por el TCL se alcanzará con un menor valor de n. En conclusión, el TCL permitirá hacer inferencia de la media poblacional a partir de la media muestral.

Entonces, como ya sabemos, si $X \in N(\mu, \sigma^2)$, donde la varianza es desconocida, el **estadístico de contraste** seguirá una distribución T de Student con $n-1$ grados de libertad. Por lo tanto, gracias a este resultado, si X no es normal y $n \geq 30$, el estadístico de contraste también seguirá, de forma aproximada, una T_{n-1}. Además, si X no difiere apenas de la normal, también podrá ser aproximada por una T_{n-1}.

Capítulo 2

Contrastes paramétricos sobre la media

Los contrastes paramétricos son aquellos en los que se conoce la forma de la distribución salvo unos parámetros sobre los que se plantean las hipótesis de inferencia. En este capítulo el objetivo será ver qué test es el óptimo para realizar un contraste sobre la media de una determinada población.

Cuando se realiza un contraste de hipótesis lo que se estudia es si aceptar o no la hipótesis nula, pero será también de gran interés el saber cuáles son las alternativas a la hipótesis propuesta, lo cual está relacionado con el concepto de potencia del test. En este capítulo nos apoyaremos sobre todo en el lema de Neyman-Pearson, que será de gran ayuda a la hora de decidir qué test conviene usar para maximizar la eficacia del contraste, en términos de la potencia del test en cuestión.

Previamente a enunciar el Lema de Neyman-Pearson será necesario el introducir varios conceptos y definiciones.

2.1. Conceptos generales de contrastes de hipótesis

Recordemos que en todo contraste de hipótesis se plantea una hipótesis nula (H_0), que se supone cierta de partida, y una alternativa (H_1), que reemplaza a la hipótesis nula cuando ésta es rechazada.

Definición 2.1. Sea $X_1, ..., X_n$ una muestra aleatoria simple extraída de una variable poblacional cuya distribución depende de un parámetro $\theta \in \Theta$. Un *test de hipótesis puro*

para contrastar H_0 frente a H_1 se puede pensar como una función φ definida como:

$$\varphi(x) = \mathrm{I}_C(x) = \begin{cases} 1, \text{ si } x \in C \\ 0, \text{ en otro caso} \end{cases}$$

donde C es la ***región crítica*** *(o de rechazo) del test*. Por lo tanto, $\varphi(x)$ representa la probabilidad de rechazar la hipótesis nula, si se obtiene el resultado muestral X.

Definición 2.2. Un **test aleatorizado** es una función del espacio muestral en $[0, 1]$ donde $\varphi(X)$ representa la probabilidad de rechazar cuando se observa la muestra X.

A la probabilidad de rechazar H_0 se le denomina ***función de potencia***, siendo ésta una función de θ definida como

$$\beta_\varphi(\theta) = E_\theta[\varphi(X)], \quad \forall \theta \in \Theta$$

Definición 2.3. Al realizar el contraste de hipótesis existen dos opciones: aceptar H_0 o bien rechazar H_0. En el momento de tomar la decisión podríamos cometer dos tipos de errores:

- **Error de tipo 1**: rechazar H_0 siendo H_0 cierta (**muy peligroso**). Siendo H_0 cierta, hay riesgo de que la muestra caiga en la región de rechazo. Para $\theta \in \Theta_0$:

$$P(\text{Error tipo 1}) = P(\text{rechazar } H_0/H_0 \text{ cierta}) = P(X \in C/H_0 \text{ cierta}) = E_\theta[\varphi(X)]$$

- **Error de tipo 2**: aceptar H_0 siendo H_0 falsa (**no tan peligroso**). Para $\theta \in \Theta_1$:

$$P(\text{Error tipo 2}) = 1 - P(X \in C/H_0 \text{ falsa}) = 1 - E_\theta[\varphi(X)]$$

Observación 2.4. Nótese que si Θ_0 tiene un único punto ($\Theta_0 = \{\theta_0\}$), en cuyo caso estamos ante una **hipótesis nula *simple***, entonces P_θ(Error tipo 1) tomará un único valor. Si esto no fuera así estaríamos ante una **hipótesis nula *compuesta***, y la P(Error tipo 1) sería una función de $\theta \in \Theta_0$. Análogamente podemos pensar que ocurre lo mismo con la hipótesis alternativa y P(Error tipo 2).

Definición 2.5. Se define el ***tamaño*** de un test φ como el valor

$$\sup \{\beta_\varphi(\theta) : \theta \in \Theta_0\}$$

Definición 2.6. Un test de hipótesis que contrasta $H_0 \in \Theta_0$ frente a $H_1 \in \Theta_1$, se dice que tiene un ***nivel de significación*** $\alpha \in [0,1]$ si su tamaño es menor o igual que α:

$$E_\theta[\varphi(X)] \leq \alpha, \quad \forall \theta \in \Theta_0$$

Definición 2.7. Dados dos test φ y φ' de nivel de significación α, se dice que φ es ***uniformemente más potente*** que φ' si:

$$\beta_\varphi(\theta) \geq \beta_{\varphi'}(\theta), \quad \theta \in \Theta_1$$

Por otra parte, un test φ se dice que es uniformemente de ***máxima potencia*** **(UMP)** dentro de una familia de test de nivel de significación α, si es uniformemente más potente que cualquier test de la familia.

2.2. Optimalidad del test de la T de Student en poblaciones normales

2.2.1. Hipótesis nula simple contra alternativa simple

En esta sección el objetivo es construir el test más potente de nivel α que contraste una hipótesis nula simple contra una alternativa simple. Para ello nos apoyaremos en el siguiente resultado:

Teorema 2.8 (Lema de Neyman-Pearson)**.** *Supongamos que $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$, ambas simples.*

1. *Cualquier test de la forma:*

$$\varphi(X_1,...,X_n) = \begin{cases} 1, & si\ f_{\theta_1}(X_1,...,X_n) > k f_{\theta_0}(X_1,...,X_n) \\ \gamma(X_1,...,X_n), & si\ f_{\theta_1}(X_1,...,X_n) = k f_{\theta_0}(X_1,...,X_n) \\ 0, & si\ f_{\theta_1}(X_1,...,X_n) < k f_{\theta_0}(X_1,...,X_n) \end{cases}$$

con $k \geq 0$, $\gamma \in [0,1]$, es UMP dentro de los test con nivel de significación igual a su tamaño para contrastar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$.

Si $k = \infty$, el test de la forma:

$$\varphi(X_1,...,X_n) = \begin{cases} 1, & si\ f_{\theta_0}(X_1,...,X_n) = 0 \\ 0, & si\ f_{\theta_0}(X_1,...,X_n) > 0 \end{cases}$$

es el más potente con $\alpha = 0$ para contrastar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$.

2. ***(Existencia)***. *Para cualquier $\alpha \in [0,1]$, existe un test φ de la forma anterior de tamaño α ($E_{\theta_0}[\varphi(x)] = \alpha$).*

3. ***(Unicidad)***. *Cualquier test UMP es de la forma descrita en a), salvo en un conjunto de medida nula bajo H_0 y bajo H_1.*

Veamos entonces qué test es UMP en el caso de realizar un test de hipótesis sobre la media en poblaciones normales, en el caso de hipótesis nula simple frente a hipótesis alternativa simple:

Problema 2.9. *Sea $X_1, ..., X_n$ una muestra aleatoria simple de $X \in N(\mu, \sigma^2)$. Se pide calcular el test UMP de nivel α para contrastar $H_0 : \mu = \mu_0$ frente a $H_a : \mu = \mu_1$ (con $\mu_0 < \mu_1$).*

Solución. Lo haremos suponiendo que σ es conocida para poder aplicar el Lema de Neymann-Pearson.

Recordemos que para un μ y σ cualquiera, la función de densidad deseada es de la forma:

$$\begin{aligned} f_{\mu,\sigma}(X_1, ..., X_n) &= \prod_{i=1}^{n} f_{\mu,\sigma}(X_i) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} \right] \\ &= (2\pi\sigma^2)^{-\mu/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned} \tag{2.1}$$

si $X_1, ..., X_n$ independientes.

Según el lema de Neyman-Pearson, y sabiendo que X es absolutamente continua, el test UMP es de la forma

$$\varphi(X_1, ..., X_n) = \begin{cases} 1, \text{ si } \lambda(X) > k \\ 0, \text{ si } \lambda(X) < k \end{cases}$$

donde k es la constante que hay que determinar para que el test tenga el nivel α, i.e., $E_{\mu_0,\sigma}\varphi(X) = \alpha$, y siendo λ el cociente:

$$\lambda(X) = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}}$$

Entonces el test UMP consiste en rechazar $H_0 : \mu = \mu_0$ a favor de $H_a : \mu = \mu_1$ si

$$\lambda(X) = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}} > k \tag{2.2}$$

y aplicando logaritmos se traduce en

$$\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu_0)^2 - \sum_{i=1}^{n}(x_i-\mu_1)^2\right] > \log k \tag{2.3}$$

Por lo tanto, el test UMP rechaza $H_0 : \mu = \mu_0$ a favor de $H_a : \mu = \mu_1$ si el estadístico

$$\sum_{i=1}^{n} (x_i - \mu_0)^2 - \sum_{i=1}^{n} (x_i - \mu_1)^2 \tag{2.4}$$

es grande, y para ello necesitamos obtener la distribución del mismo teniendo en cuenta el nivel α. Además, se puede ver que **este estadístico no depende del valor de** σ, interviniendo únicamente los valores de la muestra, μ_0 y μ_1. Para simplificar el estadístico tenemos en cuenta que la media muestral es $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, por lo que para cualquier μ

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \mu)^2 &= \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - \mu)^2 + 2\sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - \mu) \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \end{aligned} \tag{2.5}$$

ya que $(\bar{x} - \mu)\sum_{i=1}^{n} (x_i - \bar{x}) = (\bar{x} - \mu)(\sum_{i=1}^{n} x_i - n\bar{x}) = 0$. Entonces el estadístico quedaría en función de

$$\begin{aligned} &\sum_{i=1}^{n} (x_i - \mu_0)^2 - \sum_{i=1}^{n} (x_i - \mu_1)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 \\ &- \sum_{i=1}^{n} (x_i - \bar{x})^2 - n(\bar{x} - \mu_1)^2 = n\left[(\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2\right] \end{aligned} \tag{2.6}$$

Por lo que el objetivo será ver cuando $(\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2$ es grande. Definimos $R(\bar{x}) = (\bar{x} - \mu_0)^2 - (\bar{x} - \mu_1)^2$ con derivada $R'(\bar{x}) = 2(\bar{x} - \mu_0) - 2(\bar{x} - \mu_1) = 2(\mu_1 - \mu_0)$. Tenemos que $R'(\bar{x}) > 0$ para todo valor real si $\mu_1 - \mu_0 > 0$ y $R'(\bar{x}) < 0$ si $\mu_1 - \mu_0 < 0$.

Por lo tanto, si $\mu_1 > \mu_0$ el test UMP consiste en rechazar $H_0 : \mu = \mu_0$ a favor de $H_a : \mu = \mu_1$ si $\bar{X}$ es grande. En el otro caso, si $\mu_1 < \mu_0$, el test rechaza si $\bar{X}$ es pequeño.

Ahora, estandarizando, en el caso de que σ sea conocida, se tiene que

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \in N(0, 1) \tag{2.7}$$

En el caso de que σ sea desconocida, debemos sustituir σ por la cuasidesviación típica, S_c. Para que se cumpla el nivel fijado α, $\bar{X}$ será grande si $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} > t_\alpha$ ya que

$$P(\text{rechazar } H_0 : \mu = \mu_0 / \mu = \mu_1) = P\left(\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} > t_\alpha / \mu = \mu_1\right) = \alpha$$

por ser $\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \in T_{n-1}$.

Entonces la solución de Neyman-Pearson es:

$$\varphi(X_1, ..., X_n) = \begin{cases} 1, \text{ si } \bar{x} > \mu_0 + t_\alpha \dfrac{S_c}{\sqrt{n}} \\ 0, \text{ en otro caso} \end{cases} \tag{2.8}$$

Es necesario comentar que el test con varianza conocida es más potente que el test de la T de Student, pero **si no se conoce la varianza, el test de la T de Student es la mejor opción posible**.

2.2.2. Contrastes unilaterales y bilaterales

Supongamos ahora que estamos interesados en contrastar la hipótesis $H_0 : \mu \leq \mu_0$ contra $H_1 : \mu > \mu_0$ (o bien $H_0 : \mu \geq \mu_0$ contra $H_1 : \mu < \mu_0$) para algún valor de μ_0, es decir, contrastes de una sola cola. En este caso, por lo general, no existe un test UMP. De todos modos, sí existirá para las familias de verosimilitudes que poseen la siguiente propiedad:

Definición 2.10 (**Propiedad de razón monótona (MLR)**)**.** La familia de verosimilitudes $f_\theta(X)$, con $\theta \in \Theta$ se dice que tiene la propiedad de razón monótona (MLR) en el estadístico $T(X)$ si para cualquier $\theta < \theta'$ siendo $f_\theta(X)$ y $f_{\theta'}(X)$ distintas, el cociente $f_\theta(X)/f_{\theta'}(X)$ es una función creciente de $T(X)$.

En el caso de querer realizar un contraste de una cola como el dicho anteriormente, si no se verifica la propiedad MLR ya no tendrá que existir un test UMP con nivel α, para un α dado. Aún así será de gran interés encontrar el test UMP en un entorno del punto μ_0, siendo éste el que mejor rechaza H_0 cuando $\mu > \mu_0$ y μ está cerca de μ_0. Esto hace que sea necesario introducir el siguiente concepto:

Definición 2.11. Se dice que un test φ para el problema $(\alpha, \Theta_0, \Theta_1)$ es ***centrado*** si $\forall \theta \in \Theta_1$ y se cumple que

$$E_\theta \varphi(X) \geq \alpha$$

El carácter centrado será entonces necesario para obtener la optimalidad en los tests bilaterales.

En conclusión, podemos decir que aunque a veces no exista el test UMP de nivel α sí existirá el test UMP entre los centrados de nivel α. Por ende, en los contrastes unilaterales y bilaterales estaremos ante el caso que acabamos de mencionar, es decir, los test serán UMP entre los centrados.

Ejemplo 2.12. Supongamos que se quiere realizar el contraste de $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ cuando una muestra de tamaño $n = 20$ proviene de una normal $N(\mu, \sigma^2)$ con μ y σ^2 desconocidas. Entonces la función de potencia cuando $\mu_0 = 0$ es la dada en la Figura 2.1, calculada mediante simulaciones (en este caso con $\sigma = 1$ y 20000 muestras simuladas). En esta figura se puede ver que cuando $\theta = 0$ la función de potencia toma el valor α. Además, es creciente en todo su dominio, siendo ya muy proxima a 1 cuando $\theta = 1$.

