USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIVERSIDADE DA CORUÑA

Universidade Vigo

Trabajo Fin de Máster

Tests de equivalencia para comparar medias

Javier Rey Ramírez

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Tests de equivalencia para comparar medias

Título en español: Tests de equivalencia para comparar medias

English title: Equivalence tests to compare means

Modalidad: Modalidad A

Autor: Javier Rey Ramírez, Universidade de Santiago de Compostela

Director: Juan Carlos Pardo Fernández, Universidade de Vigo

Breve resumen del trabajo:

En la teoría clásica, la comparación de las medias de dos poblaciones se realiza mediante tests en los cuales la hipótesis nula establece que las medias de ambas poblaciones son iguales frente a la hipótesis alternativa de que las medias son distintas. Estos procedimientosson adecuados cuando se desea comprobar la existencia de alguna diferencia entre las medias comparadas. Sin embargo, resultan poco informativos cuando lo que se desea comprobar es la igualdad entre las cantidades comparadas. Los contrastes de equivalencia se ocupan de este problema formulando una hipótesis alternativa en la cual se establece que las medias distan menos de una cierta cantidad prefijada.

En este trabajo revisaremos la literatura sobre los contrates de equivalencia y se comprobará su funcionamiento práctico mediante estudios de simulación en R y aplicaciones. Además, se explorará la posibilidad de trabajar con tests de equivalencia para funciones de regresión.

Recomendaciones:

Otras observaciones:

Don Juan Carlos Pardo Fernández, profesor titular de la Universidade de Vigo, informa que el Trabajo Fin de Máster titulado

Tests de equivalencia
para comparar medias

fue realizado bajo su dirección por don Javier Rey Ramírez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En Vigo, a 10 de Febrero de 2020.

El director:

Don Juan Carlos Pardo Fernández

El autor:

Don Javier Rey Ramírez

Agradecimientos

A nivel académico me gustaría agradecer en primer lugar a mi tutor Juan Carlos por toda su ayuda, al grupo SiDOR por el excelente ambiente de trabajo y a

- Grupo de Referencia Competitiva de la Xunta de Galicia - Programa de Consolidación - GRC ED431C 2016/040
- FEDER. Unha maneira de facer Europa. Promover o desenvolvemento tecnolóxico, a innovación e unha investigación de calidade. Centro Singular de Investigación de Galicia 2016-2019, CINBIO - ED431G/02
- Ministerio de Economía, Industria y Competitividad - Proyectos I+D - MTM2017-089422-P

por la financiación.

A nivel personal me gustaría agradecer el apoyo a Iria, Víctor, Iago y a mi familia.

Índice general

Resumen

Resumen en español

En la teoría clásica, la comparación de las medias de dos poblaciones se realiza mediante un contraste en el cual la hipótesis nula establece que las medias de ambas poblaciones son iguales frente a la hipótesis alternativa de que son distintas. Sin embargo, cuando el interés práctico es verificar la igualdad de las medias, este tipo de contrastes presentan un inconveniente, ya que cuando el tamaño muestral es elevado la potencia aumenta y el contraste es capaz de detectar diferencias muy pequeñas entre las medias que en la práctica pueden resultar no relevantes. Surgen entonces los tests de equivalencia, en los cuales la hipótesis nula establece que las medias distan más de una cierta cantidad pre-especificada frente a la alternativa de que la diferencia entre las medias es menor que esa cantidad.

En la primera parte (capítulo 2) de este trabajo se revisa la literatura más relevante sobre los tests de equivalencia para la media de dos poblaciones y se comprobará el funcionamiento práctico de estos tests mediante estudios de simulación en R y aplicaciones. Además, se propondrá un nuevo test de equivalencia no paramétrico basado en bootstrap. En la segunda parte (capítulo 3) se estudiará la literatura relativa a los tests de equivalencia para los modelos de regresión.

English abstract

In classical theory, the comparison of the means of two populations is carried out by tests in which the null hypothesis establishes that the means of both populations are equal against the alternative hypothesis that they differ. However, when the practical interest is to verify the equality of the means, this type of tests presents an inconvenience, since when the sample size is high the power increases and the test is able to detect very small differences between the means that in practice may not be relevant. Equivalence tests deal with this problem, in which the null hypothesis establishes that the distance between the means is greater than a certain prefixed value, whereas the alternative hypothesis states that the distance is less than that value.

In the first part (chapter 2) of this dissertation we revise relevant literature related to equivalence tests for the means of two populations. The practical performance of these tests will be investigated by means of simulations in R and applications. Moreover, we will proposed a new non-parametric equivalence test based on bootstrap. In the second part (chapter 3) we will review the literature regarding equivalence tests for regression models.

Capítulo 1

Introducción

1.1. Motivación del problema

La media, junto con la varianza, es una de las medidas más utilizadas para resumir la información de una población, por ello, cuando se desee comprobar si una característica que se puede cuantificar, como podría ser la edad, difiere en dos grupos, como por ejemplo el género, estaremos hablando de comparación de medias. En la teoría clásica esto se realiza mediante un contraste que establece en la hipótesis nula que las medias de ambas poblaciones coinciden, frente a la hipótesis alternativa de que son distintas. Es decir, se realiza el contraste

$$H_0 : \mu_X - \mu_Y = 0$$

frente a la alternativa

$$H_1 : \mu_X - \mu_Y \neq 0,$$

donde μ_X y μ_Y representan las medias de las poblaciones X e Y, respectivamente. Para resolver este tipo de contrastes el procedimiento más habitual en la práctica es el t-test. Sin embargo, cuando el interés práctico sea verificar la igualdad de medias, con este tipo de contrastes no podremos afirmar que existe una fuerte evidencia de que la media de ambas poblaciones coincida. Además, cuando el tamaño muestral sea elevado, la potencia aumenta y el contraste es capaz de detectar diferencias muy pequeñas entre las medias que en la práctica pueden resultar no relevantes. Ilustraremos esto con la relación entre el tamaño muestral, la diferencia real entre las medias y la potencia del t-test de dos poblaciones normales con $\sigma^2 = 1$. Llamaremos $d = \mu_X - \mu_Y$ a la diferencia real entre las medias.

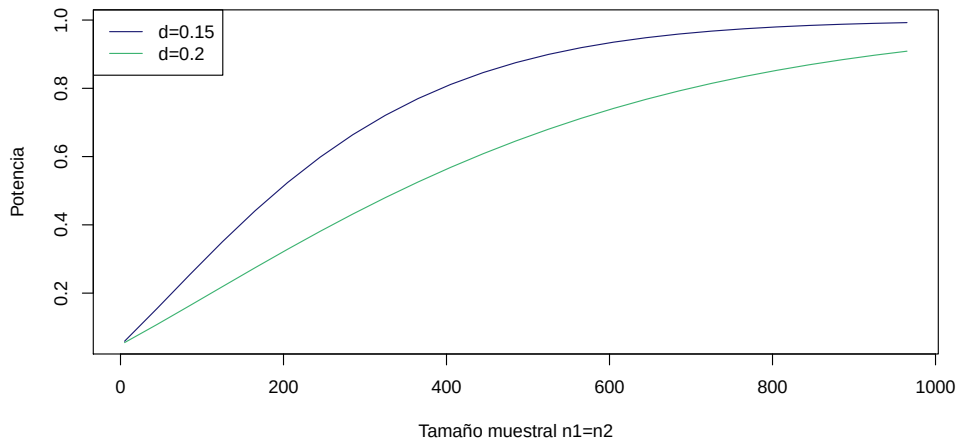

Figura 1.1: Relación entre tamaño muestral y potencia del t-test cuando las poblaciones siguen distribuciones $N(0, 1)$ y $N(d, 1)$ independientes

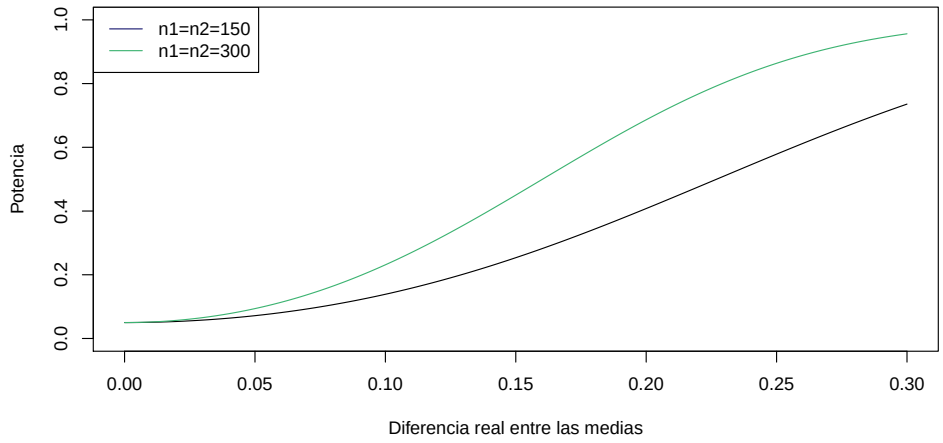

Figura 1.2: Relación entre la diferencia real entre medias y la potencia del t-test cuando las poblaciones siguen distribuciones $N(0,1)$ y $N(d,1)$ independientes

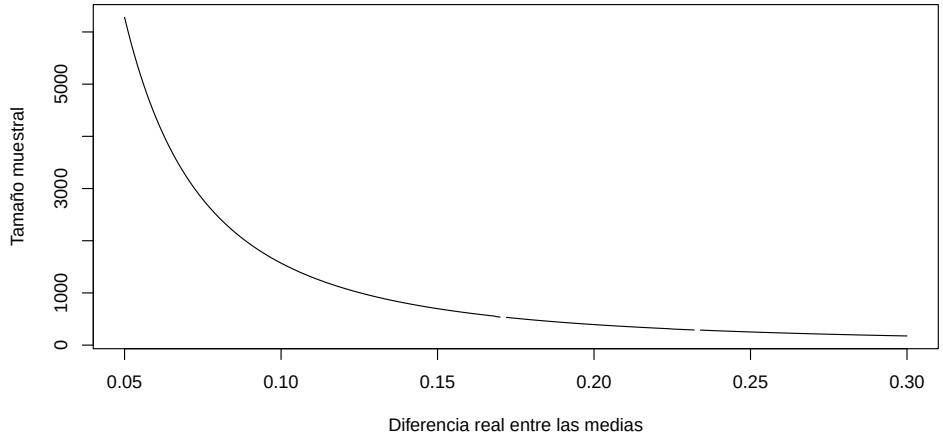

Figura 1.3: Relación entre la diferencia real entre medias y el tamaño muestral para que el t-test alcance una potencia de 0.8 cuando las poblaciones siguen distribuciones $N(0,1)$ y $N(d,1)$ independientes

En las figuras 1.1 y 1.2 podemos ver la relación entre la potencia y el tamaño muestral o la diferencia real entre las medias, respectivamente, cuando las poblaciones siguen distribuciones $N(0,1)$ y $N(d,1)$ independientes. Observamos que conforme aumentan el tamaño muestral o la diferencia real entre las medias, la potencia aumenta. Por tanto, vemos que para diferencias entre medias que pueden considerarse irrelevantes en la práctica el test clásico t-test puede alcanzar una gran potencia.

En la figura 1.3 muestra la relación entre el tamaño muestral y la diferencia real entre las medias fijada una potencia de 0.8 cuando las poblaciones siguen distribuciones $N(0,1)$ y $N(d,1)$ independientes. Una vez más observamos que conforme aumenta la diferencia entre la media de las poblaciones el tamaño muestral necesario para detectar dicha diferencia decrece.

Surgen entonces los tests de equivalencia introducidos por Westlake (1976), en los cuales la hipótesis nula establece que las medias distan más de una cierta cantidad pre-especificada frente a la alternativa de que la diferencia entre las medias es menor que esa cantidad. Más concretamente, sean X e Y dos variables aleatorias con medias μ_X y μ_Y, respectivamente, y sea $d = \mu_X - \mu_Y$ la diferencia entre las medias. Dadas

dos cantidades prefijadas d_L y d_U cumpliendo que $d_L < d_U$, el test de equivalencia contrasta la hipótesis nula

$$H_0 : d \notin [d_L, d_U],$$

frente a la alternativa

$$H_1 : d \in (d_L, d_U).$$

Con este planteamiento, el rechazo de la hipótesis nula permitiría concluir que la diferencia de medias se encuentra en el intervalo prefijado (d_L, d_U). De gran interés práctico resulta el caso $d_U = -d_L = \Delta$, donde $\Delta > 0$ es una cantidad pequeña, ya que un resultado significativo del test permitiría justificar que las medias distan menos de Δ. En este caso, podemos decir que las medias son equivalentes.

1.1.1. Tipos de intervalos de equivalencia

Un ámbito en el que el uso de los tests de equivalencia es común es el de la farmacología. Es habitual tratar de conocer si se puede remplazar un fármaco cuya eficacia ya se ha verificado por uno experimental. En este caso, llamaremos μ_R y μ_E a las medias de las poblaciones de referencia y experimentales, respectivamente. En este y otros ámbitos surgen diferentes maneras de establecer reglas para la elección del intervalo de equivalencia. Algunas de estas reglas han sido expuestas por organismos internacionales como la Food & Drug Administration (FDA) estadounidense.

Intervalos de equivalencia basados en la media

Algunas de las reglas más populares para establecer el intervalo de equivalencia es el uso de las medias de las poblaciones. Podemos dividir en dos clases este tipo intervalos.

- La primera clase de intervalos de equivalencia basados en la media viene dado por

  $$(-\phi\mu_R, \phi\mu_R),$$

  donde ϕ es un número pequeño. El número ϕ suele tomarse como 0.15 (Luzar-Stiffler y Stiffler, 2002) o 0.2 (Ocaña Rebull et al., 2008). El problema de este intervalo de equivalencia es que la media no es una medida de dispersión y si por ejemplo, la población de referencia siguiese una distribución de media 0, entonces, no existiría intervalo de equivalencia.

- La segunda clase se fundamenta en que si las medias de ambas poblaciones coinciden, entonces $\frac{\mu_E}{\mu_R} = 1$. Por tanto podemos estableces que ambas poblaciones son equivalentes si $R_L < \frac{\mu_E}{\mu_R} < R_U$. Ahora, aplicando una transformación logarítmica tenemos que el intervalo de equivalencia viene dado por

  $$(\log R_L < \log \mu_E - \log \mu_R < \log R_U).$$

  En este caso la FDA sugiere escoger los valores $R_L = 0.8$ y $R_U = 1.25$ (Center for Drug Evaluation and Research, 2001). El problema de este intervalo de equivalencia es que para aplicar los tests desarrollados suponiendo que las poblaciones siguen una distribución normal, necesitaríamos que las poblaciones siguieran una distribución lognormal. Además, no siempre es posible aplicar una transformación logarítmica.

Intervalo de equivalencia basado en la desviación típica

Otra de las posibles reglas para establecer el intervalo de equivalencia es tomar un porcentaje de la desviación de típica de la población de referencia. En este caso tenemos que el intervalo de equivalencia viene dado por

$$(-\phi\sigma_R, \phi\sigma_R),$$

donde ϕ es una proporción pequeño y σ_R es la desviación típica de la población de referencia. Food and Drug Administration (2016) sugiere tomar como $\phi = 1.5$ ya que de esta manera, cuando el tamaño muestral es de $n_1 = n_2 = 10$, se obtiene una potencia de al menos 0.8 cuando la diferencia real entre las medias está contenida en el intervalo $(-\frac{\sigma_R}{8}, \frac{\sigma_R}{8})$. Sin embargo, utilizando este valor de ϕ tenemos que la potencia también es muy elevada cuando la diferencia real entre las medias se sitúe fuera del intervalo $(-\frac{\sigma_R}{8}, \frac{\sigma_R}{8})$.

Intervalo de equivalencia basado en la distancia entre cuantiles

Hemos visto como es preferible el uso de una medida de dispersión para establecer el intervalo de equivalencia. Por tanto se propone usar como intervalo de equivalencia

$$(-0.1[Q_{95} - Q_5], 0.1[Q_{95} - Q_5])$$

donde Q_5 y Q_{95} son los cuantiles 5 % y 95 % respectivamente de la población de referencia. Es decir, tomaremos como intervalo de equivalencia el 10 % de la longitud del intervalo donde se encuentran el 90 % de la densidad. En el caso de que la población de referencia siga una distribución $N(0, 1)$, el intervalo de equivalencia sería $(-0.329, 0.329)$. En general si la población de referencia es $N(\mu_R, \sigma_R)$ entonces en intervalo de equivalencia resultante es $(-0.329\sigma_R, 0.329\sigma_R)$.

1.2. Teoría clásica

Para resolver el contraste sobre la igualdad de medias de dos poblaciones independientes hay dos alternativas muy utilizadas en la práctica. La primera conocida como el t-test está basada en la normalidad de las poblaciones y en la distribución t de Student. La segunda es un test no paramétrico conocido como el test de Wilcoxon-Mann-Whitney.

1.2.1. t-test

Uno de los tests más usados para la comparación de dos medias es el t-test introducido por Student (1908). Este busca contrastar la hipótesis nula:

$$H_0 : \mu_X - \mu_Y = 0,$$

frente a la alternativa

$$H_1 : \mu_X - \mu_Y \neq 0,$$

siendo X e Y dos poblaciones independientes que siguen una distribución normal de medias μ_X y μ_Y, respectivamente, y de igual varianza σ^2. Entonces, siendo $X_1 \ldots, X_{n_1}$ e $Y_1 \ldots, Y_{n_2}$ dos muestras aleatorias simples se utiliza el estadístico

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}}},$$

donde $\bar{X}$ y $\bar{Y}$ son las medias muestrales y s_X^2 y s_Y^2 son las cuasi-varianzas muestrales. El estadístico T sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad.

Cuando no podamos suponer que las varianzas de ambas poblaciones coinciden, Welch (1947) propuso utilizar

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

el cual sigue una distribución t de Student con f grados de libertad, siendo

$$f = \frac{\left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right)^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2-1}}.$$

1.2.2. Test de Wilcoxon-Mann-Whitney

Cuando no sea posible suponer que las poblaciones siguen una distribución normal se puede usar un contraste no paramétrico sobre la función de distribución para decidir si hay evidencias suficientes de que las localizaciones de las poblaciones son diferentes. Por tanto, siendo $X_1 \ldots, X_{n_1}$ e $Y_1 \ldots, Y_{n_2}$ dos

muestras aleatorias simples independientes de $X \sim F$ e $Y \sim G$, respectivamente, siendo F y G funciones de distribución continuas, se desea contrastar

$$H_0 : F(x) = G(x), \text{ para todo } x, \tag{1.1}$$

frente a

$$H_1 : \mathbb{P}(Y \geq X) > 0.5.$$

Para realizar el contraste (1.1) Wilcoxon (1945) y Mann y Whitney (1947) propusieron de manera independiente dos test equivalentes. Ambos se usan en la práctica como alternativa al t-test cuando no se pueda suponer que las poblaciones tienen distribución normal. En el capítulo 4 de Hollander et al. (1999) se puede ver el porqué del uso de este contraste sobre las funciones de distribución para decidir si existen diferencias significativas entre las medias de ambas poblaciones.

Estadístico de Mann-Whitney

Mann y Whitney (1947) propusieron para realizar el contraste (1.1) el siguiente estadístico,

$$D_{MW} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{(0,\infty)}(Y_j - X_i), \tag{1.2}$$

donde

$$I_A(x) = \begin{cases} 1 & si \quad x \in A, \\ 0 & si \quad x \notin A. \end{cases}$$

La distribución de (1.2) puede calcularse de manera exacta bajo la hipótesis nula. Además, asintóticamente se tiene que, bajo la hipótesis nula

$$\frac{D_{MW} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0, 1).$$

Estadístico de Wilcoxon

Wilcoxon (1945) propuso el siguiente estadístico,

$$D_W = \sum_{i=1}^{n_1} rango(Y_i), \tag{1.3}$$

donde $rango(Y_j)$ se define como

$$rango(Y_j) = \#\{X_i : X_i \leq Y_j\} + \#\{Y_l : Y_l \leq Y_j\}.$$

Los estadísticos (1.2) y (1.3) son equivalentes, ya que

$$D_W = D_{MW} + \frac{n_1(n_1 + 1)}{2}.$$

Capítulo 2

Tests de equivalencia para la media de dos poblaciones

En el capítulo 1 hemos visto como podemos realizar un contraste que nos permita decidir si la media de dos poblaciones es significativamente diferente, sin embargo, con este tipo de contrastes no podremos afirmar que hay pruebas significativas de que las medias de las poblaciones coinciden. Además, en áreas como la farmacología cuando el tamaño muestral sea elevado los tests detectarán diferencias irrelevantes en la práctica. Por ello, surge la necesidad de presentar los llamados tests de equivalencia que establecen como hipótesis alternativa la equivalencia entre las medias.

2.1. Tests basados en la normalidad de los datos

En esta sección vamos a presentar los tests de equivalencia basados en la normalidad de los datos más usados en la práctica. Estos son el *two one-sided test* (TOST), el test de Anderson-Hauck y el test de Wellek.

Sean $X_1 \ldots, X_{n_1}$ e $Y_1 \ldots, Y_{n_2}$ dos muestras aleatorias simples independientes procedentes de dos distribuciones normales de medias μ_X y μ_Y, respectivamente, y de igual varianza σ, es decir,

$$\begin{aligned} X &\sim N(\mu_X, \sigma) \\ Y &\sim N(\mu_Y, \sigma) \end{aligned} \tag{2.1}$$

Sea $d = \mu_X - \mu_Y$ la diferencia de las medias de ambas poblaciones y d_L y d_U dos cantidades prefijadas de antemano cumpliendo que $d_L < d_U$. Se quiere contrastar la hipótesis nula

$$H_0 : d \notin [d_L, d_U] \tag{2.2}$$

frente a la alternativa

$$H_1 : d \in (d_L, d_U).$$

Es decir, se quiere contrastar conjuntamente la hipótesis nulas de que la diferencia de las medias es menor que la cantidad d_L y mayor que d_U frente a la alternativa de que la diferencia de medias esté entre ambas cantidades, por tanto, cuando se rechace la hipótesis nula habrá fuerte evidencia de que la diferencia entre la media de ambas poblaciones está contenida en un intervalo prefijado. Diremos entonces que ambas medias son equivalentes.