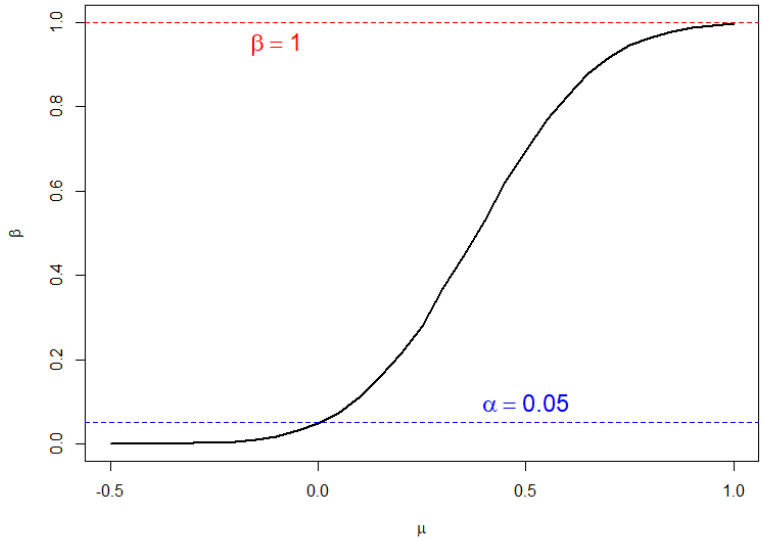

Figura 2.1: Función de potencia simulada del test de la T de Student para distintos valores de μ

Nota: para dibujar la función de potencia del test de la t de student se han realizado simulaciones ya que calcular la distribución del estadístico cuando los datos provienen de una distribución normal es un problema difícil (los comandos utilizados para la simulación en R se encuentran en el anexo).

Problema 2.13. *Supongamos que ahora el objetivo es realizar el contraste de* $H_0 : \mu = \mu_0$ *frente a* $H_1 : \mu \neq \mu_0$ *cuando* $X \in N(\mu, \sigma^2)$ *con* $\mu \in \mathbb{R}$ *y* σ^2 *desconocida.*

Solución. Procediendo análogamente a los contrastes unilaterales, se llega a que el test que rechaza $H_0 : \mu = \mu_0$ si $|\bar{x} - \mu_0|/(Sc/\sqrt{n}) > t_{n-1;\alpha/2}$ tiene un nivel de significación α si $\mu = \mu_0$.

Se pueden resumir los resultados obtenidos en el siguiente teorema:

Teorema 2.14. *Dada una muestra aleatoria simple $X \in N(\mu, \sigma^2)$ con varianza desconocida, si se desea estudiar la media, el test para contrastar H_0 contra H_1, con un nivel α, obtiene las siguientes regiones críticas:*

- $H_0 : \mu \leq \mu_0$ *frente a* $H_1 : \mu > \mu_0$ *:* $\quad T > t_{n-1;\alpha}$
- $H_0 : \mu \geq \mu_0$ *frente a* $H_1 : \mu < \mu_0$ *:* $\quad T < t_{n-1;1-\alpha}$
- $H_0 : \mu = \mu_0$ *frente a* $H_1 : \mu \neq \mu_0$ *:* $\quad |T| > t_{n-1;\alpha/2}$

siendo T el estadístico:

$$T = \frac{\bar{X} - \mu_0}{S_c}\sqrt{n} \tag{2.9}$$

2.3. Optimalidad en poblaciones no normales

En la sección anterior hemos visto cuáles son los tests óptimos para contrastar hipótesis cuando los datos provienen de una normal, pero, ¿qué ocurrirá cuando no haya normalidad? El objetivo ahora será estudiar aquellos casos en los que sabemos que los datos proceden de una distribución continua conocida (exponencial y uniforme) y comparar estos casos con el test de la t de student.

2.3.1. Caso exponencial

Problema 2.15. *Sea $X_1, ..., X_n$ una muestra aleatoria simple, donde la función de densidad es:*

$$f_\theta(x) = \frac{1}{\theta} e^{-x/\theta} \cdot I_{(0,\infty)}(x) \tag{2.10}$$

es decir, que siguen una exponencial de media θ. Encontrar el test UMP para el contraste $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ con $\theta_1 > \theta_0$ para el nivel de significación α.

Solución. En primer lugar, la función de densidad para $X_1, ..., X_n$ es de la forma.

$$f_\theta(X_1, ..., X_n) = \prod_{i=1}^{n} f_\theta(X_i) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n X_i} \cdot I_{\{X_i \geq 0\}} = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n X_i} \cdot I_{\{X_{(1)} \geq 0\}} \tag{2.11}$$

Sea θ_1 un punto arbitrario en (θ_0, ∞), pasamos a considerar el test $H_0 : \theta = \theta_0$ (simple) frente a $H_1 : \theta = \theta_1$ (simple). Por el lema de Neyman-Pearson, el test más potente a nivel

α (UMP) es:

$$\varphi(X_1,...,X_n)=\begin{cases}1,\ \text{si } f_{\theta_1}(X_1,...,X_n)>kf_{\theta_0}(X_1,...,X_n)\\ \gamma,\ \text{si } f_{\theta_1}(X_1,...,X_n)=kf_{\theta_0}(X_1,...,X_n)\\ 0,\ \text{si } f_{\theta_1}(X_1,...,X_n)<kf_{\theta_0}(X_1,...,X_n)\end{cases}$$

donde $\alpha=E_{\theta_0}\varphi(X)$. Por lo tanto, rechazamos $H_0:\theta=\theta_0$ si $f_{\theta_1}(X_1,...,X_n)>kf_{\theta_0}(X_1,...,X_n)$ para algún $k\geq 0$ $(0<k<1)$. Observemos que tanto si la muestra procede de $H_0:\theta=\theta_0$ como si procede de $H_1:\theta=\theta_1$, se tiene que $X_i\geq 0\ \forall i\in\{1,...,n\}$, por lo tanto:

$$\frac{f_{\theta_1}(X_1,...,X_n)}{f_{\theta_0}(X_1,...,X_n)}>k\Leftrightarrow\frac{\frac{1}{\theta_1^n}e^{-\frac{1}{\theta_1}\sum_{i=1}^n X_i}\cdot\ \ I_{\{X_{(1)}\geq 0\}}}{\frac{1}{\theta_0^n}e^{-\frac{1}{\theta_0}\sum_{i=1}^n X_i}\cdot I_{\{X_{(1)}\geq 0\}}}>k$$

$$\Leftrightarrow\left(\frac{\theta_1}{\theta_0}\right)\cdot e^{-\frac{1}{\theta_1}\sum_{i=1}^n X_i}\cdot e^{\frac{1}{\theta_0}\sum_{i=1}^n X_i}>k$$

$$\Leftrightarrow\left(\frac{\theta_0}{\theta_1}\right)\cdot e^{\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)\sum_{i=1}^n X_i}>k \tag{2.12}$$

y como $0<\left(\frac{\theta_0}{\theta_1}\right)^n<1$, se debe cumplir

$$e^{\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)\sum_{i=1}^n X_i}>k\left(\frac{\theta_1}{\theta_0}\right)=c \text{ para algún } c>1$$

Aplicando logaritmos

$$\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)\sum_{i=1}^n X_i>c' \text{ para algún } c'>0$$

y como $\theta_0<\theta_1$ entonces

$$\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)>0$$

Luego

$$\sum_{i=1}^n X_i>\frac{c'}{\left(\frac{1}{\theta_0}-\frac{1}{\theta_1}\right)}=c'' \text{ para algún } c''>0$$

Procedamos ahora a estudiar la distribución de $\sum_{i=2}^n X_i$:

Bajo H_0 las variables aleatorias $(X_1,...,X_n)$ están idénticamente distribuídas tal que $X_i\sim Exp(\frac{1}{\theta_0})$, entonces

$$X_i+...+X_n=\sum_{i=1}^n X_i\sim\Gamma(n,\frac{1}{\theta_0})$$

y además, si $X \sim \Gamma(a,p)$ se cumple que para cualquier $m>0$

$$mX \sim \Gamma(a, \frac{p}{m})$$

Luego, para el θ de la distribución de los X_i se tiene que

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \in \Gamma(n, \frac{1}{2}) \Rightarrow \frac{2}{\theta}\sum_{i=1}^{n} X_i \in \chi^2_{2n} \tag{2.13}$$

Así, obtenemos el valor de c'':

$$\begin{aligned} \alpha &= P_{\theta_0}\{\text{Rechazar } H_0/\theta=\theta_0\} = P_{\theta_0}\left\{\sum_{i=1}^{n} X_i > c''\right\} \\ &= P_{\theta_0}\left\{\frac{2}{\theta_0}\sum_{i=1}^{n} X_i > \frac{2}{\theta_0}c''\right\} = P_{\theta_0}\left\{\chi^2_{2n} > \frac{2}{\theta_0}c''\right\} \\ &= 1 - P_{\theta_0}\left\{\chi^2_{2n} < \frac{2}{\theta_0}c''\right\} \end{aligned} \tag{2.14}$$

$$\Rightarrow \frac{2}{\theta_0}c'' = \chi^2_{2n,\alpha} \Leftrightarrow c'' = \frac{\theta_0 \cdot \chi^2_{2n,\alpha}}{2}$$

siendo $\chi^2_{2n,\alpha}$ el cuantil que deja a su derecha una probabilidad α en la distrubución χ^2_{2n}.

Entonces el test UMP es:

$$\varphi(X_1,...,X_n) = \begin{cases} 1, \text{ si } \displaystyle\sum_{i=1}^{n} X_i > \frac{\theta_0 \cdot \chi^2_{2n,\alpha}}{2} \\ 0, \text{ si } \displaystyle\sum_{i=1}^{n} X_i < \frac{\theta_0 \cdot \chi^2_{2n,\alpha}}{2} \end{cases} \tag{2.15}$$

Con esto se podría ver que para cualquier elección de $\theta_1 \in (\theta_0, \infty)$ el valor encontrado para c'' sería siempre el mismo, lo que significa que el test es el más potente independientemente del valor de $\theta_1 \in (\theta_0, \infty)$.

Por ello, para $\theta \in (0,\infty)$, sabiendo que $X_i \in Exp(1/\theta)$ y $\frac{2}{\theta}\sum_{i=1}^{n} X_i \in \chi^2_{2n}$, la función de potencia será:

$$\begin{aligned} \beta_\varphi(\theta) = \mathbb{E}_\theta\varphi(X) &= P_\theta\left\{\sum_{i=1}^{n} X_i > c''\right\} = P_\theta\left\{\frac{2}{\theta}\sum_{i=1}^{n} X_i > \frac{2}{\theta}c''\right\} \\ &= P_\theta\left\{\chi^2_{2n} > \frac{2}{\theta}\cdot\frac{\theta_0 \cdot \chi^2_{2n,\alpha}}{2}\right\} \\ &= 1 - P_\theta\left\{\chi^2_{2n} < \frac{\theta_0 \cdot \chi^2_{2n,\alpha}}{\theta}\right\} \end{aligned} \tag{2.16}$$

Problema 2.16. *En las hipótesis del problema anterior, pero suponiendo que* H_0 *es compuesta,* $H_0 : \theta \leq \theta_0$*, para cualquier valor* $\hat{\theta}_0 \in (0, \theta_0]$ *se puede encontrar también un test uniformemente más potente de* $H_0 : \theta = \hat{\theta}_0$ *frente a* $H_1 : \theta > \theta_0$*.*

Solución. Para resolver este problema bastaría con aplicar el razonamiento de verosimilitud monótona de los tests unilaterales, o simplemente emplear que el test no depende de θ_1.

Ejemplo 2.17. Sea el caso anterior, donde se quiere contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, tomando $n = 15$, $\alpha = 0{,}05$ y $\theta_0 = 2$, y mirando en la tabla de la distribución χ^2 obtenemos que

$$c'' = \frac{43{,}773 \cdot 2}{2} \Rightarrow c'' = 43{,}773$$

Por lo tanto, el test UMP rechazaría $H_0 : \theta \leq 2$ si $X_1 + X_2 + ... + X_{15} > 43{,}773$, o bien si $\bar{x} > 2{,}918$.

En este caso, fijándonos en (2.14):

$$\alpha = \sup_{\theta \in (0,\theta_0)} P_\theta \left\{ \sum_{i=1}^{n} X_i > c'' \right\} = P_{\theta_0} \left\{ \sum_{i=1}^{n} X_i > c'' \right\} \tag{2.17}$$

(la misma que para el caso en el que $H_0 : \theta = \theta_0$) luego la función de potencia del test UMP es la representada en la Figura 2.2. En ella se puede ver que cuando $\theta = \theta_0$ la potencia toma el valor α, que era nuestro requisito para que el test fuera UMP. Además, se puede apreciar que a medida que crece el valor de θ la potencia se va acercando a 1, lo que quiere decir que según crece θ la probabilidad de rechazar la hipótesis nula también crece.

Ahora bien, si la idea es comparar el test de la T de student con el anterior, solo tendríamos que comparar ambas funciones de potencia y ver cuál es mejor. En este caso, como lo que queremos contrastar es la media, se tomará $H_0 : \mu_0 \leq 2$ y $H_1 : \mu > 2$.

Mediante simulaciones podemos aproximar la función de potencia para el test de la T de student si los datos proviniesen de la distribución exponencial. En la Figura 2.2 se ve reflejada esta situación. En primer lugar, se puede apreciar que la función de potencia dada por el test de la T de student (TS, en rojo) no respeta el nivel α, por lo que ya hay indicios de que el test no es el correcto. Además, la función de potencia del test T se encuentra siempre por debajo de la del test de Neymann-Pearson. Esto puede ser debido a que las funciones de densidad de la exponencial y de la normal difieren mucho (véase la Figura 2.3).

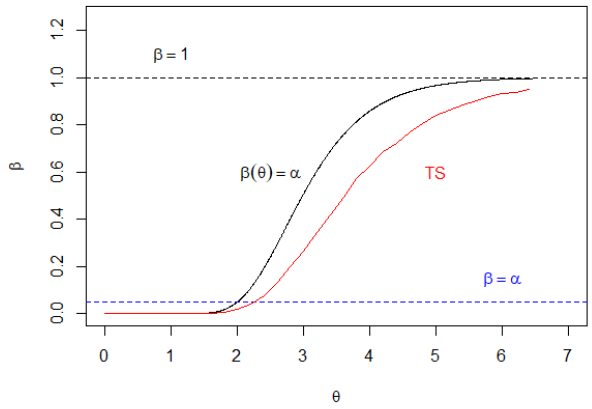

Figura 2.2: Función de potencia del test de Neyman-Pearson frente a la funciónde potencia simulada del test de la T de Student cuando la muestra sigue una distribución $Exp(\theta)$.