2.1.1. TOST

Uno de los tests de equivalencia para las medias de dos poblaciones normales más usados, conocido como TOST, fue introducido en el contexto de la bioequivalencia por Westlake (1981) y Schuirmann (1981), y ampliamente desarrollado por Schuirmann (1987).

Bajo las suposiciones (2.1) se rechazará la hipótesis nula de (2.2) si

$$T_a = \frac{\hat{d} - d_L}{se(\hat{d})} \geq t_{1-\alpha,\nu} \quad \text{y} \quad T_b = \frac{\hat{d} - d_U}{se(\hat{d})} \leq -t_{1-\alpha,\nu},$$

siendo $\hat{d} = \bar{X} - \bar{Y}$ la diferencia de las medias muestrales, $t_{1-\alpha,\nu}$ el cuantil $100(1-\alpha)$ de una distribución t de Student con $\nu = n_1 + n_2 - 2$ grados de libertad y

$$se(\hat{d}) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}},$$

donde s_X^2 y s_Y^2 son las cuasivarianzas muestrales de la primera y segunda población respectivamente. Ahora, siguiendo los pasos de Shen et al. (2015), dado $d \in (d_L, d_U)$, la potencia del test TOST se puede expresar como

$$\mathbb{P}(\text{rechazar } H_0 \text{ cuando } H_0 \text{ es falsa}) = \mathbb{P}\left(\frac{\hat{d} - d_L}{se(\hat{d})} \geq t_{1-\alpha,\nu} \quad y \quad \frac{\hat{d} - d_U}{se(\hat{d})} \leq -t_{1-\alpha,\nu}\right) =$$

$$= \int_0^R \left(\phi\left(\frac{d_U - d}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - t_{1-\alpha,\nu}\sqrt{\frac{x}{\nu}}\right) - \phi\left(\frac{d_L - d}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + t_{1-\alpha,\nu}\sqrt{\frac{x}{\nu}}\right)\right) \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} dx,$$

donde

$$R = \frac{(d_U - d_L)^2 \nu}{t_{1-\alpha,\nu}^2 2\sigma^2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1},$$

ϕ es la función de distribución de distribución normal estándar y $\Gamma(\cdot)$ es la función gamma. Por tanto, al aumentar la varianza de las poblaciones la potencia del TOST se reduce. Además, cuando aumenta la longitud del intervalo de equivalencia también aumenta la potencia.

En la figura 2.1 se muestran la curvas de potencia del test TOST para diferentes desviaciones típicas. La curva negra es la potencia del TOST cuando ambas poblaciones siguen una distribución $N(0, 0.3^2)$, la roja cuando siguen una $N(0, 0.4^2)$, la verde siguen una $N(0, 0.5^2)$ y la azul siguen una $N(0, 0.6^2)$. En la figura 2.1 se puede comprobar que para un tamaño muestral fijo de $n_1 = n_2 = 50$ al aumentar la varianza la potencia disminuye.

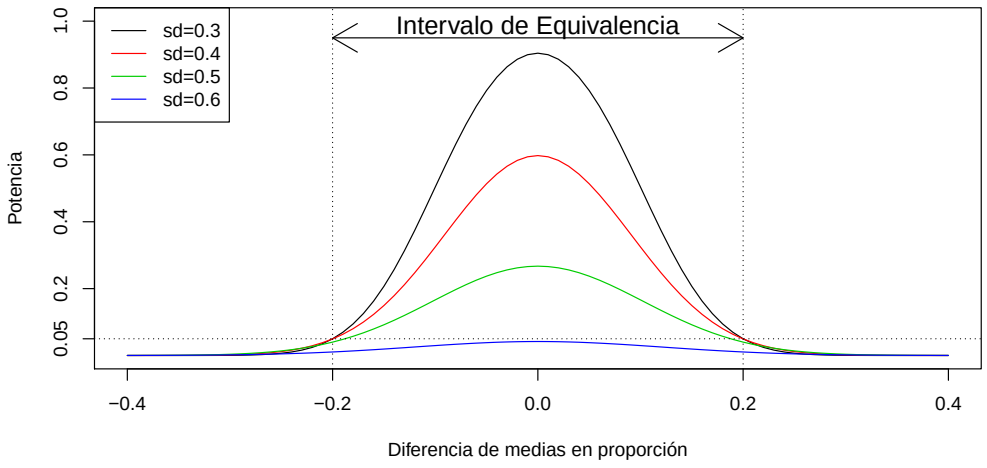

Figura 2.1: Curvas de potencia del test TOST para diferentes desviaciones típicas

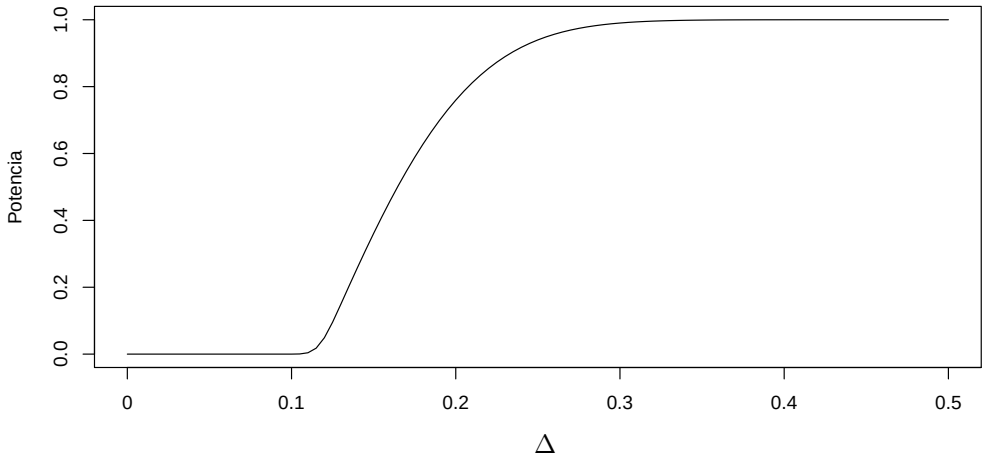

Figura 2.2: Relación entre Δ y la potencia del test TOST cuando $d = 0$ y las poblaciones siguen una distribución $N(0, 1)$ con tamaños muestrales $n_1 = n_2 = 100$

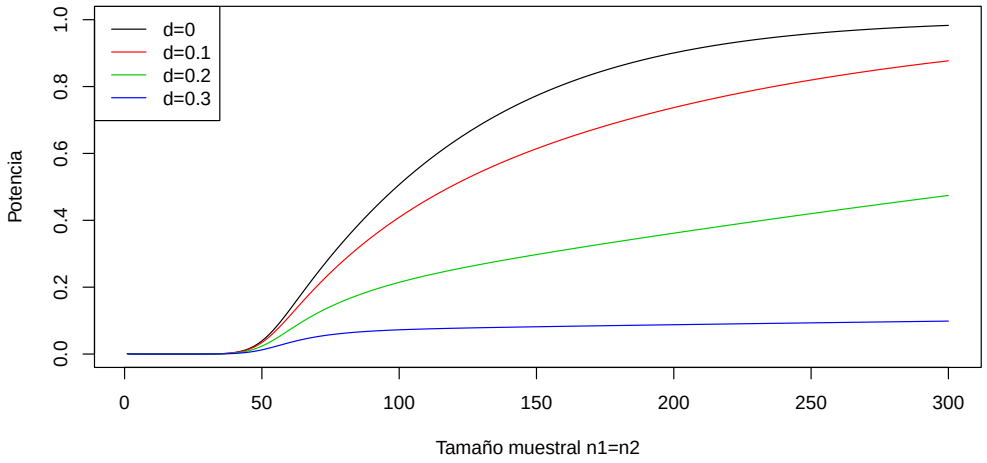

Figura 2.3: Relación entre la potencia y el tamaño muestral del test TOST cuando para diferentes valores de d cuando las poblaciones siguen una distribución $N(0, 1)$ con intervalo de equivalencia $(-0.329, 0.329)$

En la figura 2.2 se muestra la relación entre Δ, siendo $\Delta = -d_L = d_U$, y la potencia del test TOST cuando la diferencia real entre las medias es $d = 0$ y las poblaciones siguen una distribución $N(0, 1)$ con tamaños muestrales $n_1 = n_2 = 100$. En ella observamos que a medida que aumenta la longitud del intervalo de equivalencia la potencia también aumenta.

En la figura 2.3 se muestra la relación entre la potencia y el tamaño muestral del test TOST para diferentes valores de d cuando las poblaciones siguen una distribución $N(0, 1)$. Como intervalo de equivalencia hemos usado el propuesto en el capítulo 1, que, en el caso de una población $N(0, 1)$ da como resultado el intervalo $(-0.329, 0.329)$. En color negro se muestra la relación entre la potencia y el tamaño muestral cuando la diferencia real entre las medias es $d = 0$, en rojo cuando $d = 0.1$, en verde cuando $d = 0.2$ y en azul cuando $d = 0.3$. En la figura 2.3 observamos que a medida que aumenta el tamaño

muestral aumenta la potencia. Además, el test TOST para los diferentes valores de d comienza a tener potencia a partir de $n_1 = n_2 = 50$.

2.1.2. Test de Anderson-Hauck

Otro de los tests de equivalencia más importantes en la práctica es el de Anderson-Hauck introducido por Anderson y Hauck (1983).

Este, al contrario que el TOST expuesto en la sección 2.1.1, para realizar el contrate 2.2 bajo la suposición 2.1 utilizará un único estadístico

$$T_{AH} = \frac{\hat{d} - \frac{1}{2}(d_L + d_U)}{se(\hat{d})}. \tag{2.3}$$

El estadístico (2.3) sigue una distribución t de Student no central (vease sección A.1 del apéndice) con parámetro de no centralidad

$$\delta = \frac{d - \frac{1}{2}(d_L + d_U)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{2.4}$$

Se rechazará la hipótesis nula en favor de la alternativa cuando

$$C_1 < T_{AH} < C_2,$$

donde C_1 y C_2 satisfacen que

$$\mathbb{P}(C_1 < T_{AH} < C_2|_{d=d_L}) = \mathbb{P}(C_1 < T_{AH} < C_2|_{d=d_U}) = \alpha.$$

En el apéndice A de Anderson y Hauck (1983) se puede ver que C_1 y C_2 se pueden escoger tal que $C_2 = -C_1 = C$ y por tanto,

$$\mathbb{P}(|T_{AH}| < C|_{d=d_U}) = \alpha.$$

Por tanto la región de rechazo se consigue obteniendo el valor C. Si el parámetro de no centralidad es conocido Anderson y Hauck sugieren que es más informativo calcular un p-valor ρ, es decir, siendo t un valor observado de T_{AH} el p-valor, ρ que viene dado por

$$\rho = \mathbb{P}(|T_{AH}| < |t||_{d=d_U}).$$

En la práctica suponer que se conoce el parámetro de no centralidad δ no es muy realista, por tanto, cuando el parámetro de no centralidad δ es desconocido los autores plantean tres aproximaciones, una basada en una t de Student no central, otra basada e la distribución normal y otra basada en una t de Student. En el estudio de simulación de Anderson y Hauck (1983) se puede ver que la mejor aproximación en términos de potencia es la basada en una distribución t de Student, por tanto, esta es la que usaremos. Para esto, se necesita estimar el parámetro de no centralidad δ por

$$\hat{\delta} = \frac{d_U - \frac{1}{2}(d_L + d_U)}{se(\hat{d})} = \frac{d_U - d_L}{2se(\hat{d})}.$$

Ahora, podemos reescribir ρ como

$$\begin{aligned} \rho &= \mathbb{P}(|T_{AH}| < |t||_{d=d_U}) = \mathbb{P}(-|t| < T_{AH} < |t||_{d=d_U}) \\ &= \mathbb{P}(-|t| - \hat{\delta} < T_{AH} - \hat{\delta} < |t| - \hat{\delta}|_{d=d_U}) = F_t(|t| - \hat{\delta}) - F_t(-|t| - \hat{\delta}), \end{aligned}$$

siendo F_t la función de distribución de una t de Student.