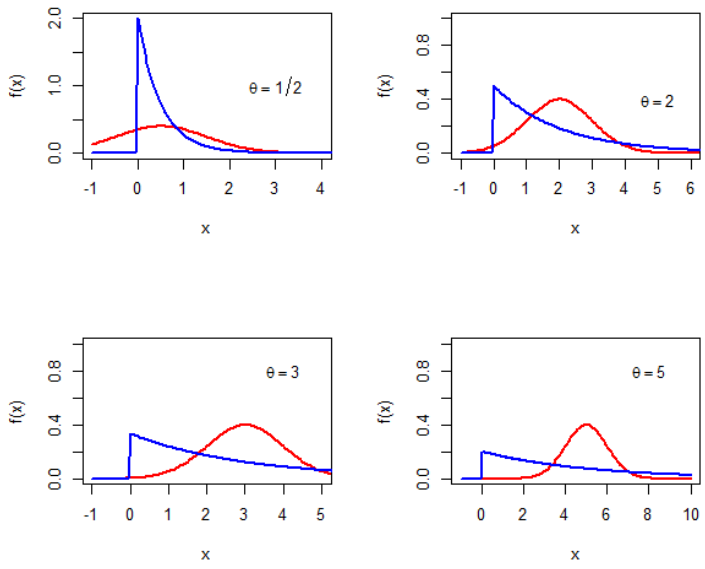

Figura 2.3: Funciones de densidad de $Exp(\theta)$ (azul) y $N(\mu = \theta, \sigma^2 = 1)$ (rojo) para distintos valores de θ.

Por ello, podemos decir que el test de la T de student es pésimo para hacer contrastes de hipótesis sobre la media cuando los datos provienen de una distribución exponencial, con un tamaño muestral de $n = 15$.

2.3.2. Caso uniforme

Problema 2.18. *Sea X una muestra aleatoria simple con X_i independientes entre sí siguiendo una distribución uniforme en el intervalo $(0,\theta)$ donde θ es desconocida. Encontrar el test más potente para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ con $\theta_1 > \theta_0$.*

Solución. Por el Lema de Neyman-Pearson, el test φ más potente a un nivel α para el problema planteado deberá ser de la forma:

$$\varphi(X_1,...,X_n) = \begin{cases} 1, \text{ si } f_{\theta_1}(X_1,...,X_n) > k f_{\theta_0}(X_1,...,X_n) \\ \gamma, \text{ si } f_{\theta_1}(X_1,...,X_n) = k f_{\theta_0}(X_1,...,X_n) \\ 0, \text{ si } f_{\theta_1}(X_1,...,X_n) < k f_{\theta_0}(X_1,...,X_n) \end{cases} \tag{2.18}$$

y cumplir que

$$E_{\theta_0}\varphi(X_1,...,X_n) = \alpha$$

Ahora, $\forall i \in \{1,...,n\}$ la función de densidad es

$$f_\theta(x_i) = \begin{cases} 1/\theta, \text{ si } x_i \in (0,\theta) \\ 0, \text{ en otro caso} \end{cases} \tag{2.19}$$

y por ser $X_1,...,X_n$ independientes

$$f_{\theta_0}(X_1,...,X_n) = \prod_{i=1}^{n} f_{\theta_0}(x_i) = \prod_{i=1}^{n} \frac{1}{\theta_0}.\ I_{\{x_i \in (0,\theta_0)\}}$$

$$f_{\theta_1}(X_1,...,X_n) = \prod_{i=1}^{n} f_{\theta_1}(x_i) = \prod_{i=1}^{n} \frac{1}{\theta_1}.\ I_{\{x_i \in (0,\theta_1)\}}$$

Luego, considerando $X_{(1)} = \text{mín}(X_1,...,X_n)$ y $X_{(n)} = \text{máx}(X_1,...,X_n)$

$$f_{\theta_0}(X_1,...,X_n) = \begin{cases} \dfrac{1}{(\theta_0)^n}, \text{ si } 0 \le X_{(1)} \le X_{(n)} \le \theta_0 \\ 0, \text{ en otro caso} \end{cases} \tag{2.20}$$

$$f_{\theta_1}(X_1,...,X_n) = \begin{cases} \dfrac{1}{(\theta_1)^n}, \text{ si } 0 \le X_{(1)} \le X_{(n)} \le \theta_1 \\ 0, \text{ en otro caso} \end{cases} \tag{2.21}$$

Se pueden dar entonces cuatro casos:

I. $X_{(1)} < 0 \Longrightarrow f_{\theta_0}(X) = 0$ y $f_{\theta_1}(X) = 0$

II. $X_{(1)} \ge 0$ y $X_{(n)} \le \theta_0 \Longrightarrow f_{\theta_0}(X) = \frac{1}{(\theta_0)^n}$ y $f_{\theta_1}(X) = \frac{1}{(\theta_1)^n}$

III. $X_{(1)} \ge 0$ y $\theta_0 < X_{(n)} \le \theta_1 \Longrightarrow f_{\theta_0}(X) = 0$ y $f_{\theta_1}(X) = \frac{1}{(\theta_1)^n}$

IV. $X_{(1)} \geq 0$ y $X_{(n)} > \theta_1 > \theta_0 \Longrightarrow f_{\theta_0}(X) = 0$ y $f_{\theta_1}(X) = 0$

Ahora veamos para qué valores de k se cumple $f_{\theta_1}(X) > kf_{\theta_0}(X)$, $f_{\theta_1}(X) = kf_{\theta_0}(X)$ o bien $f_{\theta_1}(X) < kf_{\theta_0}(X)$ según nos encontremos en los casos i), ii), iii) o iv):

Caso	$f_{\theta_1}(X) > kf_{\theta_0}(X)$	$f_{\theta_1}(X) = kf_{\theta_0}(X)$	$f_{\theta_1}(X) < kf_{\theta_0}(X)$
i)	Nunca	Siempre	Nunca
ii)	si $k < (\theta_0/\theta_1)^n$	si $k = (\theta_0/\theta_1)^n$	si $k > (\theta_0/\theta_1)^n$
iii)	Siempre	Nunca	Nunca
iv)	Nunca	Siempre	Nunca

Para satisfacer 2.18, mirando en la tabla anterior y separando por casos:

En el caso iii) tenemos que $\varphi(X) = 1$ independientemete del valor escogido para k, es decir, que rechaza siempre. En los casos i) y iv) podemos escoger $\varphi(X) = 1$ ó $\varphi(X) = 0$ sea cual sea el valor de k, ya que estos casos se encuentran fuera de los soportes. Por ello, nos limitaremos a partir de ahora a estudiar únicamente los casos ii) y iii). En el caso ii) separaremos en tres subcasos:

1. $k < (\theta_0/\theta_1)^n \Longrightarrow \varphi(X) = 1$
2. $k = (\theta_0/\theta_1)^n \Longrightarrow \varphi(X) = 1$ ó $\varphi(X) = 0$
3. $k > (\theta_0/\theta_1)^n \Longrightarrow \varphi(X) = 0$

Ahora bien, bajo H_0:

$$P_{\theta_0}\left[\text{Caso } iii)\right] = 0 \tag{2.22}$$

Por lo tanto, aunque rechazásemos una observación que cayese en el caso iii), el nivel siempre sería 0. Luego, un rechazo en el caso ii) a nivel α debería ser posible. Como $k > (\theta_0/\theta_1)^n$ no lo permite y $k < (\theta_0/\theta_1)^n$ siempre rechaza, entonces necesitaremos que $k = (\theta_0/\theta_1)^n$ para obtener algún α con $(0 < \alpha < 1)$, es decir, **aleatorizaremos el caso ii) con probabilidad α de rechazo**.

Veamos entonces que valor toma φ en el caso ii) para verificar

$$E_{\theta_0}\varphi(X) = \alpha$$

Un modo de proceder es tomar $\varphi(X) = \gamma$ para todos los $X_1, ..., X_n$ en el caso ii):

$$E_{\theta_0}\varphi(X) = \gamma \cdot P_{\theta_0}\left[\text{Caso ii)}\right] + 1 \cdot P_{\theta_0}\left[\text{Caso } i), \text{Caso } iii), \text{Caso } iv)\right] = \gamma + 0 = \gamma \tag{2.23}$$

Se toma entonces $\gamma = \alpha$.

Así, el test más potente con nivel α para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$ está dado por:

$$\varphi(X_1, ..., X_n) = \begin{cases} 1, \text{ si } X_{(n)} = \text{máx}(X_1, ..., X_n) > \theta_0 \\ \gamma, \text{ si } X_{(n)} < \theta_0 \end{cases} \tag{2.24}$$

La función de potencia del test se calcula de la siguiente manera:

para $0 < \theta \leq \theta_0$

$$\begin{aligned} E_\theta \varphi(X) &= \alpha \cdot P_\theta \left[X_{(n)} \leq \theta_0\right] + 1 \cdot P_\theta \left[X_{(n)} \in (\theta > \theta_0)\right] \\ &= \alpha \cdot 1 + 1 \cdot 0 = \alpha \end{aligned}$$

y para $\theta > \theta_0$

$$\begin{aligned} E_{\theta_0} \varphi(X) &= \alpha \cdot P_\theta \left[X_{(n)} \leq \theta_0\right] + 1 \cdot P_\theta \left[X_{(n)} \in (\theta_0, \theta)\right] \\ &= \alpha \cdot \left(\frac{\theta_0}{\theta}\right)^n + 1 \cdot \left(1 - \left(\frac{\theta_0}{\theta}\right)^n\right) = 1 - (1 - \alpha)\left(\frac{\theta_0}{\theta}\right)^n \end{aligned}$$

Por lo que

$$E_{\theta_0} \varphi(X) = \alpha \quad \text{y} \quad E_{\theta_1} \varphi(X) = 1 - (1 - \alpha)\left(\frac{\theta_0}{\theta}\right)^n \tag{2.25}$$

Además, el Lema de Neyman-Pearson asegura que cualquier otro test φ^* con $E_{\theta_0} \varphi^*(X) = \alpha$ tiene potencia

$$E_{\theta_1} \varphi^*(X) \leq E_{\theta_1} \varphi(X)$$

Ejemplo 2.19. Supongamos que se quiere encontrar el test más potente a un nivel $\alpha = 0{,}05$ para contrastar $H_0 : \theta = 1$ frente a $H_1 : \theta = 3/2$, donde el tamaño de la muestra es $n = 5$. Entonces la función de potencia para los distintos valores de θ tendrá la forma que se muestra en la Figura 2.4.

En efecto se puede ver que $\forall \theta \in \Theta_1$ la potencia del test de la T de student (TS, en verde) es menor que la potencia para el test de Neyman-Pearson (en rojo). Esto era esperable ya que el de Neyman-Pearson es el más potente al nivel α fijado. Además, la potencia de este test crece muy rápido, lo que quiere decir que con alejarse un poco de θ_0, enseguida el test rechazará la hipótesis nula. Por ejemplo, fijándonos en la figura, si θ toma un valor cercano a $3/2$ el test más potente rechazará con una probabilidad próxima a 1, sin embargo, el test de la T de student rechazará con menos de un 50 % de posibilidades.

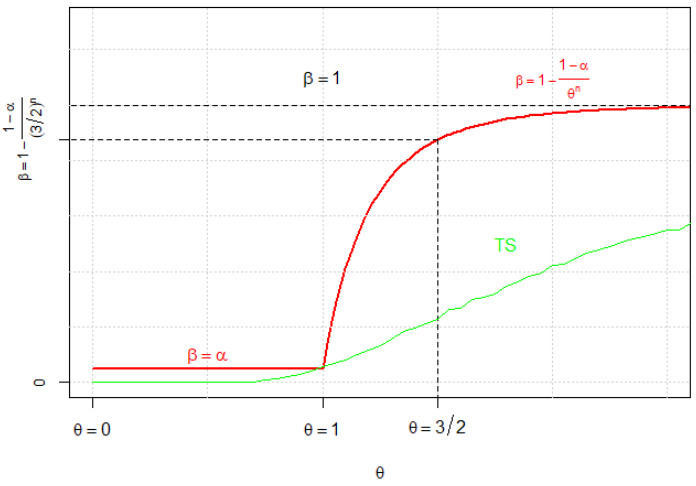

Figura 2.4: Funciones de potencia. En rojo, la del test de Neymann-Pearson, y en verde, la del test T (realizada mediante simulaciones).

Observación 2.20. Recordemos que si $X \sim Unif(a, b)$ se tiene que $E[X] = \frac{a+b}{2}$. En el ejemplo se está considerando que $X \sim Unif(0, \theta_0 = 1)$, con lo cual $E[X] = 1/2$. Por lo tanto se realizará el contraste $H_0 : \mu \leq 1/2$ frente a $H_1 : \mu > 1/2$ (el código de R para la realización de la Figura 2.4 se encuentra en el anexo, donde se ha considerado una cantidad de 10000 muestras simuladas con un tamaño $n = 5$).

Supongamos ahora que el tamaño muestral es mayor, considerando $n = 15$ y $n = 50$. En la Figura 2.5 se muestra cómo la potencia del test de la T de Student mejora notablemente según aumenta el tamaño muestral, pero aún así seguirá siendo peor que la del test de Neymann-Pearson, ya que a medida que se mejore el tamaño muestral, también lo hará su función de potencia, superando con creces la potencia del test T.

Como conclusión se puede decir que, a diferencia del caso exponencial, el test de la T de student no funciona tan mal para hacer contraste de medias sobre una distribución uniforme en el intervalo $(0, \theta)$. Aún así, está lejos de ser el test óptimo.

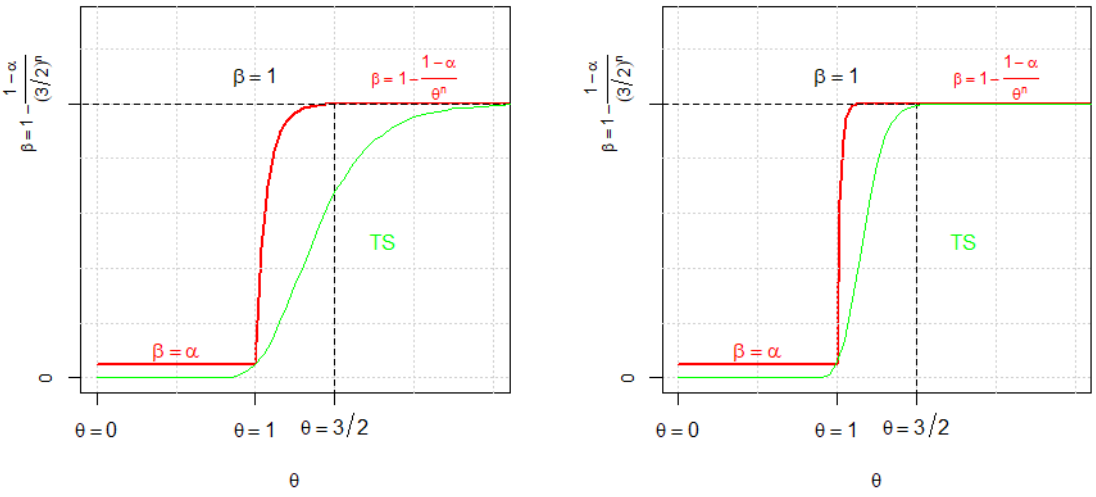

Figura 2.5: Comparativa del test de Neymann-Pearson (en rojo) frente al test de la T de Student (en verde) para el caso de una población uniforme en el intervalo $(0, \theta)$ donde los tamaños muestrales son $n = 15$ (izquierda) y $n = 50$ (derecha).