Cuando el intervalo de equivalencia es simétrico, es decir, $d_U = -d_L = \Delta$, Rocke (1984) desarrolló de manera independiente un test basado en el estadístico

$$F = \frac{\hat{d}^2}{se(\hat{d})^2},$$

cuya distribución es una F de Snedecor no central (vease sección A.3 del apéndice) con 1 y $n_1 + n_2 - 2$ grados de libertad de parámetro de no centralidad

$$\lambda = \frac{n_1 n_2 \Delta^2}{(n_1 + n_2)\sigma^2}.$$

Martín Andrés (1990) ilustró que el test de Rocke es un caso particular del test de Anderson-Hauck ya que $T_{AH}^2 = F$ y ${t'}^2_{n_1+n_2-2}(\lambda) = F'_{1,n_1+n_2-2}(\lambda)$ si $d_U = -d_L = \Delta$.

2.1.3. Corrección de Welch

En los apartados 2.1.1 y 2.1.2 teníamos la hipótesis poco realista de que la varianza de ambos grupos coincidía. Dannenberg et al. (1994) propusieron un método basado en la corrección de Welch para el t-test. Este se basa en substituir $se(\hat{d})$ por

$$se_w(\hat{d}) = \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

y los grados de libertad $n_1 + n_2 - 2$ por

$$f = \frac{\left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right)^2}{\frac{\left(\frac{s_X^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_Y^2}{n_2}\right)^2}{n_2 - 1}}.$$

2.1.4. Test de Wellek

Wellek (2002) presentó en el capítulo 6 un test UMPI (uniformemente más potente invariante) para contrastar 2.2 bajo la suposición 2.1. Este se basa en el estadístico

$$T_W = \frac{\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}\hat{d}}{\sqrt{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2}}$$

La región crítica del test viene dada por

$$\{\tilde{C}_1 < T_W < \tilde{C}_2\},$$

siendo $\tilde{C}_1$ y $\tilde{C}_2$ las soluciones del siguiente sistema de ecuaciones

$$\left.\begin{aligned} G_{\tilde{d}_L}(C_2) - G_{\tilde{d}_L}(C_1) &= \alpha \\ G_{\tilde{d}_U}(C_2) - G_{\tilde{d}_U}(C_1) &= \alpha \end{aligned}\right\}$$

siendo $G_\delta(.)$ la función de distribución de una t de Student de $n_1 + n_2 - 2$ grados de libertad no central con parámetro de no centralidad δ, $\tilde{d}_L = d_L\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ y $\tilde{d}_U = d_U\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ los extremos del intervalo de equivalencia d_L y d_U respectivamente ponderados por $\sqrt{\frac{n_1 n_2}{n_1+n_2}}$.

Cuando el intervalo de equivalencia es simétrico, es decir, $d_U = -d_L = \Delta$, la región crítica admite la siguiente representación,

$$\left\{T_W^2 < F'^{-1}_{1,n_1+n_2-2,\alpha}\left(\frac{n_1 n_2 \Delta^2}{n_1 + n_2}\right)\right\},$$

donde $F'^{-1}_{\nu_1,\nu_2,\alpha}(\psi)$ es el cuantil α de una distribución F de Snedecor no central con ν_1 y ν_2 grados de libertad y parámetro de no centralidad ψ.

En el caso de que el intervalo de equivalencia sea simétrico y el tamaño de ambas poblaciones coincida, esto es, $d_U = -d_L = \Delta$ y $n_1 = n_2 = n$, la potencia del test cuando la diferencia real es $d = 0$ se puede calcular como

$$\beta_{n,\alpha} = 2F_t\left(F'^{-1}_{1,n_1+n_2-2,\alpha}\left(\frac{n^2\Delta^2}{2}\right)\right) - 1.$$

2.2. Tests no paramétricos

En el apartado 2.1 para realizar el contraste de equivalencia suponíamos que las muestras procedían de dos poblaciones normales independientes. En este apartado relajaremos las hipótesis sobre las poblaciones de ambas muestras suprimiendo la suposición de normalidad, para ello, propondremos dos tests no paramétricos, el primero de ellos basado en el bootstrap y el segundo inspirado en el test de Wilcoxon-Mann-Whitney.

2.2.1. Test de tipo Mann-Whitney

Wellek (1996) propuso un test no paramétrico para distribuciones continuas basado en el test de Wilcoxon-Mann-Whitney en la forma de Mann-Whitney. Entonces, siendo $X_1 \ldots, X_{n_1}$ e $Y_1 \ldots, Y_{n_2}$ dos muestras aleatorias simples independientes de $X \sim F$ e $Y \sim G$, respectivamente, con F y G funciones de distribución continuas. Denotemos por $\gamma(F,G) = P(X > Y)$. En este caso el contraste de equivalencia se plantea como

$$H_0 : \gamma(F,G) \notin \left[\frac{1}{2} - \varepsilon_1, \frac{1}{2} + \varepsilon_2\right] \tag{2.5}$$

frente a

$$H_1 : \gamma(F,G) \in \left(\frac{1}{2} - \varepsilon_1, \frac{1}{2} + \varepsilon_2\right),$$

donde ε_1 y ε_2 son dos valores prefijados de antemano. Es decir, queremos contrastar la hipótesis nula de que $\gamma(F,G)$ no está contenida en el intervalo $[\frac{1}{2} - \varepsilon_1, \frac{1}{2} + \varepsilon_2]$ frente a la alternativa de que sí lo está. Cuando rechacemos la hipótesis nula diremos que las localizaciones de ambas poblaciones son equivalentes. Un estimador razonable de γ será de la forma del estadístico de Mann-Whitney para dos muestras independientes, es decir,

$$\hat{\gamma} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{(0,\infty)}(X_i - Y_j).$$

A continuación daremos un procedimiento para calcular la distribución asintótica de dicho estadístico. Este se basará en la solución UMP para el siguiente problema

$$\begin{aligned} &H_0 : \theta \leq \theta_1 \ \text{ o } \ \theta \geq \theta_2, \\ &H_1 : \theta > \theta_1 \ \text{ y } \ \theta < \theta_2, \end{aligned} \tag{2.6}$$

donde θ denota la media de una distribución normal con varianza conocida σ^2. Sea $Z_1, \ldots, Z_k$ una muestra de dicha normal y θ_0 un valor fijo contenido en el intervalo (θ_1, θ_2) tal que cualquier $N(\theta, \sigma^2)$ con $\theta \in (\theta_1, \theta_2)$ es equivalente a la distribución objetivo $N(\theta_0, \sigma^2)$. En Wellek (1993) se puede ver que para este problema existe una solución UMP de nivel α cuya zona de rechazo es

$$\left\{ k\frac{(\bar{Z} - \frac{\theta_1+\theta_2}{2})^2}{\sigma^2} < {\chi'}_1^{2\,-1}\left(\sqrt{k}\frac{\theta_2 - \theta_1}{2\sigma}\right)\right\},$$

donde $\bar{Z}$ denota la media muestral de $Z_1, \ldots, Z_k$ y ${\chi'}_1^{2\,-1}(\psi)$ es el cuantil α de una distribución ${\chi'}^2$ no central (vease sección A.2 del apéndice) con un grado de libertad con parámetro de no centralidad $\psi > 0$.

Ahora, siguiendo el tercer capítulo de Randles y Wolfe (1979) conocemos para cualquier F y G la varianza de $\hat{\gamma}$,

$$\sigma^2_{\hat{\gamma}} = \frac{1}{n_1 n_2}\left[\gamma - (n_1 + n_2 - 1)\gamma^2 + (n_1 - 1)\gamma_{FFG} + (n_2 - 2)\gamma_{FGG}\right],$$

donde,

$$\gamma_{FFG} = \mathbb{P}(X_{i_1} > Y_j, X_{i_2} > Y_j) \quad \text{y} \quad \gamma_{FGG} = \mathbb{P}(X_i > Y_{j_1}, X_i > Y_{j_2}).$$

Para estimar estas funciones se proponen los siguientes estadísticos.

$$\hat{\gamma}_{FFG} = \binom{n_1}{2}^{-1} \frac{1}{n_2} \sum_{i_1=1}^{n_1-1} \sum_{i_2=i_1+1}^{n_1} \sum_{j=1}^{n_2} I_{(0,\infty)}(X_{i_1} - Y_j) I_{(0,\infty)}(X_{i_2} - Y_j),$$

y,

$$\hat{\gamma}_{FGG} = \binom{n_2}{2}^{-1} \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j_1=1}^{n_2-1} \sum_{j_2=j_1+1}^{n_2} I_{(0,\infty)}(X_i - Y_{j_1}) I_{(0,\infty)}(X_i - Y_{j_2}).$$

Por el corolario 3.49 de Randles y Wolfe (1979) los estimadores $\hat{\gamma}$, $\hat{\gamma}_{FFG}$ y $\hat{\gamma}_{FGG}$ son consistentes para γ, γ_{FFG} y γ_{FGG} respectivamente, por tanto,

$$\hat{\sigma}^2_{\hat{\gamma}} = \frac{1}{n_1 n_2} \left[\hat{\gamma} - (n_1 + n_2 - 1)\hat{\gamma}^2 + (n_1 - 1)\hat{\gamma}_{FFG} + (n_2 - 2)\hat{\gamma}_{FGG} \right],$$

es un estimador consistente de $\sigma^2_{\hat{\gamma}}$. Ahora, dada la normalidad asintótico del estadístico de Mann-Whitney y la consistencia de $\hat{\sigma}^2_{\hat{\gamma}}$ implica que

$$\frac{\hat{\gamma} - \gamma}{\hat{\sigma}_{\hat{\gamma}}} \xrightarrow{d} N(0,1)$$

cuando $0 < \gamma < 1$, n_1,$n_2 \to \infty$ y $\frac{n_1}{n_2} \to \lambda > 0$.

Volviendo al problema original y tomando $\theta = \frac{(\gamma - [\frac{1}{2} + \frac{\varepsilon_2 - \varepsilon_1}{2}])}{\sigma_{\hat{\gamma}}}$ y $\bar{\varepsilon} = \frac{\varepsilon_1 + \varepsilon_2}{2}$ tenemos que

$$\begin{aligned} H_0 &: \theta \notin \left[-\frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}}, \frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}} \right], \\ H_1 &: \theta \in \left(-\frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}}, \frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}} \right), \end{aligned} \tag{2.7}$$

es asintóticamente equivalente al problema (2.5) cuando tenemos una sola observación $Z = \frac{(\hat{\gamma} - [\frac{1}{2} + \frac{\varepsilon_2 - \varepsilon_1}{2}])}{\sigma_{\hat{\gamma}}}$ precedente de la distribución $N(\theta, 1)$. Ahora bien, el problema (2.7) es un caso particular de (2.6) con $k = 1$, $\theta_1 = -\frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}}$, $\theta_2 = \frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}}$ y $\sigma = 1$. Por tanto se obtiene un test UMP de nivel α para el problema reducido utilizando la siguiente región crítica

$$\left\{ \frac{(\hat{\gamma} - \frac{1}{2} - \frac{\varepsilon_2 - \varepsilon_1}{2})^2}{\sigma^2_{\hat{\gamma}}} < {\chi_1^2}^{-1} \left(\frac{\bar{\varepsilon}}{\sigma_{\hat{\gamma}}} \right) \right\}.$$

Por último, podemos substituir en la expresión anterior $\sigma^2_{\hat{\gamma}}$ por $\hat{\sigma}^2_{\hat{\gamma}}$ para obtener un test asintóticamente válido para el problema (2.5) como podemos ver en el apéndice de Wellek (1996).