Capítulo 3

Contrastes sobre medidas de posición central

En capítulos anteriores hemos visto algunos test para realizar contrastes sobre la media de una población. Es de interés el preguntarse si podemos realizar contrastes sobre alguna otra medida de posición central y ver qué relación tiene con la media.

Las medidas de posición nos facilitan información sobre los datos que estamos analizando. Si nos encontrásemos en un contexto no paramétrico, donde la población de la que provienen los datos es desconocida, el hacer contrastes sobre la media sería un problema complicado. Entre los test no paramétricos, uno de gran importancia es el **test de la mediana**, ya que la mediana es una medida con propiedades más universales que la media.

En este capítulo estudiaremos primeramente si en poblaciones simétricas, en las cuales la mediana coincide con la media, hay alguna relación entre el test de la mediana y el test de la T de Student al realizar contrastes sobre la media.

En segundo lugar se expondrá el **test de Wilcoxon de rangos con signos (RSW)** y, para terminar, se estudiarán las funciones de potencia de dichos tests comparándolas con la del test T.

3.1. Test de la Mediana

3.1.1. Presentación del test

Supongamos que una muestra aleatoria simple procede de una distribución continua desconocida. El objetivo es ver si la mediana de esa distribución (m) toma un valor m_0. Se tratará entonces de contrastar la hipótesis nula

$$H_0 : m = m_0 \tag{3.1}$$

Si m_0 fuera la mediana de esta distribución teórica, en una muestra procedente de esta población, la mitad de las observaciones se encontrarían por encima de la mediana y la otra mitad por debajo, salvo una pequeña desviación aleatoria. Entonces, si la división que m_0 induce en la muestra deja porcentajes de elementos por encima y por debajo muy distintos al 50 %, habría evidencias significativas para rechazar H_0. Con ayuda del estadístico

$$T = \text{nº de observaciones muestrales menores que } m_0 \tag{3.2}$$

el test consistirá en rechazar $H_0 : m = m_0$ si T es muy diferente de $n/2$. Ahora bien, para saber cuándo T difiere mucho nos fijaremos que bajo H_0 su distribución es una $Binomial(n, 1/2)$. Luego el problema se resuelve como el contraste de una proporción, ya que la hipótesis nula es equivalente a $H_0 : p = 1/2$ con $p = P(X < m_0)$, siendo la proporción muestral $\hat{p} = T/n$. Por lo tanto, si n es pequeño se resolverá con las tablas de la binomial y si n es grande se recurrirá a la aproximación normal:

$$\frac{T/n - 1/2}{\sqrt{\frac{1/2(1-1/2)}{n}}} \sim N(0, 1)$$

3.1.2. Extensión a otros cuantiles

El siguiente teorema muestra que la idea del test de la mediana se puede extender a cualquier cuantil, y dependiendo del tipo de hipótesis el test será UMP o UMP entre los centrados.

Definición 3.1. Dada una constante k, se define a la variable aleatoria $S_n^+(X)$ como el número de elementos de $\{X_1 - k, ..., X_n - k\}$ que son positivos, es decir:

$$S_n^+(X) = \sum_{i=1}^{n} I_{(k,\infty)}(X_i) \tag{3.3}$$

Teorema 3.2. *Sea una población absolutamente continua, cuyo cuantil de orden p es c_p.*

Si se desea contrastar $H_0 : c_p \leq k$ *frente a* $H_1 : c_p > k$*, el test*

$$\varphi(X_1, ..., X_n) = \begin{cases} 1, \ si \ S_n^+(X) > c \\ \gamma, \ si \ S_n^+(X) = c \\ 0, \ si \ S_n^+(X) < c \end{cases} \tag{3.4}$$

es UMP, donde las constantes $\gamma \in \mathbb{R}$ *y* $c \in \mathbb{N}$ *se hallan mediante la ecuación siguiente, que refleja que el nivel del test ha de ser* α*.*

$$\sum_{r=c+1}^{n} \binom{n}{r}(1-p)^r p^{n-r} + \gamma \binom{n}{c}(1-p)^c p^{n-c} = \alpha \tag{3.5}$$

Por otro lado, si se quiere contrastar $H_0 : c_p = k$ *contra* $H_1 : c_p \neq k$*, el test*

$$\varphi(X_1, ..., X_n) = \begin{cases} 1, \ si \ S_n^+(X) < c_1 \ \acute{o} \ S_n^+(X) > c_2 \\ \gamma_i, \ si \ S_n^+(X) = c_i \ con \ (i = 1, 2) \\ 0, \ si \ c_1 < S_n^+(X) < c_2 \end{cases} \tag{3.6}$$

es UMP entre los centrados, donde c_i, γ_i *son constantes determinadas a partir de las ecuaciones:*

$$\sum_{r=c_1+1}^{c_2-1} \binom{n}{r}(1-p)^r p^{n-r} + \sum_{i=1}^{2}(1-\gamma_i)\binom{n}{c_i}(1-p)^{c_i} p^{n-c_i} = 1-\alpha \tag{3.7}$$

$$\sum_{r=c_1+1}^{c_2-1} \binom{n-1}{r-1}(1-p)^r p^{n-r} + \sum_{i=1}^{2}(1-\gamma_i)\binom{n-1}{c_i-1}(1-p)^{c_i} p^{n-c_i} = 1-\alpha \tag{3.8}$$

Nota*: Para muestras de tamaño grande se puede utilizar la aproximación por la distribución normal.*

Observación 3.3. Nótese que en el teorema anterior se supone que la probabilidad de que $X_i - k = 0$ para algún i es nula, ya que la población es absolutamente continua. Nos podemos preguntar entonces qué ocurre para aquellas observaciones las cuales verifican $X_i - k = 0$. La solución sería realizar el test dos veces: una incluyendo en S_n^+ los $X_i = k$ y en la otra sin incluírlos. Si no hay apenas diferencias en las áreas de las colas de ambos la solución será casi la misma. Por el contrario, si las diferencias son notorias, el test no es fiable.

El teorema 3.2 es aplicable a cualquier cuantil, por lo que también podrá ser usado para la prueba de la mediana. A continuación se mostrarán dos ejemplos: el primero, sobre la mediana, donde el tamaño muestral será 10 (moderado, no se podrá usar la aproximación normal); y el segundo, también sobre la mediana, pero esta vez con un tamaño muestral de 80, dónde sí se aplicará la aproximación normal.

Ejemplo 3.4. Supongamos que en una auditoría la mediana de las cuentas financieras que cada trabajador es capaz de supervisar a la semana está establecida en 6 cuentas. En la auditoría se hace un experimento con el fin de mejorar la eficacia de las horas trabajadas, consistente en introducir mejoras de software en el sistema informático. Estas nuevas mejoras se aplican al software de 10 trabajadores, y se obtiene que estos trabajadores supervisaron 4,5,7,7,8,9,11,12,13,15 cuentas respectivamente. Se quiere contrastar la hipótesis H_0 de que $C_{0,5} \leq 6$ contra $H_1 : C_{0,5} > 6$, a un nivel $\alpha = 0{,}1$.

Solución. Aplicando el teorema 3.2, la igualdad (3.5) da lugar a :

$$\sum_{r=c+1}^{10} \binom{10}{r} \left(1 - \left(\frac{1}{2}\right)\right)^r \left(\frac{1}{2}\right)^{10-r} + \gamma \binom{10}{c} \left(1 - \left(\frac{1}{2}\right)\right)^c \left(\frac{1}{2}\right)^{10-c}$$

$$= \sum_{r=c+1}^{10} \binom{10}{r} 2^{-10} + \gamma \binom{10}{c} 2^{-10} = 0,10$$

obteniéndose las constantes $c = 7$ y $\gamma = 0{,}39$. Como $S_n^+(x) = 8$, se rechaza la hipótesis nula a favor de la alternativa, a un nivel $\alpha = 0{,}1$, de que la mediana es mayor al implementar el nuevo software. Es, decir, en términos de la mediana, el nuevo software mejora la efectividad de las horas trabajadas.

Ejemplo 3.5. En un gimnasio la mediana del número de veces que cada socio va al gimnasio (m_0) está establecida en 3. El objetivo de los dueños es determinar si la mediana es superior a 3, y para ello se le pregunta a 80 socios, de los cuales 50 van más de 3 veces al gimnasio. ¿Se puede afirmar con pruebas significativas que la mediana es superior a 3 a un nivel $\alpha = 0{,}05$?

Solución. En este caso, como el tamaño muestral es bastante grande ($n = 80$), utilizaremos la aproximación normal.

Será equivalente a contrastar $H_0 : p \leq 1/2$ contra $H_1 : p > 1/2$, siendo $p = P(X > m_0)$.

El estadístico de contraste es

$$\frac{50/80 - 1/2}{\sqrt{\frac{1/2(1-1/2))}{80}}} = 2,2361$$

y se cumple que

$$3,35 > 1,6448 = z_\alpha$$

Además, el nivel crítico es 0.0127, luego hay pruebas significativas para rechazar la hipótesis nula y concluir que la mediana es mayor que 3 a un nivel $\alpha = 0{,}05$ (a un nivel $\alpha = 0{,}01$ ya no se podría rechazar H_0).

3.2. Test de la mediana vs. test de la t de Student en poblaciones simétricas

Una vez que se ha explicado el test de la mediana para el contexto no paramétrico, podríamos pensar que en una población desconocida la mediana podría estar muy cerca de la media real de dicha población, pero esto sólo ocurriría en determinados tipos de poblaciones. Un requisito que garantiza que la media y la mediana coincide es que las funciones de densidad sean simétricas.

3.2.1. La distribución de Laplace

Definición 3.6. Una variable aleatoria X posee una distribución de Laplace(μ, λ) si tiene como función de densidad:

$$f_{(\mu,\lambda)}(x) = \frac{1}{2\lambda} \cdot e^{-\frac{|x-\mu|}{\lambda}} = \begin{cases} \frac{1}{2\lambda} e^{-\frac{\mu-x}{\lambda}}, & \text{si } x < \mu \\ \frac{1}{2\lambda} e^{-\frac{x-\mu}{\lambda}}, & \text{si } x \geq \mu \end{cases} \tag{3.9}$$

donde μ es un parámetro de localización y $\lambda > 0$ un parámetro de escala. Por ser esta distribución simétrica, las medidas de posición central (media, mediana y moda) toman el mismo valor, μ.

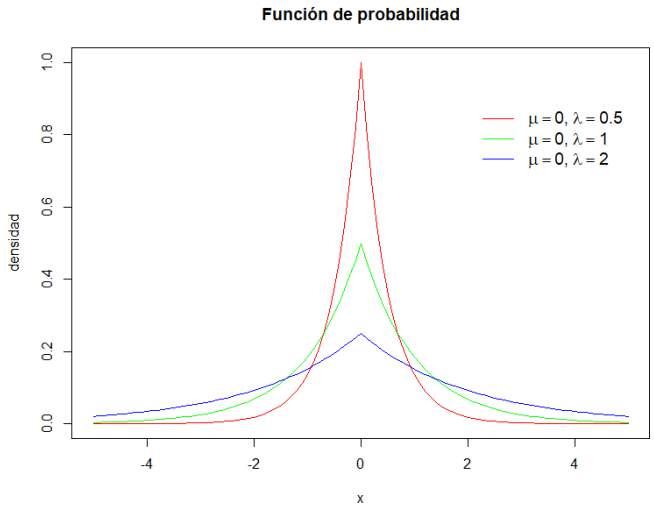

Figura 3.1: Función de probabilidad de la distribución de Laplace para $\mu = 0$ y distintos valores $\lambda = 0.5, 1, 2$.

Ejemplo 3.7. Se considera que la muestra X de tamaño $n = 30$ sigue una distribución de Laplace de parámetros μ desconocido y $\lambda = 1$, y se quiere realizar el contraste $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ con $\mu_0 = 0$. Entonces, podríamos entender que realizar un contraste sobre la media sería equivalente a hacer el mismo contraste sobre la mediana.

Realizando simulaciones obtenemos que las funciones de potencia de ambos test son las representadas en la Figura 3.2. Se puede apreciar que en efecto ambas funciones de potencia son muy semejantes y que respetan muy bien el nivel α cuando $\mu = \mu_0$. Además, por la forma de las funciones, parece que ambos test no son nada malos a la hora de realizar el contraste. Por ejemplo, si $\mu = 1$, ambos rechazarían casi con toda probabilidad la hipótesis nula.

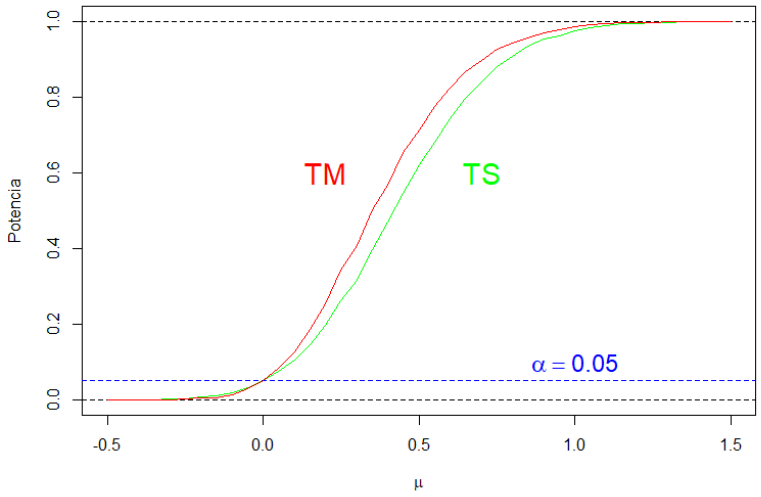

Figura 3.2: Función de potencia para el test de la t de student (TS, en verde) y para el test de la mediana (TM, en rojo) cuando $X \in Laplace(0, 1)$

Por otro lado, es curioso ver que a pesar de que ambos test realizan decentemente el contraste, el test de la mediana es algo más potente (aunque se hayan realizado simulaciones para su representación, ésta no va a variar apenas de la función real, ya que estamos hablando de un número de simulaciones muy elevado, 10000). Por ello, en este caso, el test recomendado para realizar el contraste sería el de la mediana.

3.2.2. Caso normal

Ya hemos visto en el ejemplo anterior que cuando las observaciones provienen de una población simétrica, realizar un contraste sobre la mediana es equivalente a realizar un contraste sobre la media. Además, el test de la mediana se comporta muy bien en estas situaciones, pero, ¿qué ocurre si los datos proviniesen de una normal, que también es simétrica? Pues bien, en el capítulo anterior se ha probado que el test de la T de Student es el óptimo para esta situción, lo que se traduce en que la función de potencia es mejor en el caso de la prueba T de Student. Aún así, es de interés ver qué diferencia hay entre estos tests.

Consideremos el ejemplo anterior, pero ahora suponiendo que los X_i siguen una distribución normal. En este caso las funciones de potencia para dichos test aparecen representadas en la Figura 3.3.