2.2.2. Test basado en el bootstrap

En este apartado se propondrá un test basado en el bootstrap para realizar el contraste (2.2) suponiendo que el intervalo de equivalencia es simétrico, es decir, se quiere contrastar la hipótesis nula

$$H_0 : |d| \geq \Delta,$$

frente a la alternativa

$$H_1 : |d| < \Delta.$$

Sin embargo, dado que la función valor absoluto no es diferenciable, elevaremos los términos al cuadrado obteniendo un test equivalente. Por tanto, queremos contrastar la hipótesis nula

$$H_0 : d^2 \geq \Delta^2, \tag{2.8}$$

frente a la alternativa

$$H_1 : d^2 < \Delta^2.$$

Sean $X_1 \ldots, X_{n_1}$ e $Y_1 \ldots, Y_{n_2}$ dos muestras aleatorias simples independientes tales que $X_i \sim F$ y $Y_i \sim G$ para todo $i = 1, \ldots, n_1$ y $j = 1, \ldots, n_2$ siendo F e G dos distribuciones totalmente desconocidas. Para presentar un algoritmo bootstrap que nos permita realizar el contraste (2.8) necesitamos estimar las medias μ_X y μ_Y bajo la restricción $d^2 = \Delta^2$, es decir, tenemos que encontrar dos cantidades $\hat{\mu}_X$ y $\hat{\mu}_Y$ tales que minimicen

$$\begin{aligned} &\sum_{i=1}^{n_1}(X_i-\hat{\mu}_X)^2+\sum_{j=1}^{n_2}(Y_i-\hat{\mu}_Y)^2, \\ &s.a:(\hat{\mu}_X-\hat{\mu}_Y)^2=\Delta^2. \end{aligned} \tag{2.9}$$

Como es una optimización con restricciones de igualdad la resolveremos por el método de los multiplicadores de Lagrange, por tanto, tenemos que resolver el siguiente sistema de ecuaciones

$$\left.\begin{array}{l} \frac{\partial}{\partial\hat{\mu}_X}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \\ \frac{\partial}{\partial\hat{\mu}_Y}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \\ \frac{\partial}{\partial\lambda}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \end{array}\right\}$$

siendo

$$\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=\sum_{i=1}^{n_1}(X_i-\hat{\mu}_X)^2+\sum_{j=1}^{n_2}(Y_i-\hat{\mu}_Y)^2+\lambda((\hat{\mu}_X-\hat{\mu}_Y)^2-\Delta^2).$$

Ahora procedemos a resolver el sistema,

$$\frac{\partial}{\partial\hat{\mu}_X}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \iff -2\sum_{i=1}^{n_1}(X_i-\hat{\mu}_X)+2\lambda(\hat{\mu}_X-\hat{\mu}_Y)=0, \tag{2.10}$$

$$\frac{\partial}{\partial\hat{\mu}_Y}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \iff -2\sum_{j=1}^{n_2}(Y_i-\hat{\mu}_Y)-2\lambda(\hat{\mu}_X-\hat{\mu}_Y)=0, \tag{2.11}$$

$$\frac{\partial}{\partial\lambda}\mathscr{L}(\hat{\mu}_X,\hat{\mu}_Y,\lambda)=0 \iff (\hat{\mu}_X-\hat{\mu}_Y)^2=\Delta^2. \tag{2.12}$$

Para ello despejamos e igualamos λ de las ecuaciones (2.10) y (2.11) obteniendo que

$$\sum_{i=1}^{n_1}(X_i-\hat{\mu}_X)=-\sum_{j=1}^{n_2}(Y_i-\hat{\mu}_Y) \iff \hat{\mu}_Y=\frac{n_1\bar{X}+n_2\bar{Y}-n_1\hat{\mu}_X}{n_2}. \tag{2.13}$$

Ahora substituyendo $\hat{\mu}_Y$ de la ecuación (2.13) en la ecuación (2.12) obtenemos la siguiente ecuación cuadrática,

$$\left(\frac{n_1+n_2}{n_2}\right)^2\hat{\mu}_X^2+-2\frac{(n_1+n_2)(n_1\bar{X}+n_2\bar{Y})}{n_2^2}\hat{\mu}_X+\left(\frac{n_1\bar{X}+n_2\bar{Y}}{n_2}\right)^2-\Delta^2=0. \tag{2.14}$$

A continuación, resolviendo (2.14) obtenemos dos posibles soluciones,

$$\text{Solución 1:}\quad \hat{\mu}_X=\tfrac{n_1\bar{X}+n_2\bar{Y}+n_2\Delta}{n_1+n_2},\quad \hat{\mu}_Y=\tfrac{n_1\bar{X}+n_2\bar{Y}-n_1\Delta}{n_1+n_2}, \tag{2.15}$$

$$\text{Solución 2:}\quad \hat{\mu}_X=\tfrac{n_1\bar{X}+n_2\bar{Y}-n_2\Delta}{n_1+n_2},\quad \hat{\mu}_Y=\tfrac{n_1\bar{X}+n_2\bar{Y}+n_1\Delta}{n_1+n_2}. \tag{2.16}$$

Por último, para determinar si las posibles soluciones (2.15) y (2.16) son mínimos locales debemos ver si el determinante de la hessiana orlada es menor que cero,

$$\begin{vmatrix} 0 & 2(\hat{\mu}_X-\hat{\mu}_Y) & -2(\hat{\mu}_X-\hat{\mu}_Y) \\ 2(\hat{\mu}_X-\hat{\mu}_Y) & 2n_1+2\lambda & -2\lambda \\ -2(\hat{\mu}_X-\hat{\mu}_Y) & -2\lambda & 2n_2+2\lambda \end{vmatrix} = -8(n_1+n_2)(\hat{\mu}_X-\hat{\mu}_Y)^2<0.$$

Si en la ecuación (2.14) despejamos $\hat{\mu}_X$ podemos encontrar de manera análoga otras dos soluciones al problema (2.9),

$$\text{Solución 3:} \quad \hat{\mu}_X = \frac{n_1\bar{X}+n_2\bar{Y}+n_1\Delta}{n_1+n_2}, \quad \hat{\mu}_Y = \frac{n_1\bar{X}+n_2\bar{Y}-n_2\Delta}{n_1+n_2}, \tag{2.17}$$

$$\text{Solución 4:} \quad \hat{\mu}_X = \frac{n_1\bar{X}+n_2\bar{Y}-n_1\Delta}{n_1+n_2}, \quad \hat{\mu}_Y = \frac{n_1\bar{X}+n_2\bar{Y}+n_2\Delta}{n_1+n_2}. \tag{2.18}$$

Por tanto, hemos encontrado 4 soluciones (2.15), (2.16), (2.17), (2.18) para el problema (2.9).

Ahora, estamos en condiciones de presentar el algoritmo bootstrap.

1. Calculamos las medias muestrales $\bar{X}$ y $\bar{Y}$, los estimadores (2.15) $\hat{\mu}_X$ y $\hat{\mu}_Y$ y las funciones de distribución empíricas centradas $\hat{F}_{n_1}$ y $\hat{G}_{n_2}$ de las poblaciones X e Y respectivamente.

2. Calculamos el estadístico $\hat{d}^2 = (\bar{X} - \bar{Y})^2 - \Delta^2$ y los estimadores $\tilde{\mu}_X$ y $\tilde{\mu}_Y$ dados por

$$\tilde{\mu}_X = \begin{cases} \bar{X} & si \quad \hat{d}^2 \geq \Delta^2 \\ \hat{\mu}_X & si \quad \hat{d}^2 < \Delta^2 \end{cases}$$

$$\tilde{\mu}_Y = \begin{cases} \bar{Y} & si \quad \hat{d}^2 \geq \Delta^2 \\ \hat{\mu}_Y & si \quad \hat{d}^2 < \Delta^2 \end{cases}$$

3. Calculamos el estadístico bootstrap de la siguiente manera.

   i. Se arrojan una remuestras bootstrap $X_1^* \ldots, X_{n_1}^*$ e $Y_1^* \ldots, Y_{n_2}^*$ de $\hat{F}_{n_1}$ y $\hat{G}_{n_2}$ respectivamente.

   ii. Se calcula $X_1^{**} \ldots, X_{n_1}^{**}$ e $Y_1^{**} \ldots, Y_{n_2}^{**}$ siendo

$$X_i^{**} = X_i^* + \tilde{\mu}_X \quad \forall i = 1, \ldots, n_1,$$

$$Y_j^{**} = Y_j^* + \tilde{\mu}_Y \quad \forall j = 1, \ldots, n_2.$$

   iii. Se calcula $\hat{d}^{2*} = (\bar{X}^{**} - \bar{Y}^{**})^2 - \Delta^2$

4. Ahora, repetimos B veces el paso 3 para generar $\hat{d}_1^{2*}, \ldots, \hat{d}_B^{2*}$ estadísticos bootstrap.

5. Por último, siendo $\hat{d}_{(1)}^{2*} \leq \ldots \leq \hat{d}_{(B)}^{2*}$ los estadísticos bootstrap ordenados, rechazaremos la hipótesis nula si

$$\hat{d}^2 < \hat{d}_{(\lfloor B\alpha \rfloor)}^{2*}.$$

   Es decir, rechazaremos la hipótesis nula si $\hat{d}^2$ es menor que el elemento $\lfloor B\alpha \rfloor$ de los estadísticos bootstrap ordenados.

2.3. Estudio de simulación

En este apartado vamos a comparar mediante un estudio de Monte Carlo los tests para la equivalencia de medias de dos poblaciones estudiados hasta ahora bajo dos situaciones distintas, una diseñada para que nos permita comparar el test de tipo Mann-Whitney y otra situación más acercada a la realidad utilizando como intervalo de equivalencia el propuesto en el capítulo 1.

2.3.1. Situación para comparar el test de tipo Mann-Whitney

Para poder comparar los tests expuestos en la literatura con el nuevo test basado en el bootstrap se simularán poblaciones normales, $N(0,1)$, uniformes, $U\left(-\frac{1}{2},\frac{1}{2}\right)$ y exponenciales, $Exp(1)$, bajo las hipótesis nula y alternativa. Esto nos permitirá comparar la significación y potencia de los tests con un nivel de significación de $\alpha = 0.05$. Para poder realizar la simulación necesitamos conocer la relación entre los intervalos de equivalencia planteados en 2.2, 2.8 y 2.5, es decir, los intervalos de equivalencia de los tests basados en la normalidad de los datos, el test basado en el bootstrap y el test de tipo Mann-Whitney. Mientras que la relación entre los tests basados en la normalidad y el bootstrap es inmediata, la relación con el test tipo Mann-Whitney no es tan sencilla. Por tanto, siendo $X_i = X_i^0 + \nu$ y $Y_j = Y_j^0$ donde X_i^0 y Y_j^0 muestras procedentes de poblaciones F_0 y G_0 respectivamente, siguiendo Wellek (1996) se debe escoger un ν tal que

$$w_+(\nu) = \int_{-\infty}^{\infty} [1 - F_0(y-\nu)] dG_0(y)$$

tome un valor fijado de antemano w^* tal que $w^* \in (0,1)$. En general no es factible resolver esta integral, sin embargo, para las distribuciones que vamos a estudiar existen las siguientes expresiones explícitas si F_0 y G_0 coinciden.

$$\begin{array}{ll} N(0,1) & w_+(\nu) = \phi\left(\frac{\sqrt{2}}{2}\nu\right) \\ U\left(-\frac{1}{2},\frac{1}{2}\right)^{\text{I}} & w_+(\nu) = \begin{cases} \frac{1}{2} + \nu(1-\frac{\nu}{2}), & \text{si} \quad \nu \in \left[0,\frac{1}{2}\right) \\ 1 & \text{si} \quad \nu \geq \frac{1}{2} \end{cases} \\ Exp(1) & w_+(\nu) = 1 - \Psi_4(\text{máx}\{-4\nu, 0\}) \end{array}$$

donde ϕ y Ψ_4 son las funciones de distribución de una normal estándar y una χ^2 con cuatro grados de libertad respectivamente.

Significación

Comenzaremos aproximando la significación, es decir, la probabilidad de aceptar la hipótesis nula cuando esta es verdadera. Para realizar esta tarea simularemos dos poblaciones normales estándar, dos uniformes entre menos un medio y un medio y dos exponenciales de parámetro uno. Concretamente se simularán 1000 muestras $X_i = X_i^0 + \nu$ y $Y_j = Y_j^0$ donde X_i^0 y Y_j^0 proceden de poblaciones F_0 y G_0, respectivamente, siendo estas unas poblaciónes de las mencionadas anteriormente. Las muestras simuladas serán de tamaños 15, 30 y 50. Por último, para realizar las simulaciones se tomaron los valores para el test de tipo Mann-Whitney $\varepsilon_1 = \varepsilon_2 = 0.2$, por tanto, tenemos que los valores de d_L y d_U se obtienen despejando

$$w_+(d_L) = 0.3,$$

$$w_+(d_U) = 0.7.$$

Realizando las operaciones oportunas obtenemos los intervalos de equivalencia para las distintas distribuciones que vamos a estudiar. En la tabla 2.1 se muestran las distribuciones que siguen las poblaciones que vamos a usar en el estudio de simulación 2.3.1 y sus correspondientes intervalos de equivalencia.

Una vez vista la relación entre el intervalo de equivalencia del test Mann-Whitney $\left(\frac{1}{2}-\varepsilon_1, \frac{1}{2}+\varepsilon_2\right)$ y el resto de tests, estamos en condiciones de ver la significación de los tests.

En la tabla 2.2 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando las poblaciones F_0 y G_0 sigan una distribución $N(0,1)$. En la parte de la izquierda se muestra la proporción de rechazos cuando $d = -0.7416$ y en la derecha se muestra la proporción de rechazos cuando $d = 0.7416$. En la tabla 2.2 se puede observar que para los diferentes tamaños muestrales todos los test presentan

[I] Para $\nu < 0$ se puede usar la relación $w_+(-\nu) = 1 - w_+(\nu)$ valida para cualquier familia de localización con función de densidad simétrica.

Poblaciones		Intervalo de equivalencia
$F_0 \sim N(0,1)$	$G_0 \sim N(0,1)$	(-0.7416, 0.7416)
$F_0 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$	$G_0 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$	(-0.2254, 0.2254)
$F_0 \sim Exp(1)$	$G_0 \sim Exp(1)$	(-0.5108, 0.5108)

Tabla 2.1: Distribuciones que siguen las poblaciones usadas en la simulación y sus correspondientes intervalos de equivalencia

	$d = -0.7416$						$d = 0.7416$				
(n_1, n_2)	TOST	H-A	W	M-W	Boot	(n_1, n_2)	TOST	H-A	W	M-W	Boot
(15,15)	0.047	0.059	0.055	0.048	0.063	(15,15)	0.041	0.047	0.038	0.046	0.050
(15,30)	0.054	0.057	0.049	0.045	0.065	(15,30)	0.044	0.062	0.050	0.050	0.069
(30,15)	0.038	0.040	0.046	0.038	0.046	(30,15)	0.058	0.061	0.050	0.048	0.065
(15,50)	0.056	0.059	0.055	0.044	0.062	(15,50)	0.054	0.056	0.055	0.48	0.062
(50,15)	0.046	0.047	0.046	0.038	0.055	(50,15)	0.041	0.041	0.046	0.043	0.053
(30,30)	0.059	0.059	0.063	0.048	0.063	(30,30)	0.045	0.046	0.043	0.042	0.048
(30,50)	0.037	0.037	0.034	0.031	0.044	(30,50)	0.042	0.042	0.043	0.039	0.043
(50,30)	0.052	0.052	0.056	0.050	0.052	(50,30)	0.058	0.058	0.059	0.057	0.063
(50,50)	0.063	0.063	0.067	0.059	0.064	(50,50)	0.049	0.049	0.053	0.046	0.050

Tabla 2.2: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0, G_0 \sim N(0,1)$

una proporción de rechazos en tanto por uno similar al nivel de significación $\alpha = 0.05$, por tanto, bajo normalidad todos los tests aceptan la hipótesis nula cuando esta es verdadera.

En la tabla 2.3 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando las poblaciones F_0 y G_0 sigan una distribución $U\left(-\frac{1}{2}, \frac{1}{2}\right)$. En la parte de la izquierda se muestra la proporción de rechazos cuando $d = -0.7416$ y en la derecha se muestra la proporción de rechazos cuando $d = 0.7416$. En la tabla 2.3 se puede ver que, salvo el test de Wellek, todos aproximan de manera correcta el nivel de significación $\alpha = 0.05$. El test de Wellek tiene una proporción de rechazos muy por debajo del nivel de significación y por tanto en este caso se comporta de manera muy conservadora.

En la tabla 2.4 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando las poblaciones F_0 y G_0 sigan una distribución $Exp(1)$. En la parte de la izquierda se muestra la proporción de rechazos cuando $d = -0.7416$ y en la derecha se muestra la proporción de rechazos cuando $d = 0.7416$. En la tabla 2.4 se puede ver que en los test de Anderson-Hauck, Wellek y de tipo Mann-Whitney respetan el nivel de significación. El TOST cuando uno de los tamaños muestrales es 15 la proporción de rechazos es ligeramente inferior al nivel de significación. Por último el test basado en el bootstrap presenta una proporción de rechazos cercano a 0.1 cuando $n_1 = 50$ y $n_2 = 15$ si $d = -0.5108$ y cuando $n_1 = 15$ y $n_2 = 50$ si $d = 0.5108$.

	$d = -0.2254$						$d = 0.2254$				
(n_1, n_2)	TOST	A-H	W	M-W	Boot	(n_1, n_2)	TOST	A-H	W	M-W	Boot
(15,15)	0.051	0.054	0.003	0.044	0.061	(15,15)	0.052	0.060	0.006	0.043	0.065
(15,30)	0.043	0.044	0.001	0.040	0.057	(15,30)	0.047	0.047	0.002	0.051	0.056
(30,15)	0.060	0.060	0.004	0.051	0.068	(30,15)	0.064	0.064	0.002	0.047	0.066
(15,50)	0.052	0.052	0.001	0.050	0.058	(15,50)	0.056	0.056	0.002	0.048	0.062
(50,15)	0.051	0.052	0.001	0.051	0.062	(50,15)	0.043	0.044	0.004	0.051	0.057
(30,30)	0.053	0.053	0.001	0.044	0.059	(30,30)	0.052	0.052	0.001	0.039	0.056
(30,50)	0.061	0.061	0.000	0.059	0.067	(30,50)	0.061	0.061	0.000	0.049	0.067
(50,30)	0.042	0.042	0.000	0.035	0.051	(50,30)	0.055	0.055	0.000	0.046	0.053
(50,50)	0.054	0.054	0.000	0.050	0.056	(50,50)	0.031	0.043	0.000	0.037	0.043

Tabla 2.3: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0, G_0 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$

	$d = -0.5108$						$d = 0.5108$				
(n_1, n_2)	TOST	A-H	W	M-W	Boot	(n_1, n_2)	TOST	A-H	W	M-W	Boot
(15,15)	0.014	0.062	0.050	0.052	0.063	(15,15)	0.008	0.048	0.044	0.052	0.047
(15,30)	0.023	0.060	0.055	0.056	0.055	(15,30)	0.020	0.065	0.062	0.056	0.053
(30,15)	0.019	0.061	0.052	0.053	0.055	(30,15)	0.015	0.047	0.043	0.052	0.057
(15,50)	0.022	0.051	0.047	0.050	0.099	(15,50)	0.024	0.059	0.055	0.043	0.044
(50,15)	0.030	0.062	0.056	0.055	0.049	(50,15)	0.023	0.053	0.054	0.049	0.098
(30,30)	0.043	0.057	0.051	0.048	0.060	(30,30)	0.029	0.045	0.043	0.042	0.049
(30,50)	0.037	0.037	0.034	0.031	0.044	(30,50)	0.047	0.052	0.063	0.046	0.048
(50,30)	0.043	0.048	0.041	0.043	0.037	(50,30)	0.039	0.046	0.043	0.038	0.056
(50,50)	0.052	0.052	0.051	0.047	0.054	(50,50)	0.041	0.041	0.057	0.044	0.045

Tabla 2.4: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0, G_0 \sim Exp(1)$

Potencia

A continuación realizaremos varias simulaciones que nos permitan aproximar la curva potencia bajo diferentes poblaciones. Para realizar esta tarea se simularán 5000 muestras procedentes las mismas poblaciones vistas en 2.3.1 con tamaños muestrales de 25 y 40 de tal forma que $X_i = X_i^0 + \nu * \xi$ y $Y_j = Y_j^0$ donde X_i^0 y Y_j^0 proceden de las anteriores poblaciones y $\xi \in [-1.3, 1.3]$

En los gráficos se muestra las aproximaciones de la curva de potencia bajo las diferentes poblaciones. En negro tenemos el TOST, en verde el test de Anderson-Hauck, en rojo el test de Wellek, en rosa el test de tipo Mann-Whitney y en azul el test basado en el bootstrap. La intersección de las lineas punteadas marcan los extremos de la hipótesis nula.

En las figuras 2.4 y 2.5 se puede ver una aproximación de la curva potencia cuando los datos proceden de poblaciones $N(0, 1)$ cuando los tamaños muestrales son $n_1 = n_2 = 25$ y $n_1 = n_2 = 40$, respectivamente. En ella se puede observar que la potencia de todos los tests es muy similar siendo el test basado en el bootstrap el que tiene una mayor potencia y el test de tipo Mann-Whitney el que una potencia menor.