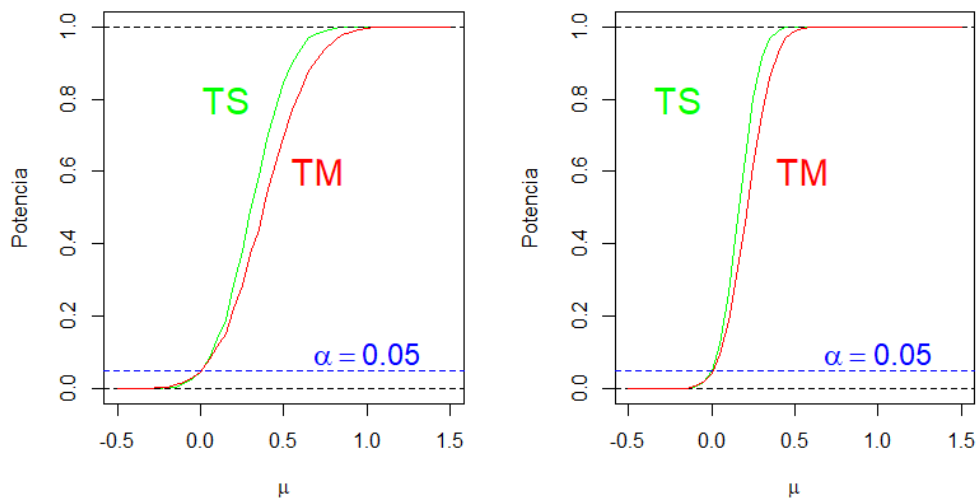

Figura 3.3: Funciones de potencia (simuladas) para el test de la T de student (TS, en verde) y para el test de la mediana (TM, en rojo) cuando $X \in N(\mu, 1)$. A la izquierda, tamaño muestral $n = 30$ y a la derecha, $n = 100$.

A diferencia del caso en que los datos provenían de una distribución de Laplace, ahora se puede ver cómo el test de la mediana sigue siendo bastante bueno, aunque sobrepasado por el test de la T de student. Además, si aumentara el tamaño muestral (en la figura, con $n = 100$), la diferencia seguiría siendo perceptible. Evidentemente, si el tamaño muestral fuera muy grande, la diferencia entre ambos tests sería mínima, en términos absolutos (si

dilatásemos el eje de μ el aspecto cambia).

Por otra parte es necesario añadir que este caso no es como el anterior. El test de la mediana siempre respetará el nivel, sin embargo, el test de la T de Student respetará el nivel cuando la distribución sea normal; cuando no es normal podría no respetar dicho nivel.

3.3. Contraste de Wilcoxon de rangos con signos (RSW)

El test de la mediana descrito anteriormente considera si las observaciones muestrales están por encima o por debajo de un valor de referencia, sin tener en cuenta las diferencias de las observaciones a dicho valor, lo que ocasiona una pérdida de información. El test de Wilcoxon de rangos con signos salva esa limitación, ya que provee un test de localización (y simetría) alternativo que tiene en cuenta las magnitudes de estas diferencias.

Sean $X_1, ..., X_n$ independientes e idénticamente distribuídas siguiendo una distribución continua F, simétrica respecto de la mediana m. El problema trata de contrastar $H_0 : m = m_0$ frente a las hipótesis alternativas usuales (unilaterales y bilaterales). Sin pérdida de generalidad, se asume que $m_0 = 0$, luego $F(-x) = 1 - F(x)\ \forall x \in \mathbb{R}$. Para contrastar $H_0 : F(0) = 1/2$ ó $H_0 : m = 0$, en primer lugar ordenamos $|X_1|, |X_2|, ..., |X_n|$ en un orden de magnitud creciente, asignando ***rangos***. Por ejemplo, si $n = 3$ y $|X_2| < |X_1| < |X_3|$, los rangos asignados serán: $|X_1|$ es 2, $|X_2|$ es 1 y $|X_3|$ es 3. Definimos

- T^+ = suma de los rangos procedentes de X_i's positivos.
- T^- = suma de los rangos procedentes de X_i's negativos.

Por lo tanto, bajo H_0 se puede esperar que T^+ y T^- sean iguales. Notemos que

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \tag{3.10}$$

luego ofrecerán un criterio equivalente.

El estadístico T^+ (o T^-) es conocido como el ***estadístico de Wilcoxon***. Un valor elevado de T^+ señalará que las mayores diferencias poseen un signo positivo, haciendo creer que la población no es simétrica respecto a m, estando ésta desplazada hacia la derecha, por lo que rechazaríamos $H_0 : m \leq 0$ en favor de la alternativa $H_1 : m > 0$. Un valor pequeño de T^- indicaría que las mayores diferencias tienen lugar por debajo de m, sugiriendo que está desplazada hacia la izquierda. Un análisis similar se aplica a las otras dos hipótesis alternativas:

- $H_0 : m \leq 0$, $H_1 : m > 0$: el test rechaza H_0 si $T^+ > c_1$
- $H_0 : m \geq 0$, $H_1 : m < 0$: el test rechaza H_0 si $T^+ < c_2$
- $H_0 : m = 0$, $H_1 : m \neq 0$: el test rechaza H_0 si $T^+ < c_3$ ó $T^+ > c_4$

El estadístico T de Wilcoxon consiste en la suma de todos los rangos, una vez asignado a cada uno de ellos el signo de la diferencia $X_i - m$. Denotando $r(|X_i|) = r_i$ como el rango de $|X_i|$ $(i = 1, ..., n)$ y a

$$Z_i = \begin{cases} 1, \text{ si } X_i > 0 \\ 0, \text{ si } X_i < 0 \end{cases} \tag{3.11}$$

escribiremos $T^+ = \sum_{i=1}^{n} r_i Z_i$ y $T^- = \sum_{i=1}^{n} (1 - Z_i) r_i$. Con esto,

$$\begin{aligned} T = T^+ - T^- &= -\sum_{i=1}^{n} r_i + 2\sum_{i=1}^{n} r_i Z_i \\ &= 2\sum_{i=1}^{n} r_i Z_i - \frac{n(n+1)}{2}. \end{aligned} \tag{3.12}$$

Por lógica, bajo la hipótesis nula de simetría respecto a m, la mitad de los rangos provendrían de diferencias de signo positivo y la otra mitad de signo negativo. Además, la probabilidad de ser positivo o negativo debería ser de 1/2. Por ello, el rango i-ésimo tomará valores i y $-i$ con probabilidad 1/2, por lo que tendrá esperanza 0, y lo mismo le pasará al estadístico T. La varianza de T coincidirá con la suma de los cuadrados de los n primeros números naturales:

$$Var(T) = \frac{n(n+1)(2n+1)}{24} \tag{3.13}$$

Ahora bien, existen dos formas de realizar la prueba de Wilcoxon. La primera consiste en utilizar la aproximación normal del estadístico T^+, ya que, para un n grande se satiface la condición de Lindeberg y se tiene que

$$\frac{T^+ - E_{H_0} T^+}{\sqrt{[n(n+1)(2n+1)]/24}} \longrightarrow N(0, 1) \tag{3.14}$$

Este estadístico utiliza solo los rangos de las diferencias positivas $(X_i - m)$, luego su suma será igual (en promedio) a la mitad de la suma de los n primeros números naturales:

$$E_{H_0}(T^+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4} \tag{3.15}$$

y su varianza es igual a la cuarta parte de la del estadístico T:

$$Var(T^+) = \frac{n(n+1)(2n+1)}{24} \tag{3.16}$$

Además, para resolver el contraste existen unas tablas que poseen los valores críticos. Si el estadístico muestral de Wilcoxon estuviera por debajo del valor crítico de dichas tablas, se rechazaría la hipótesis nula de simetría respecto de m_0. De todos modos, para tamaños muestrales elevados conviene realizar la aproximación del estadístico por la normal, ya que se simplificarían mucho los cálculos y el resultado sería practicamente el mismo.

Por otra parte, el contraste también podría resolverse utilizando el estadístico T, con la aproximación normal a su distribución, de media nula y varianza la descrita en (3.13).

En conclusión, la prueba consistente en el estadístico T^+ rechaza $H_0 : m \leq 0$ a favor de $H_1 : m > 0$ para un $\alpha \in (0, 1)$ fijado, cuando $T^+ \geq k$. El valor de k se puede determinar sabiendo la distribución del estadístico T^+ para tamaños muestrales pequeños tal que

$$P_{H_0}(T^+ \geq k) = \alpha \tag{3.17}$$

Observación 3.8. El test de Wilcoxon también puede ser usado para contrastar la simetría de una población respecto de un punto dado. Sea $X_1, ..., X_n$ una muestra aleatoria donde X_i son independientes e idénticamente distribuidas siguiendo una distribución continua F. Podemos definir la hipótesis nula mediante

$$H_0 : m = m_0 \text{ y } F \text{ simétrica respecto de } m_0 \tag{3.18}$$

y la alternativa como

$$H_1 : m \neq m_0 \text{ , o } F \text{ asimétrica} \tag{3.19}$$

Ejemplo 3.9. Se consideran los resultados que han sacado en un examen un grupo de 18 alumnos. Se quiere contrastar que dichas calificaciones proceden de una población simétrica con una mediana igual a 60.

Notas del examen	
Calificaciones	65 52 72 61 57 59 62 70 62 64 62 71 55 62 53 57 67 62
Diferencias $X_i - m$	5 -8 12 1 -3 -1 2 10 2 4 2 11 -5 2 -7 -3 7 2
Rangos	11,5 -15 17 1,5 -8,5 -1,5 5 15 5 10 5 16 -11,5 5 -13,5 -8,5 13,5 5

A la hora de asignar rangos, cuando una magnitud está repetida para la diferencia $X_i - m$, se asigna a cada una el promedio de los rangos que le corresponden a las observaciones. En este caso, la menor diferencia observada es 1, que aparece dos veces, a las que les

corresponderían los rangos 1 y 2. El promedio de ellas es $1'5$, que asignamos a cada una de las dos observaciones. La siguiente menor diferencia es 2, que aparece en cinco ocasiones, cuyo promedio es 5, luego asignamos 5 a estas cinco observaciones. Así sucesivamente hasta hallar todos los rangos.

El estadístico T^+ es igual a

$$T^+ = 11,5 + 17 + 1,5 + 5 + 15 + 5 + 10 + 5 + 16 + 5 + 13,5 + 5 = 109,5 \tag{3.20}$$

Mirando en la tabla, H_0 es rechazada a un nivel $\alpha = 0{,}05$ si $T^+ \geq 124$. Como $109,5 < 124$, aceptamos $H_0 : m = 60$ y F es simétrica respecto de 60.

Veamos que si aproximamos por la normal el estadístico T también llegaríamos al mismo resultado:

$$T^- = -15 - 8,5 - 1,5 - 11,5 - 13,5 - 8,5 = -58,5 \tag{3.21}$$

Entonces $T = T^+ - T^- = 109,5 - 58,5 = 51$, con esperanza cero y varianza :

$$Var(T) = \frac{18(18+1)(2 \cdot 18+1)}{6} = 2109 \tag{3.22}$$

Utilizamos el estadístico T tipificado (que sigue una normal estándar en muestras grandes):

$$\frac{T - E(T)}{\sqrt{Var(T)}} = \frac{51 - 0}{\sqrt{2109}} = 1,111 \tag{3.23}$$

y vemos que no aporta un nivel crítico menor del 5 %.

En conclusión, no se rechza la hipótesis nula de que la mediana es superior a 60, y además tampoco se rechaza la hipótesis de que la distribución sea simétrica alrededor de esta mediana.

3.4. Test RSW vs. test de la T de Student bajo distribuciones simétricas

En esta sección nuestro objetivo será ver qué ocurriría si comparásemos el test de la T de Student contra una aproximación del test de Wilcoxon de rangos con signos, que se describirá a continuación, en el caso de poblaciones simétricas. Para ello reutilizaremos el Ejemplo 3.7 de la sección 3.2.

Es importante señalar que la función de potencia exacta de la prueba del test de Wilcoxon no se puede obtener ya que la distribución exacta del estadístico T^+ no se puede escribir, bajo la hipótesis alternativa, de forma explícita. Por otra parte, será también

inviable realizar una aproximación por el Teorema del Límite Central dado que bajo H_1, T^+ no se puede representar como suma de variables aleatorias independientes. De todos modos, la potencia de este test puede ser aproximada para valores de m próximos a cero de la siguiente manera (Hettmansperger, 1984, pp. 59-62):

$$\begin{aligned}\beta(m) &= P(\text{Rechazar } H_0/H_1 \text{ cierta}) \\ &= P(T^+ \geq k) \\ &= 1 - \phi\left(\frac{k - E_{H_1}(T^+)}{\sqrt{Var_{H_1}(T^+)}}\right) \\ &\approx 1 - \phi\left(z_\alpha - \frac{n(n-1)m\int_{-\infty}^{\infty} f^2(x)\,dx + nmf(0)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right).\end{aligned} \tag{3.24}$$

con $m > 0$. Ahora bien, si los datos siguieran una distribución $N(0, \sigma^2)$ es fácil ver que

$$f(0) = \frac{1}{\sqrt{2\pi}\sigma} \tag{3.25}$$

y (3.24) quedaría así

$$P(T^+ \geq k) = 1 - \phi\left(z_\alpha - \sqrt{12}\left(\int f^2(x)\,dx\right) m\sqrt{n}\right) \tag{3.26}$$

siendo ésta la función de potencia asintótica local del test RSW.

3.4.1. Caso doble exponencial (Laplace)

Como ya habíamos visto anteriormente en este capítulo, el test de la mediana se comporta muy bien cuando se trata con poblaciones simétricas, y es de esperar que el test RSW se comporte igual o incluso mejor.

Trabajando con el Ejemplo 3.7 podremos ver en la Figura 3.4 que forma adquiere la función de potencia para el test de RSW aproximada por la fórmula (3.24), cuando trabajamos con una distribución de Laplace de parámetros $\mu = 0$ y $\lambda = 1$. En primera instancia hay que decir que la función de potencia aproximada para el test RSW se asemeja bastante a la real, por trabajar con valores cercanos a cero, ya que de no ser así la fórmula propuesta en (3.24) sería errónea.

En la Figura se puede apreciar cómo el test de Wilcoxon se parece bastante al de la mediana, siendo ambos mejores que la prueba T de Student. Además, es de interés ver que a medida que crece el tamaño de la muestra, más se parecen las funciones de potencia para ambos tests.