En las figuras 2.6 y 2.7 se puede ver una aproximación de la curva potencia cuando los datos proceden de poblaciones $U\left(-\frac{1}{2}, \frac{1}{2}\right)$ cuando los tamaños muestrales son $n_1 = n_2 = 25$ y $n_1 = n_2 = 40$, respectivamente. Se observa que la potencia del TOST, el test de Anderson-Hauck y el test basado en el bootstrap es muy similar siendo el test basado en el bootstrap el que tiene una mayor potencia. El test de tipo Mann-Whitney presenta una potencia ligeramente inferior y el test de Wellek no tiene potencia.

En las figuras 2.8 y 2.9 se puede ver una aproximación de la curva potencia cuando los datos proceden de poblaciones $Exp(1)$ cuando los tamaños muestrales son $n_1 = n_2 = 25$ y $n_1 = n_2 = 40$, respectivamente. El test de tipo Mann-Whitney presenta una potencia notablemente superior al resto. Cuando los tamaños muestrales son $n_1 = n_2 = 25$ los tests de Anderson-Hauck y el basado en el bootstrap presentan una mayor potencia que el TOST, y este, a su vez presenta una mayor potencia que el test de Wellek. Cuando los tamaños muestrales son $n_1 = n_2 = 40$ todos los test salvo el de tipo Mann-Whitney tienen una potencia casi idéntica.

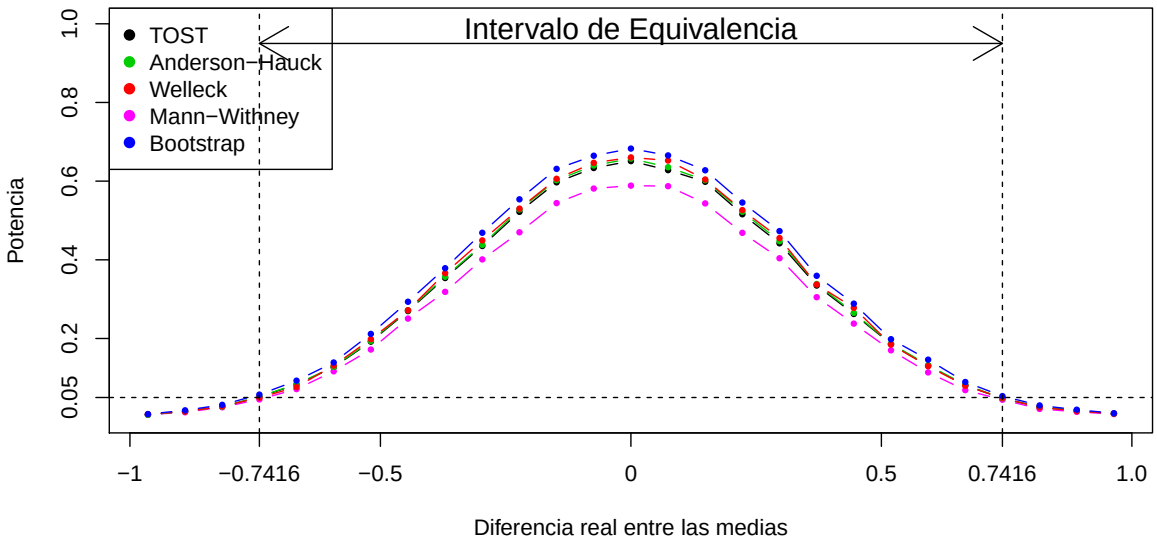

Figura 2.4: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim N(0,1)$ con tamaños muestrales $n_1 = n_2 = 25$

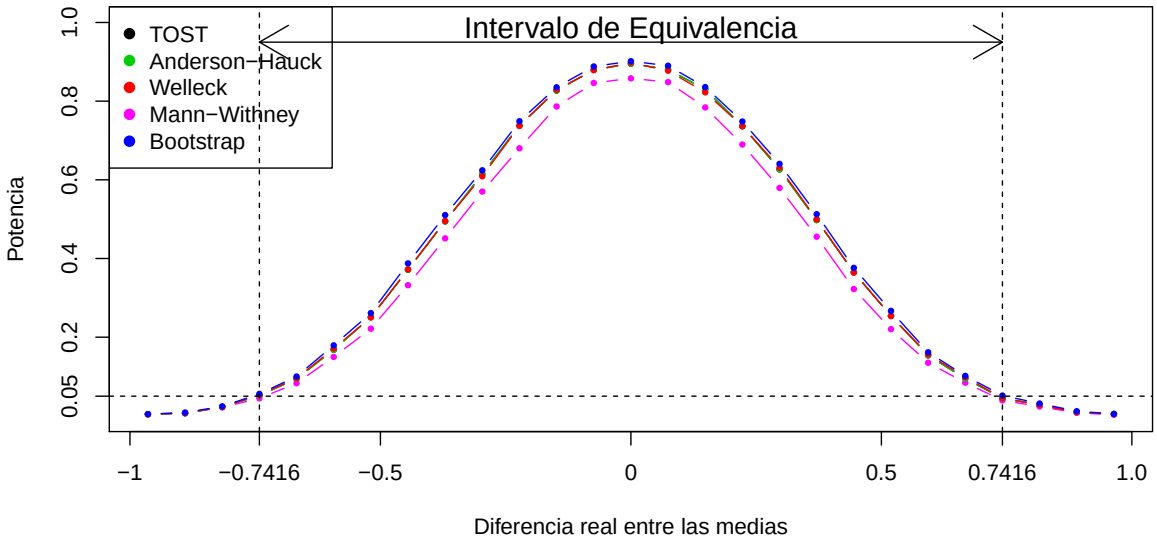

Figura 2.5: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim N(0,1)$ con tamaños muestrales $n_1 = n_2 = 40$

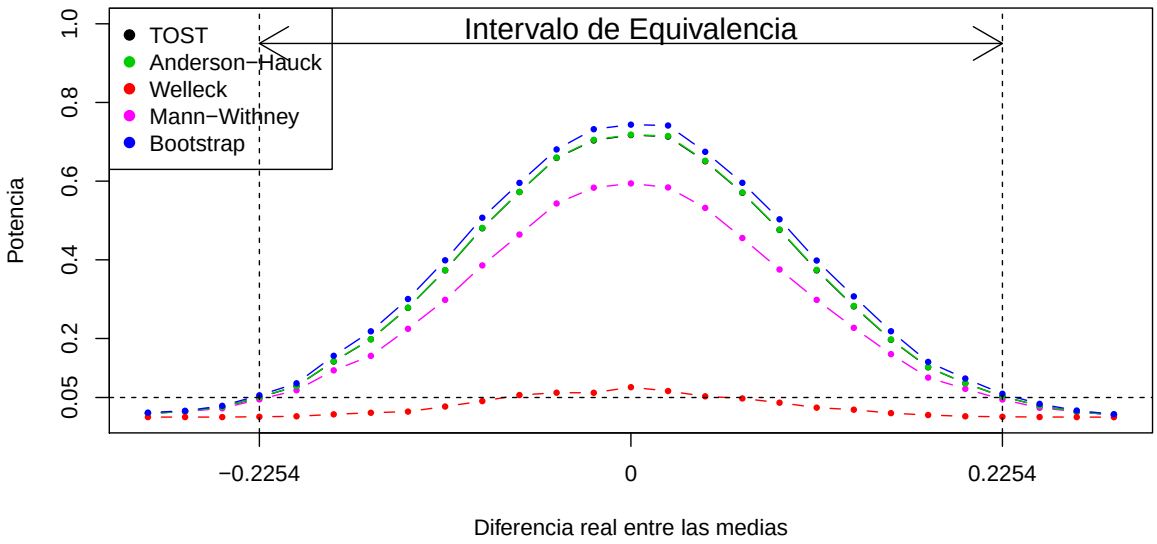

Figura 2.6: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$ con tamaños muestrales $n_1 = n_2 = 25$

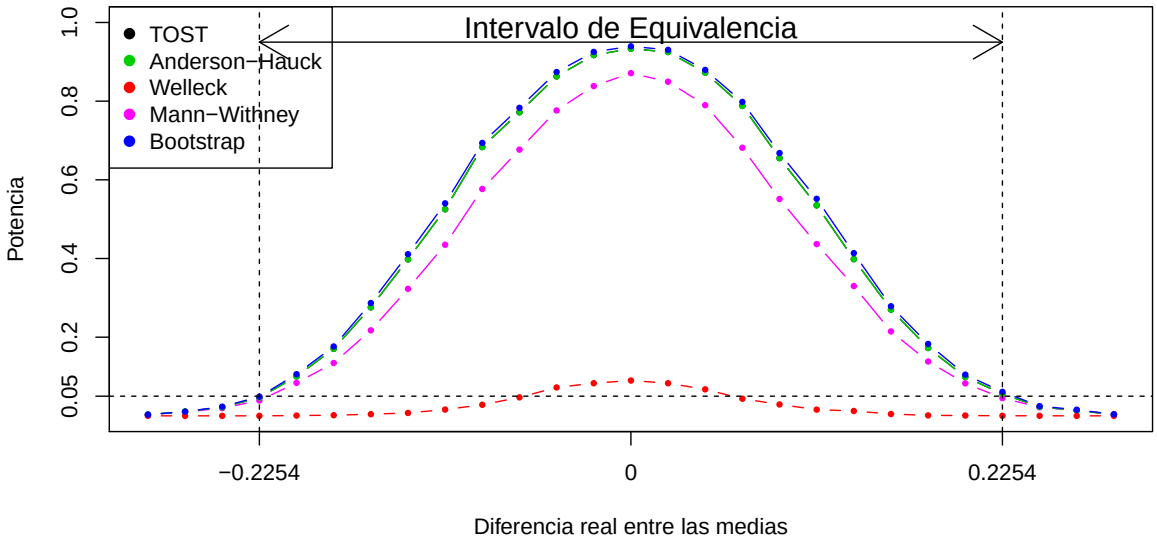

Figura 2.7: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$ con tamaños muestrales $n_1 = n_2 = 40$

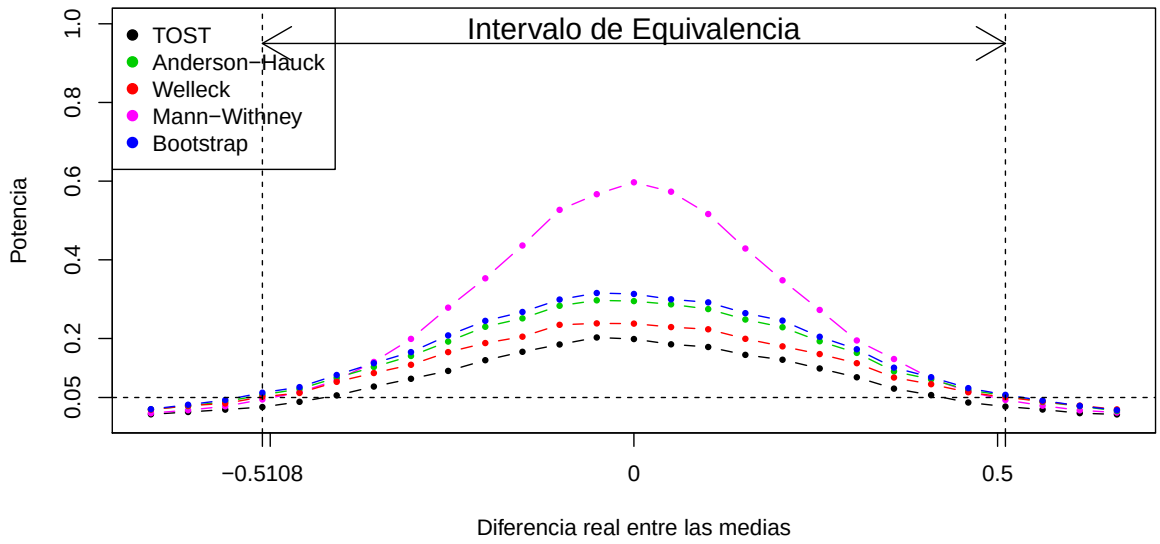

Figura 2.8: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim Exp(1)$ con tamaños muestrales $n_1 = n_2 = 25$