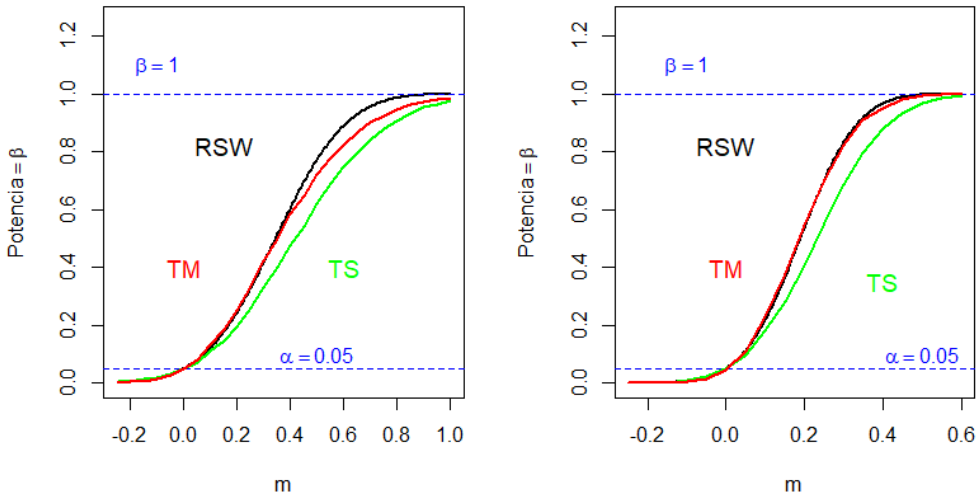

Figura 3.4: Funciones de potencia simuladas para el test de la T de student (TS, en verde), para el test de la mediana (TM, en rojo), y función de potencia para el test RSW (en negro) calculada mediante la fórmula 3.24, cuando $X \in Laplace(\mu = 0, \lambda = 1)$. Tamaños muestrales $n = 30$ y $n = 100$.

3.4.2. Caso Normal

Al igual que en el test de la mediana, supondremos ahora que se sigue una distribución normal, y veremos cómo funciona el test RSW comparado con el test de la T de Student. Para ello, nos apoyaremos de nuevo en el Ejemplo 3.7.

Se puede ver ahora en la Figura 3.5 que en efecto el test de la T de Student es el óptimo en cuanto a poblaciones normales se refiere. Por otro lado, al contrario que en el caso anterior, ahora el test RSW se ve superado por el test de la mediana. Además, el test de RSW se ve ahora muy afectado por el tamaño muestral, siendo la mejora notable cuando pasamos de un tamaño muestral de 30 a 100.

Observación 3.10. En la Figura 3.5, para la realización de la función de potencia del test RSW (en amarillo), se ha utilizado la aproximación dada en (3.26). En negro se ha representado la aproximación mediante la fórmula (3.24), como hemos hecho en el caso de la distribución de Laplace. Asimismo, como se puede ver en el gráfico, con $n = 30$ las funciones son casi iguales, y con $n = 100$ están prácticamente superpuestas.

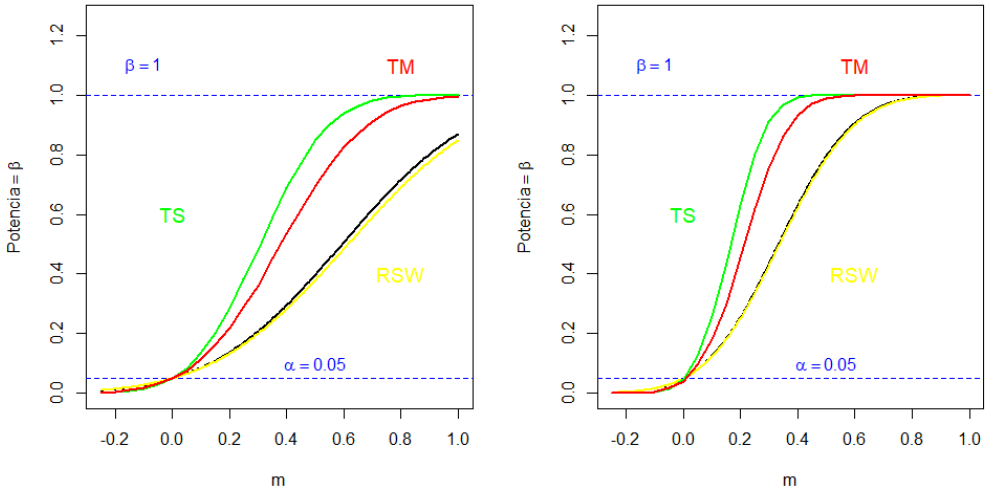

Figura 3.5: Funciones de potencia (simuladas) para el test de la T de student (TS, en verde), para el test de la mediana (TM, en rojo), y funciones de potencia aproximadas para el test RSW, en amarillo y en negro, calculadas mediante las fórmulas (3.26) y (3.24), respectivamente, cuando $X \in N(\mu = 0, \sigma^2 = 1)$. A la izquierda, tamaño muestral $n = 30$ y a la derecha, $n = 100$.

Como conclusión se puede decir que bajo distribuciones normales el test T es más potente que los test no paramétricos de la mediana y RSW. Por otra parte, es necesario reiterar que a pesar de utilizar poca información, estos tests son capaces de ser bastante potentes bajo hipòtesis de distribuciones simétricas, y en ciertas ocasiones superar al test T.

Sin embargo, si el objetivo es el estudio en distribuciones no simétricas, estos tests podrían remarcar la no simetría de dichas distribuciones, pero sin obtener resultados fiables sobre la media, algo que sí ocurre cuando hay simetría.

Capítulo 4

Comparación de dos poblaciones

En este capítulo abordaremos el problema de contrastar si dos poblaciones poseen características similares, viendo si es posible sacar conclusiones sin que haya un grado de duda severo. Para ello, dividiremos en dos casos: paramétricos y no paramétricos.

En primer lugar estudiaremos el caso normal, separando en casos dependiendo de si tienen igual o distinta variabilidad, y de si las muestras son emparejadas o independientes. La solución a este problema ya se dio en el capítulo 1, por lo que recordaremos los tests de tipo T de Student resultantes de manera breve.

En segundo lugar introduciremos dos pruebas para situaciones no paramétricas: el **test de los signos** y el test de **Wilcoxon-Mann-Whitney**.

4.1. Situaciones paramétricas. El caso normal.

Las siguientes tablas recogen las diferentes situaciones que se describieron en el primer capítulo, para el caso de dos poblaciones:

Muestras emparejadas

Consideremos dos variables aleatorias X e Y suponiendo que $X \in N(\mu_1, \sigma_1^2)$ e $Y \in N(\mu_2, \sigma_2^2)$ con σ_1 y σ_2 desconocidas, que se miden simultáneamente en cada individuo de una muestra de n individuos independientes.

Hipótesis nula	**Hipótesis alternativa**	**Estadístico de contraste**	**Región crítica**
$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 \neq \mu_2$	$t_0 = \frac{\bar{X}-\bar{Y}}{S_c/\sqrt{n}}$	$\lvert t_0 \rvert > t_{\alpha/2}$
$H_0 : \mu_1 \leq \mu_2$	$H_1 : \mu_1 > \mu_2$		$t_0 > t_\alpha$
$H_0 : \mu_1 \geq \mu_2$	$H_1 : \mu_1 < \mu_2$		$t_0 < t_\alpha$

Muestras independientes

Suponemos dos poblaciones normales con respectivas medias y varianzas: $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$. Se extrae de manera independiente una muestra aleatoria simple de cada población.

$\sigma_1 = \sigma_2$ desconocidas			
Hipótesis nula	**Hipótesis alternativa**	**Estadístico de contraste**	**Región crítica**
$H_0 : \mu_1 - \mu_2 = \delta_0$	$H_1 : \mu_1 - \mu_2 \neq \delta_0$	$t_0 = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$\lvert t_0\rvert > t_{n_1+n_2-2,\alpha/2}$
$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_1 : \mu_1 - \mu_2 > \delta_0$		$t_0 > t_{n_1+n_2-2,\alpha}$
$H_0 : \mu_1 - \mu_2 \geq \delta_0$	$H_1 : \mu_1 - \mu_2 < \delta_0$		$t_0 < t_{n_1+n_2-2,\alpha}$

$\sigma_1 \neq \sigma_2$ desconocidas (Behrens-Fisher)			
Hipótesis nula	**Hipótesis alternativa**	**Estadístico de contraste**	**Región crítica**
$H_0 : \mu_1 - \mu_2 = \delta_0$	$H_1 : \mu_1 - \mu_2 \neq \delta_0$	$t_0 = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S_p\sqrt{\frac{S_{c1}^2}{n_1}+\frac{S_{c2}^2}{n_2}}}$	$\lvert t_0\rvert > t_{\gamma,\alpha/2}$
$H_0 : \mu_1 - \mu_2 \leq \delta_0$	$H_1 : \mu_1 - \mu_2 > \delta_0$		$t_0 > t_{\gamma,\alpha}$
$H_0 : \mu_1 - \mu_2 \geq \delta_0$	$H_1 : \mu_1 - \mu_2 < \delta_0$		$t_0 < t_{\gamma,\alpha}$

Denotando γ como el entero más proximo al cálculo de

$$\frac{\left(S_{c1}^2/n_1 + S_{c2}^2/n_2\right)^2}{\frac{\left(S_{c1}^2/n_1\right)^2}{n_1-1} + \frac{\left(S_{c2}^2/n_2\right)^2}{n_2-1}}$$

4.2. Test de los signos

Un test usado frecuentemente es el test de los signos. Este test es muy útil para juzgar la efectividad de diversos tratamientos. Supongamos que en un grupo de n individuos se le efectúa a cada uno la medición de dos variables variables aleatorias, en diferentes condiciones. Por ejemplo, antes y despúes del tratamiento con un tipo de fármaco en diversos pacientes para reducir la presencia de células cancerígenas. Una pregunta obvia es si hubo alguna mejora en los pacientes después del tratamiento, es decir, si la segunda variable que se ha medido es de algún modo más pequeña que la primera.

A la hora de tomar una decisión nos apoyaremos en las observaciones de los n individuos independientes, recogidas como una muestra aleatoria simple de la forma

$$(X_1, Y_1), ..., (X_n, Y_n) \tag{4.1}$$

Para la comparación de los resultados de la variable X con los de Y se procede realizando el siguiente recuento:

$$T = \text{ Número de veces que X es mayor que Y} \tag{4.2}$$

luego

$$T = \sum_{i=1}^{n} I_{\{X_i > Y_i\}} \tag{4.3}$$

El estadístico T se puede ver como el número de signos positivos en las diferencias $X_1 - Y_1, ..., X_n - Y_n$. Si ambas variables X e Y fueran muy parecidas sería lógico que $P(X > Y) = 1/2$, valiendo T aproximadamente $n/2$. De no ser así, una variable sería mayor que la otra.

Por lo tanto, se puede plantear un problema de contraste de hipótesis de la forma

$$H_0 : P(X > Y) = 1/2 \tag{4.4}$$

que se resuelve por el procedimiento binomial visto para el test de la mediana en el tercer capítulo.

Ejemplo 4.1. En una empresa automovilística se está experimentando con un nuevo tipo de aceite para mejorar el rendimiento de los motores de sus coches. Con el objetivo de evaluar el rendimiento de los coches se han realizado dos pruebas en un circuito: una con el aceite estándar y otra con la nueva fórmula de aceite, obteniéndose los siguientes tiempos mínimos de vuelta en circuito (en segundos):

Coche	Aceite estándar	Aceite experimental
1	305	297
2	301	295
3	292	293
4	291	286
5	308	305
6	303	290
7	307	299
8	288	289

¿Se puede decir que la nueva fórmula de aceite mejora el rendimiento de los motores?

Solución. Suponiendo que los X_i son los tiempos de los coches con aceite estándar y los Y_i los tiempos con el aceite experimental, el problema se plantea definiendo las hipótesis

$$H_0 : P(Y > X) \geq 1/2 \qquad H_1 : P(Y > X) < 1/2$$

pues queremos probar que el tiempo con el aceite experimental es menor.

Haciendo el recuento,

$$T = \sum_{i=1}^{8} I_{\{X_i > Y_i\}} = 6$$

En este caso rechazaremos $H_0 : P(Y > X) \geq 1/2$ si T es mucho mayor que $n \cdot 0'5$. Como $n = 8$, se puede ver que $T = 6 > n/2 = 4$, luego hay indicios de que H_0 podría ser falsa. Mediante el siguiente código en R

```
n=8
exitos=6
binom.test(exitos,n,p=0.5,alternative="greater")
```

podremos obtener el valor crítico. En este caso se arroja un valor crítico de 0.1445, por lo que a los niveles usuales de significación (0.01, 0.05, 0.10) no podremos rechazar la hipótesis nula de que con el aceite experimental los tiempos de vuelta mínimos en circuito sean mayores que con el aceite estándar.

4.3. Test de Wilcoxon-Mann-Whitney

Ahora el objetivo será la comparación de dos variables aleatorias independientes entre sí. Al contrario que la situación anterior, nos encontramos ante dos muestras $X_1, ..., X_m$ e $Y_1, ..., Y_n$.

En 1945 Wilcoxon propuso el siguiente procedimiento:

1. Crear la muestra combinada juntando las dos muestras y ordenada de menor a mayor.

2. Hallar los rangos correspondientes.

3. Determinar la suma W de los rangos R_j de los Y_j en la muestra combinada para $j = 1, ..., n$, esto es

$$W = \sum_{j=1}^{n} R_j \tag{4.5}$$

A W se le conoce como estadístico de Wilcoxon.

Posteriormente en 1947 Mann y Whitney publicaron otro trabajo que da solución a este mismo problema, y aunque se tratase otro diferente al de Wilcoxon, los tests resultantes son equivalentes. Por ello a día de hoy se le conoce como test de Wilcoxon-Mann-Whitney.

El **estadístico de Mann-Whitney** se define de la siguiente forma:

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} T_{ij} \text{ con } T_{ij} = \begin{cases} 1, \text{ si } Y_j > X_i \\ 0, \text{ en otro caso} \end{cases} \tag{4.6}$$

Como se puede ver, este estadístico consiste en un recuento de todos los pares X_i, Y_j (nótese que hay $m \cdot n$ posibles) en los que Y_j es mayor que X_i.

Así pues, consideraremos que las dos variables son semejantes si el estadístico de Mann-Whitney toma un valor moderado, y por el contrario, que Y tiende a ser mayor que X si el estadístico tomase un valor elevado, o bien que Y es por lo general menor que X si este esadístico tomara un valor reducido.

Con esto, la hipótesis nula que vamos a contrastar será la misma que para el test de los signos:

$$H_0 : P(Y > X) = 1/2 \tag{4.7}$$

siendo en este caso X e Y variables independientes, a diferencia del caso anterior en el que estaban emparejadas.

Ahora bien, para realizar el contraste en base al estadístico de Mann-Whitney será necesario conocer su distribución bajo H_0. Esa distribución es conocida y se dispone de una tabla específica para tamaños muestrales reducidos. Esta tabla contiene los valores que dan lugar a rechazar en un contraste unilateral del tipo "rechazar si U es grande".

En el resto de casos recurriremos a las siguientes propiedades del estadístico de Mann-Whitney:

Propiedad 1: Denotando $U_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{Y_j > X_i\}}$ y a $U_{n,m} = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{X_i > Y_j\}}$ se cumplirá (salvo empates) la siguiente igualdad:

$$U_{m,n} + U_{n,m} = mn \tag{4.8}$$

Propiedad 2: Bajo la hipótesis nula de que X e Y poseen la misma distribución, el estadístico de Mann-Whitney tiene distribución simétrica en torno a $\frac{mn}{2}$. En particular,

$$P(U_{m,n} = k) = P(U_{m,n} = mn - k) \quad \forall k \in \{0, 1, ..., mn\} \tag{4.9}$$

Propiedad 3: Bajo la hipótesis nula de que X e Y tienen la misma distribución, $U_{m,n}$ y $U_{n,m}$ tienen la misma distribución, lo cual es consecuencia de las propiedades 1 y 2.

Por otro lado, se puede comprobar fácilmente que se cumple la igualdad; que relaciona los estadísticos de Wilcoxon y de Mann-Whitney:

$$U = W - \frac{n(n+1)}{2} \tag{4.10}$$

Ejemplo 4.2. Supongamos que se tienen los datos diagnósticos de 9 pacientes, de los cuáles 5 ($= n$) son mujeres y 4 ($= m$) son hombres. Todos ellos fueron diagnosticados con una determinada enfermedad. El objetivo es estudiar si existen diferencias entre las mujeres y los hombres que padecen dicha enfermedad, en términos de la edad. Las muestras son las siguientes:

$$\text{Mujeres (X): } 21, 19, 17, 28, 25$$

$$\text{Hombres (Y) : } 20, 12, 18, 13$$

Se definen las hipótesis:

$$H_0 : P(X > Y) \leq 1/2$$

$$H_1 : P(X > Y) > 1/2$$

Ordenando las edades

Edad:	12	13	17	18	19	20	21	25	28
Grupo:	Y	Y	X	Y	X	Y	X	X	X
Rango:	1	2	3	4	5	6	7	8	9

Con estos datos se obtiene que

$$W = \sum_{j=1}^{n} R_j = 3 + 5 + 7 + 8 + 9 = 32$$

luego

$$U = W - \frac{n(n+1)}{2} = 32 - 15 = 17$$

La prueba U de Mann-Whitney está disponible en R mediante la función *wilcoxon.test*:

```
> x=c(21,19,17,28,25)
> y=c(20,12,18,13)
```

```
> wilcox.test(x,y,alternative="greater")

Wilcoxon rank sum exact test

data:  x and y
W = 17, p-value = 0.05556
alternative hypothesis: true location shift is greater than 0
```

que efectivamente ofrece un valor de 17 para el estadístico de Mann-Whitney (en R se denota por W), como ya habíamos visto. Además, se obtiene un nivel crítico de 0.05556, por lo que a un nivel de significación del 0.10 se puede rechazar la hipótesis nula (pero no al 0.05).

En el caso de que los **tamaños muestrales** m **y** n **sean grandes**, se podrá efectuar la **aproximación por la distribución normal**:

$$\frac{U - mn/2}{\sqrt{mn(m+n+1)/12}} \sim N(0,1) \tag{4.11}$$

En consecuecia es equivalente usar un estadístico o el otro para el contraste. De todos modos, el estadístico de Wilcoxon es más sencillo de calcular que el de Mann-Whitney. Por este motivo es recomendable introducir en el algoritmo el cálculo del estadístico de Wilcoxon, hallar el estadístico de Mann-Whitney, y posteriormente valorar si es significativo para el contraste deseado.

4.3.1. Función de potencia para la prueba basada en el estadístico W

Supongamos dos variables aleatorias independientes entre sí definidas como en la sección anterior provenientes de distribuciones $F(x)$ y $G(y) = F(y - \theta)$ continuas.

Bajo la hipótesis nula $H_0 : \theta = 0$, la esperanza y la varianza para el estadístico W serán:

$$E(W) = \frac{n\,[(m+n)+1]}{2} \qquad Var(U) = \frac{mn\,[(m+n)+1]}{12} \tag{4.12}$$

y bajo la hipótesis alternativa $H_1 : \theta > 0$, la esperanza y la media para el estadístico W resultan:

$$E(W) = mnp_1 + \frac{n(n+1)}{2} \tag{4.13}$$

$$Var(W) = mn(p_1 - p_1^2) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2) \tag{4.14}$$

siendo

$$p_1 = P(Y > X) = \int F(y)g(y)\,dy = \int (1 - G(x))f(X)\,dx \tag{4.15}$$

$$p_2 = P(Y_1 > X_1, Y_2 > X_1) = \int (1 - G(x))^2 f(x)\, dx \tag{4.16}$$

$$p_3 = P(Y_1 > X_1, Y_1 > X_2) = \int (F(y))^2 g(y)\, dy \tag{4.17}$$

Además, el estadístico estandarizado tiene una distribución aproximada normal estándar:

$$\frac{W - E(W)}{\sqrt{Var(W)}} \longrightarrow N(0,1) \tag{4.18}$$

Así pues, con los elementos anteriores, la función de potencia basada en el estadístico de Wilcoxon W se puede aproximar (al igual que el de Mann-Whitney, visto anteriormente) por la distribución normal, y su expresión puede escribirse como:

$$\beta_W(\theta) = P_\theta(W \geq c) \approx \phi\left(\sqrt{\frac{12mn}{m+n}} f^*(0)\theta - z_\alpha\right) \tag{4.19}$$

donde α es el nivel de significación, z_α es la abscisa que cumple $\phi(z_\alpha) = 1 - \alpha$, c es tal que $P_{H_0}(W \geq c) = \alpha$ y

$$f^*(0) = \int_{-\infty}^{\infty} f^2(x)\, dx \tag{4.20}$$

Ejemplo 4.3. Consideremos dos muestras que provienen de poblaciones normales y que sean independientes, de tamaño muestral $n = 30$. La función de potencia para contrastar la hipótesis $H_0 : P(X \leq Y) \geq 1/2$ frente a $H_1 : P(X > Y) > 1/2$ a un nivel $\alpha = 0{,}05$ está representada en la Figura 4.1 (para el caso de la normal estándar). Como se puede observar, la potencia del test WMW es bastante buena, rechazando H_0 casi con toda probabilidad cuando $\theta = 1$, y además respetando el nivel α cuando $\theta = 0$.

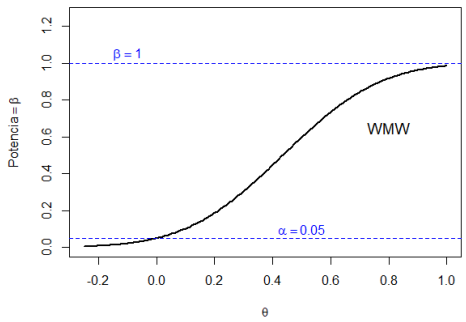

Figura 4.1: Función de potencia aproximada por la fórmula (4.19) para el test WMW cuando $X \in N(\mu = 0, \sigma^2 = 1)$ y $n = 30$ (muestras emparejadas).

4.3.2. Test T vs. Wilcoxon-Mann-Whitney

A diferencia de los test paramétricos, la potencia en los test no paramétricos es difícil de obtener, como ocurre en los test mencionados anteriormente.

Según (Dudewicz, Mishra, 1988, pp. 667-668), la eficiencia del test de Wilcoxon-Mann-Whitney relativa al test T para alternativas normales es en el caso asintótico de $3/\pi \approx 0{,}96$ y para el resto de casos es $\geq 0{,}864$. Además, en el caso normal (donde deberíamos usar el test T), si usáramos el test de Wilcoxon-Mann-Whitney, estaríamos descartando aproximadamente un 4 % de la información de la muestra; y para algunas alternativas no normales, podríamos descartar fácilmente un 14 % de la información. De todos modos, para otro tipo de alternativas el test de Wilcoxon-Mann-Whitney puede ser realmente mejor que el test T.

Por ello, la recomendación sería la siguiente: **si se tuviera una distribución normal (u otra distribución específica) para los datos, se debería usar un procedimiento paramétrico; y en caso de duda significativa, utilizar procedimientos no paramétricos**. Aún así, hay que remarcar que discernir entre estos dos casos es una cuestión para nada trivial, que sigue siendo objeto de estudio e investigación en la comunidad científica.

Código de R

Capítulo 1

Figura 1.1 : Función de densidad de la T de Student con k grados de libertad.

```
x=seq(-5,5,by=0.05)     # rejilla de valores
y_1=dt(x,1)
y_2=dt(x,2)
y_3=dt(x,5)
y_4=dt(x,10)
y_5=dt(x,1000)
y_6=dnorm(x)
plot(x,y_1,type="l",ylim=c(0,0.5),ylab=(expression(paste(f(x)))),
        xlab="x",col="yellow",lwd=2)
lines(x,y_2,type="l",col="orange",lwd=2)
lines(x,y_3,type="l",col="red",lwd=2)
lines(x,y_4,type="l",col="violet",lwd=2)
lines(x,y_5,type="l",col="blue",lwd=2)

legend(2, 0.4, expression(paste(k==1),
        paste(k==2), paste(k==5),paste(k==10),paste(k==Inf)),
        lty=c(1,1,1,1,1),bty="n",
        col=c("yellow","orange","red","violet","blue"),cex=1)

    # como se puede ver, cuando k tiende a infinito,
    # la t de Student tiene #la misma forma que la normal
lines(x,y_6,type="l",col="green",lwd=2)
```

Capítulo 2

Figura 2.1 : Función de potencia simulada del test T cuando la muestra procede de una distribución normal con $\sigma = 1$ para distintos valores de μ.

```
# H_0: mu<=0    Hipótesis nula
# H_a: mu>0     Hipótesis alternativa

n=20            # Tamaño muestral
ns=20000        # Número de simulaciones

vmedia=seq(-1/2,1,by=0.05)
vpotencia=c()
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=rnorm(n,mean=vmedia[imedia])

    # Test de la T de Student
    t=t.test(x,mu=0,alternative="greater")
    nct[is]=t$p.value

  }
  sum(nct<0.05)/ns  # Proporción de rechazos (T-Student)
  vpotencia[imedia]=sum(nct<0.05)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia[imedia],"\n")
}
plot(vmedia,vpotencia,ylim=c(0,1),type="l",
        xlab=expression(paste(mu)),ylab=expression(paste(beta)),
        lwd=2)
abline(h=0.05,lty=2,col="blue")
abline(h=1,lty=2,col="red")
text(0.5,0.1,expression(paste(alpha==0.05)),cex=1.5,col="blue")
text(-0.1,0.95,expression(paste(beta==1)),cex=1.5,col="red")
```

Figura 2.2 : Función de potencia del test de Neyman-Pearson frente a la función de potencia simulada del test de la T de Student cuando la muestra sigue una distribución $Exp(\theta)$.

```
# Representación de la función de potencia del
# test UMP de Neymann-Pearson

theta_0=2
n=15
theta=seq(0,6.5,length=1000)
alfa=0.05
chi=qchisq(1-alfa,df=2*n)
chi
potencia=1-pchisq(chi*theta_0/theta,df=2*n)
plot(theta,potencia,ylim=c(0,1.25),xlim=c(0,7),
        xlab=expression(theta),ylab=expression(beta),type="l")
abline(h=alfa,lty=2,col="blue")
abline(h=1,lty=2)

text(6,alfa+0.1,expression(paste(beta==alpha)),cex=1, col="blue")
text(x=1, y=1.1, expression(paste(beta==1)))
text(x=2.5, y=0.6, expression(paste(beta(theta)==alpha)))


# función de potencia para el test de la T de student
set.seed(123456)

# H_0: mu<=2
# H_a: mu>2

n=15
ns=20000

vmedia=seq(0,6.5,by=0.2)
#vmedia=c(1,2,3)
vpotencia=c()
```

```
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=rexp(n,rate=1/vmedia[imedia])

    # Test de la T de Student
    t=t.test(x,mu=2,alternative="greater")
    nct[is]=t$p.value

  }
  sum(nct<0.05)/ns  # Proporción de rechazos (T-Student)
  vpotencia[imedia]=sum(nct<0.05)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia[imedia],"\n")
}
lines(vmedia,vpotencia,type="l",col="red")
text(5,0.6,expression(paste(TS)), col="red")
```

Figura 2.3 : distribución Exp(theta=1/2,2,3,5) frente a N((mu=1/2,2,3,5),1)

```
# Rejilla de valores para el eje X
x <- seq(-1, 10, 0.05)
par(mfrow = c(2, 2))

# theta=1/2
plot(x, dnorm(x, mean = 1/2, sd = 1), type = "l",
     ylim = c(0, 2),xlim=c(-1,4),ylab = "f(x)", lwd = 2, lty=1, col = "red")

lines(x, dexp(x,rate = 2), col = "blue", lty = 1, lwd = 2)
text(3,1,expression(paste(theta==1/2)),cex=1)

# theta=2
plot(x, dnorm(x, mean = 2, sd = 1), type = "l",
     ylim = c(0, 1),xlim=c(-1,6),ylab = "f(x)", lwd = 2, lty=1, col = "red")

lines(x, dexp(x,rate = 1/2), col = "blue", lty = 1, lwd = 2)
```

```
text(5,0.4,expression(paste(theta==2)),cex=1)

# theta=3
plot(x, dnorm(x, mean = 3, sd = 1), type = "l",
     ylim = c(0, 1),xlim=c(-1,5), ylab = "f(x)", lwd = 2, lty=1, col = "red")

lines(x, dexp(x,rate = 1/3), col = "blue", lty = 1, lwd = 2)
text(4,0.8,expression(paste(theta==3)),cex=1)

# theta=5
plot(x, dnorm(x, mean = 5, sd = 1), type = "l",
     ylim = c(0, 1),xlim=c(-1,10), ylab = "f(x)", lwd = 2, lty=1, col = "red")

lines(x, dexp(x,rate = 1/5), col = "blue", lty = 1, lwd = 2)
text(8,0.8,expression(paste(theta==5)),cex=1)
```

Figura 2.4 : Función de potencia del test de Neymann-Pearson y del test T (simulada) para cuando la muestra sigue una distribución $Unif(0,\theta)$)

```
#Función de potencia para el test de Neymann-Pearson

alpha=0.05
 n=5

theta_0=1
theta_1=1.5

f1 <- function(x) alpha+0*x
f3 <- function(x) 1-((1-alpha)/x^n)

curve(expr=f3, from=theta_0, to=3,type="l",
      col='red',lwd=2 ,xlab = expression(theta),ylab="",
      xlim=c(0, 2.5),ylim=c(0,1.3),xaxt = "n",yaxt = "n")

curve(expr=f1, from=0, to=theta_0, type="l",
```

```
      col="red", lwd=2,add=TRUE)

grid()      # Añade guías en el fondo

segments(-1, 1, 3, 1, lty="dashed")
segments(-1, 1-((1-alpha)/theta_1^n), theta_1,
            1-((1-alpha)/theta_1^n) , lty="dashed")
segments(theta_1, -1, theta_1, 1-((1-alpha)/theta_1^n) ,
            lty="dashed")
text(2,1.1,(expression(paste(beta==1-frac(1-alpha,theta^n)))),
            cex=0.8, col="red")
text(x=1, y=1.1, expression(paste(beta==1)))
text(x=1/2, y=0.1, expression(paste(beta==alpha)),col="red")

axis(1,at=theta_1,labels=c(expression(paste(theta==3/2))))
axis(1,at=theta_0,labels=c(expression(paste(theta==1))))
axis(1,at=0,labels=c(expression(paste(theta==0))))
axis(2,at=0,labels=c(0),cex.axis=0.8)
axis(2,at=1-((1-alpha)/theta_1^n),
        labels=c(expression(paste(beta==1-frac(1-alpha,(3/2)^n)))),
        cex.axis=0.8)


# función de potencia para el test de la T de student

set.seed(123456)

# H_0: mu<=1/2
# H_a: mu>1/2
n=5

ns=10000
theta_0=1

vmedia=seq(0,3,by=0.05)
vpotencia=c()
```

```
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=runif(n,0,vmedia[imedia])     # datos normales

    # Test de la T de Student
    t=t.test(x,mu=theta_0/2,alternative="greater")
    # la media de la uniforme es (theta_0=1)/2

    nct[is]=t$p.value

  }
  sum(nct<0.05)/ns  # Proporción de rechazos (T-Student)
  vpotencia[imedia]=sum(nct<0.05)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia[imedia],"\n")
}
lines(vmedia,vpotencia,type="l",col="green")
#abline(h=1,lty=2)
text(1.8,0.5,expression(paste(TS)), col="green")
```

Figura 2.5: Exactamente lo mismo que en la figura anterior sólo que cambiando el tamaño muestral por $n = 15$ y $n = 50$.

Capítulo 3

Figura 3.1 : Función de densidad para la distribución de Laplace para $\mu = 0$ y distintos valores $\lambda = 1/2, 1, 2$

```
library(LaplacesDemon)
# DENSIDAD DE LAPLACE

x <- seq(from=-5, to=5, by=0.1)
plot(x, dlaplace(x,0,0.5), ylim=c(0,1), type="l",
        main="Función de probabilidad",
        ylab="densidad", col="red")
lines(x, dlaplace(x,0,1), type="l", col="green")
```

```
lines(x, dlaplace(x,0,2), type="l", col="blue")
legend(2, 0.9, expression(paste(mu==0, ", ", lambda==0.5),
        paste(mu==0, ", ", lambda==1), paste(mu==0, ", ", lambda==2)),
        lty=c(1,1,1),bty="n", col=c("red","green","blue"),cex=1.2)
```

Figura 3.2 : Funciones de potencia simuladas para el test de la T de student y para el test de la mediana cuando $X \in Laplace(0,1)$

```
set.seed(123456)

# H_0: mu<=0
# H_a: mu>0

n=30
ns=10000
mu_0=0
alfa=0.05

vmedia=seq(-0.5,1.5,by=0.05)
vpotencia_t=c()
vpotencia_m=c()
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=rlaplace(n,vmedia[imedia],1)

    # Test de la T de Student
    t=t.test(x,mu=mu_0,alternative="greater")
    # la media de la uniforme es (theta_0=1)/2

    nct[is]=t$p.value
    # Test de la mediana
    est=sum(x>0)
    ncm[is]=1-pbinom(est-1,n,prob=0.5)
  }
```

```
  sum(nct<alfa)/ns  # Proporción de rechazos (T-Student)
  vpotencia_t[imedia]=sum(nct<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_t[imedia],"\n")

  sum(nct<alfa)/ns  # Proporción de rechazos (Mediana)
  vpotencia_m[imedia]=sum(ncm<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_m[imedia],"\n")
}
plot(vmedia,vpotencia_t,type="l",col="green",
        ylab = "Potencia",xlab=expression(paste(mu)))
abline(h=1,lty=2)
abline(h=0,lty=2)
abline(h=alfa,lty=2,col="blue")
text(1,0.1,expression(paste(alpha==0.05)),cex=1.5,col="blue")
text(0.7,0.6,expression(paste(TS)), col="green",cex=1.8)

lines(vmedia,vpotencia_m,type="l",col="red")
text(0.2,0.6,expression(paste(TM)), col="red",cex=1.8)
```

Figura 3.3 : Funciones de potencia (simuladas) para el test de la T de student y para el test de la mediana cuando $X \in N(\mu, 1)$. Tamaños muestrales $n = 30$ y $n = 100$.

```
set.seed(123456)
par(mfrow = c(1, 2))
# H_0: mu<=0
# H_a: mu>0
n=30        # ó n=100
ns=10000
mu_0=0
alfa=0.05

vmedia=seq(-0.5,1.5,by=0.05)
vpotencia_t=c()
vpotencia_m=c()
for (imedia in 1:length(vmedia)){
  nct=c()
```

```
  ncm=c()
  for (is in 1:ns){
    x=rnorm(n,mean=vmedia[imedia])

    # Test de la T de Student
    t=t.test(x,mu=mu_0,alternative="greater")
    nct[is]=t$p.value

    # Test de la mediana
    est=sum(x>0)
    ncm[is]=1-pbinom(est-1,n,prob=0.5)
  }
  sum(nct<alfa)/ns  # Proporción de rechazos (T-Student)
  vpotencia_t[imedia]=sum(nct<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_t[imedia],"\n")

  sum(nct<alfa)/ns  # Proporción de rechazos (Mediana)
  vpotencia_m[imedia]=sum(ncm<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_m[imedia],"\n")
}
plot(vmedia,vpotencia_t,type="l",col="green",
        ylab = "Potencia",xlab=expression(paste(mu)))
abline(h=1,lty=2)
abline(h=0,lty=2)
abline(h=alfa,lty=2,col="blue")
text(1,0.1,expression(paste(alpha==0.05)),cex=1.5,col="blue")
text(-0.2,0.8,expression(paste(TS)), col="green",cex=1.8)

lines(vmedia,vpotencia_m,type="l",col="red")

text(0.55,0.6,expression(paste(TM)), col="red",cex=1.8)
```

Figura 3.4 : Funciones de potencia simuladas para el test de la T de student, test de la mediana y función de potencia para el test RSW (en negro) calculada mediante la fórmula 3.24 cuando $X \in Laplace(\mu = 0, \lambda = 1)$. Tamaños muestrales $n = 30$ y $n = 100$.

```
library(LaplacesDemon)

# función de potencia para el test de RSW :

n=100
theta=seq(-0.25,0.6,length=200) # hasta 1 en n=30
alpha=0.05
zeta_alpha=qnorm(1-alpha)
#en la normal estándar deja probabilidad de alpha a la derecha

integrand <- function(x) {(1/2)*exp(-2*x)}    # laplace(mu=0,lambda=1)

# multiplicamos por 2 por ser par respecto de 0.
int=integrate(integrand,lower=0,upper = Inf)
integral=int$value
# utilizamos $value para que nos de el valor estimado,
# sino devuelve un valor no numérico.

denominador=sqrt((n*(n+1)*(2*n+1))/(24))
potencia=1-pnorm(zeta_alpha-(n*(n-1)*theta*integral
                    +n*theta*dlaplace(0,0,1))/denominador)

plot(theta,potencia,ylim=c(0,1.25),type="l",lwd=2,xlab="m",
        ylab=expression(paste("Potencia"==beta)))

text(0,0.82,expression(paste(RSW)),cex=1.2)
abline(h=0.05,lty=2,col="blue")
abline(h=1,lty=2,col="blue")
text(0.5,0.1,expression(paste(alpha==0.05)),cex=1,col="blue")
text(-0.1,1.1,expression(paste(beta==1)),cex=1,col="blue")
```

```
# Función de potencia para el test de la T de student y el test
# de la mediana en el caso en que los datos siguieran una
# Laplace(mu=0,scale=1)

set.seed(123456)
# H_0: mu<=0
# H_a: mu>0
n=100

ns=20000
mu_0=0
alfa=0.05

vmedia=seq(-0.25,0.6,by=0.05)
vpotencia_t=c()
vpotencia_m=c()
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=rlaplace(n,vmedia[imedia],1)
    # Test de la T de Student
    t=t.test(x,mu=mu_0,alternative="greater")
    nct[is]=t$p.value
    # Test de la mediana
    est=sum(x>0)
    ncm[is]=1-pbinom(est-1,n,prob=0.5)
  }
  sum(nct<alfa)/ns  # Proporción de rechazos (T-Student)
  vpotencia_t[imedia]=sum(nct<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_t[imedia],"\n")

  sum(nct<alfa)/ns  # Proporción de rechazos (Mediana)
  vpotencia_m[imedia]=sum(ncm<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_m[imedia],"\n")
}
```

```
lines(vmedia,vpotencia_t,type="l",col="green",lwd=2)
text(0.4,0.35,expression(paste(TS)), col="green",cex=1.2)

lines(vmedia,vpotencia_m,type="l",lwd=2,col="red")
text(0,0.4,expression(paste(TM)), col="red",cex=1.2)

### Análogo para el caso en el que n=30 ###
```

Figura 3.5 : Funciones de potencia (simuladas) para el test de la T de student (TS, en verde), test de la mediana (TM, en rojo) y test RSW (aproximada mediante la fórmula (3.26)) (en amarillo) cuando $X \in N(\mu = 0, \sigma^2 = 1)$. Tamaño muestral $n = 30$ y $n = 100$.

```
### Hacemos lo mismo que en caso anterior pero para cuando los datos
provienen de una N(0,1) ###

## Función de potencia del test RSW por la fórmula (caso normal):

n=100       # ó n=30
theta=seq(-0.25,1,length=200)
alpha=0.05
zeta_alpha=qnorm(1-alpha)
integrand <- function(x) {dnorm(x)^2}
int=integrate(integrand,lower=0,upper = Inf)

integral=int$value
denominador=sqrt((n*(n+1)*(2*n+1))/(24))
potencia=1-pnorm(zeta_alpha-(n*(n-1)*theta*integral
                +n*theta*dnorm(0))/denominador)


potencia_2=1-pnorm(zeta_alpha-sqrt(12)*integral*theta*sqrt(n))


plot(theta,potencia,ylim=c(0,1.25),type="l",lwd=2,xlab="m",
        ylab=expression(paste("Potencia"==beta)))
lines(theta,potencia_2,type="l",col="yellow",lwd=2)
```

```
text(0.6,0.4,expression(paste(RSW)),cex=1.2,col="yellow")

abline(h=0.05,lty=2,col="blue")
abline(h=1,lty=2,col="blue")

text(0.5,0.1,expression(paste(alpha==0.05)),cex=1,col="blue")
text(-0.1,1.1,expression(paste(beta==1)),cex=1,col="blue")

# función de potencia para el test de la T de student y el
# test de la mediana en el caso en que los datos siguieran
# una laplace(mu=0,scale=1)

set.seed(123456)

# H_0: mu<=0
# H_a: mu>0

#n=100
ns=20000
mu_0=0
alfa=0.05

vmedia=seq(-0.25,1,by=0.05)   #hasta 0.6 en n=100 y hasta 1 en n=30
vpotencia_t=c()
vpotencia_m=c()
for (imedia in 1:length(vmedia)){
  nct=c()
  ncm=c()
  for (is in 1:ns){
    x=rnorm(n,mean=vmedia[imedia])
    # Test de la T de Student
    t=t.test(x,mu=mu_0,alternative="greater")
    nct[is]=t$p.value
    # Test de la mediana
    est=sum(x>0)
```

```
    ncm[is]=1-pbinom(est-1,n,prob=0.5)
  }
  sum(nct<alfa)/ns  # Proporción de rechazos (T-Student)
  vpotencia_t[imedia]=sum(nct<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_t[imedia],"\n")

  sum(nct<alfa)/ns  # Proporción de rechazos (Mediana)
  vpotencia_m[imedia]=sum(ncm<alfa)/ns
  cat('vmedia',vmedia[imedia],'potencia',vpotencia_m[imedia],"\n")
}
lines(vmedia,vpotencia_t,type="l",col="green",lwd=2)
text(0,0.6,expression(paste(TS)), col="green",cex=1.2)

lines(vmedia,vpotencia_m,type="l",lwd=2,col="red")
text(0.6,1.1,expression(paste(TM)), col="red",cex=1.2)
```

Capítulo 4

Figura 4.1 : Función de potencia aproximada por la fórmula (4.19) para el test WMW cuando $X \in N(\mu = 0, \sigma^2 = 1)$ y $n = 30$ (muestras emparejadas).

```
n=30
m=30

theta=seq(-0.25,1,length=200)
alpha=0.05
zeta_alpha=qnorm(1-alpha) # en la normal estándar deja probabilidad de
                          # alpha a la derecha
integrand <- function(x) {dnorm(x)^2}

int=integrate(integrand,lower=-Inf,upper = Inf)
integral=int$value

potencia=pnorm(-zeta_alpha+sqrt((12*m*n)/(m+n))*integral*theta)

plot(theta,potencia,ylim=c(0,1.25),type="l",lwd=2,
```

```
    xlab=expression(paste(theta)),ylab=expression(paste("Potencia"==beta)))
text(0.8,0.65,expression(paste(WMW)),cex=1.2)

abline(h=0.05,lty=2,col="blue")
abline(h=1,lty=2,col="blue")

text(0.5,0.1,expression(paste(alpha==0.05)),cex=1,col="blue")
text(-0.1,1.05,expression(paste(beta==1)),cex=1,col="blue")
```

Bibliografía

[1] Canavos, G.C. (1988). *Probabilidad y Estadística. Aplicaciones y métodos*. McGraw-Hill/Interamericana de Mexico, S.A. de C.V..

[2] Cristóbal, J.A.C. (2003). *Lecciones de inferencia estadística*. Universidad de Zaragoza, Tercera edición.

[3] Dudewicz, E.J. & Mishra, S.N. (1988). *Modern mathemathical statistics*. John Wiley & Sons, Inc..

[4] Freijeiro González, L., González Manteiga, W. & Sánchez Sellero, C.A., *Apuntes de la asignatura de Probabilidad y estadística*. Grado en matemáticas, USC, curso 2019/2020.

[5] Hettmansperger, T.P. (1984). *Statistical Inference based on Ranks*. John Wiley & Sons.

[6] Hodges, J.L. (1955). *The Annals of Mathemathical Statistics*. Institute of Mathematical Statistics, Vol. 26, No. 3, 523-527.

[7] Lehman, E.L., (1975). *Nonparametric statistical methods based on ranks*. San Francisco: Holden-Day.

[8] Rohatgi, V.K. (1976). *Probability theory and mathematical statistics*. John Wiley & Sons, Inc..

[9] Sánchez Sellero, C.A., *Apuntes de la asignatura Estadística avanzada para la Optometría, Contrastes no paramétricos*. Máster en Optometría, USC, curso 2020/2021.

[10] Vergara Morales, M. (2007). *Impacto de la asimetría de la distribución muestreada en la potencia asintótica de la prueba del rango signado de Wilcoxon* [Tesis de maestría].