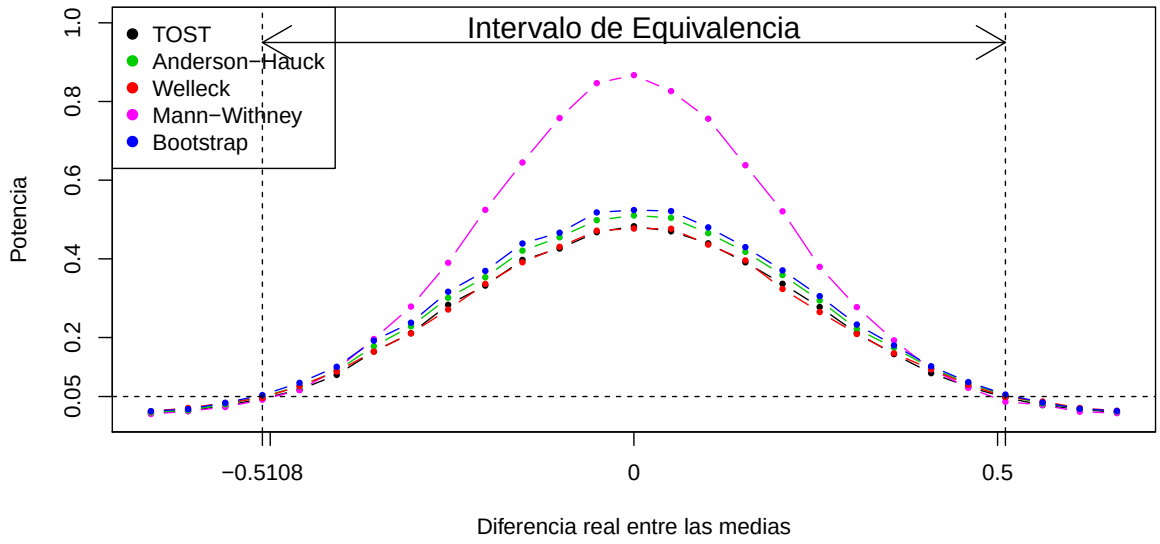

Figura 2.9: Aproximación de las curvas de potencia de los tests cuando $F_0, G_0 \sim Exp(1)$ con tamaños muestrales $n_1 = n_2 = 40$

2.3.2. Situación más realista

En el anterior apartado hemos visto las limitaciones del test de tipo Mann-Whitney para ser aplicado a una situación realista, por ello, a continuación vamos a realizar simulaciones en escenarios que nos permitan comparar la potencia de los tests en una situación más plausible. Por tanto, vamos a considerar dos poblaciones diferentes F_0 y G_0 independientes y con la misma media. Tomaremos como población de referencia F_0 y consideraremos como intervalo de equivalencia el propuesto en el capítulo 1.

En la tabla 2.5 se muestran las distribuciones que siguen las poblaciones que vamos a usar en el estudio de simulación 2.3.2 y sus correspondientes intervalos de equivalencia.

Significación

Igual que hicimos en el subapartado 2.3.1 comenzaremos aproximando la significación. Para ello simularemos 1000 muestras $X_i = X_i^0 + \nu$ y $Y_j = Y_j^0$ donde $X_i^0 \sim F_0$ y $Y_j^0 \sim G_0$ siguen alguna de las poblaciones expuestas en la tabla 2.5 y ν tomará uno de los valores extremos del intervalo de equivalencia. Las muestras simuladas serán de tamaños 30 y 50.

En la tabla 2.6 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando la población F_0 siga una distribución $N\left(\frac{5}{4}, \frac{9}{16}\right)$ y G_0 siga una $U\left(0, \frac{5}{2}\right)$. En la parte izquierda se muestra la proporción de rechazos cuando $d = -0.2849$ y en la parte derecha cuando $d = 0.2849$. En la tabla 2.6 podemos observar que el test TOST presenta una proporción de rechazos muy inferior al nivel de significación, el test de Wellek presenta una proporción de rechazos ligeramente inferior al nivel de significación y los tests de Anderson-Hauck y el basado en el bootstrap presenta una proporción de rechazos similar al nivel de significación.

En la tabla 2.7 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando la población F_0 siga una distribución $Exp(\frac{10}{9})$ y G_0 siga una $Weibull\left(\frac{3}{2}, 1\right)$. En la parte izquierda se muestra la proporción de rechazos cuando $d = -0.2649$ y en la parte derecha cuando $d = 0.2649$. En la tabla 2.7 se observa que el test TOST cuando los tamaños muestrales $n_1 = n_2 = 30$ o la diferencia entre las medias es $d = 0.2649$ presenta una proporción de rechazos muy inferior al del nivel de significación. Además, el test TOST para el resto de casos y el test de Wellek presentan una proporción de rechazos ligeramente inferior al nivel de significación. Los tests de Anderson-Hauck y el basado en el bootstrap presentan una proporción de rechazos similar al nivel de significación.

En la tabla 2.8 se muestra la proporción de rechazos en los extremos de la hipótesis nula cuando la población F_0 siga una distribución $Log - N(0, \frac{1}{2})$ y G_0 siga una $Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$. En la parte izquierda se muestra la proporción de rechazos cuando $d = -0.1836$ y en la parte derecha cuando $d = 0.1836$. En la tabla 2.8 se puede observar que el test TOST no presenta ningún rechazo. Los tests de Anderson-Hauck, Wellek y el basado en el bootstrap presentan una proporción de rechazos similar al nivel de significación.

Poblaciones		Intervalo de equivalencia
$F_0 \sim N\left(\frac{5}{4}, \frac{9}{16}\right)$	$G_0 \sim U\left(0, \frac{5}{2}\right)$	(-0.2849,0.2849)
$F_0 \sim Exp(\frac{10}{9})$	$G_0 \sim Weibull\left(\frac{3}{2}, 1\right)$	(-0.2649,0.2649)
$F_0 \sim Log - N\left(0, \frac{1}{2}\right)$	$G_0 \sim Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$	(-0.1836,0.1836)

Tabla 2.5: Distribución de las poblaciones usadas en la simulación y sus correspondientes intervalos de equivalencia.

	$d = -0.2849$					$d = 0.2849$			
(n_1, n_2)	TOST	H-A	W	Boot	(n_1, n_2)	TOST	H-A	W	Boot
(30,30)	0.001	0.046	0.030	0.045	(30,30)	0	0.054	0.028	0.052
(30,50)	0.001	0.046	0.027	0.044	(30,50)	0.003	0.058	0.032	0.061
(50,30)	0.002	0.046	0.026	0.048	(50,30)	0.001	0.042	0.025	0.041
(50,50)	0.017	0.042	0.022	0.046	(50,50)	0.021	0.053	0.029	0.056

Tabla 2.6: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0 \sim N\left(\frac{5}{4}, \frac{9}{16}\right)$ y $G_0 \sim U\left(0, \frac{5}{2}\right)$

	$d = -0.2649$					$d = 0.2649$			
(n_1, n_2)	TOST	H-A	W	Boot	(n_1, n_2)	TOST	H-A	W	Boot
(30,30)	0.004	0.036	0.029	0.039	(30,30)	0.009	0.056	0.028	0.057
(30,50)	0.029	0.041	0.028	0.038	(30,50)	0	0.057	0.032	0.059
(50,30)	0.027	0.049	0.032	0.052	(50,30)	0.003	0.050	0.021	0.050
(50,50)	0.033	0.037	0.025	0.039	(50,50)	0.005	0.051	0.029	0.052

Tabla 2.7: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0 \sim Exp(\frac{10}{9})$ y $G_0 \sim Weibull\left(\frac{3}{2}, 1\right)$

	$d = -0.1836$					$d = 0.1836$			
(n_1, n_2)	TOST	H-A	W	Boot	(n_1, n_2)	TOST	H-A	W	Boot
(30,30)	0	0.044	0.033	0.045	(30,30)	0	0.038	0.026	0.034
(30,50)	0	0.052	0.043	0.055	(30,50)	0	0.039	0.041	0.040
(50,30)	0	0.068	0.050	0.065	(50,30)	0	0.045	0.036	0.038
(50,50)	0	0.056	0.046	0.053	(50,50)	0	0.053	0.051	0.052

Tabla 2.8: Proporción de rechazos en los extremos de la hipótesis nula cuando $F_0 \sim Log - N(0, \frac{1}{2})$ y $G_0 \sim Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$

Potencia

A continuación aproximaremos las curvas de potencia para los diferentes tests. Para esto simularemos 5000 muestras procedentes de las poblaciones expuestas en la tabla 2.5 con tamaños muestrales $n_1 = n_2 = 100$ de tal forma que $X_i = X_i^0 + \nu$ y $Y_j = Y_j^0$ donde $X_i^0 \sim F_0$ y $Y_j^0 \sim G_0$ y $\nu \in (-1.3\Delta, 1.3\Delta)$ siendo Δ el extremo superior del intervalo de equivalencia. En negro se muestra la aproximación de la curva de potencia del test TOST, en verde la del test de Anderson-Hauck, en rojo la del test de Wellek y en azul la del test basado en el bootstrap.

En la figura 2.10 se muestra la aproximación de las curvas de potencia cuando la población F_0 siga una distribución $N\left(\frac{5}{4}, \frac{9}{16}\right)$ y G_0 siga una $U\left(0, \frac{5}{2}\right)$ con $n_1 = n_2 = 100$. En ella podemos ver que los tests TOST, Anderson-Hauck y el basado en el bootstrap presentan una potencia muy similar. Además, observamos que el test de Wellek es el que presenta una menor potencia.

En la figura 2.11 se muestra la aproximación de las curvas de potencia cuando la población F_0 siga una distribución $Exp\left(\frac{10}{9}\right)$ y G_0 siga una $Weibull\left(\frac{3}{2}, 1\right)$ con $n_1 = n_2 = 100$. En la figura 2.11 ocurre una situación similar a la de la figura 2.10 donde los tests TOST, Anderson-Hauck y el basado en el bootstrap presentan una potencia similar mientras que el test de Wellek tiene una potencia bastante inferior.

En la figura 2.12 se muestra la aproximación de las curvas de potencia cuando la población F_0 siga una distribución $Log - N\left(0, \frac{1}{2}\right)$ y G_0 siga una $Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$ con $n_1 = n_2 = 100$. En la figura 2.12 observamos que los tests tiene una baja potencia. Esto se debe a que la desviación típica de la distribución $Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$ es elevada. Además, observamos que los tests de Anderson-Hauck y el basado en el bootstrap tiene una potencia muy similar. El test de Wellek presenta una potencia ligeramente inferior con respecto a los tests anteriores y el test TOST no presenta potencia.

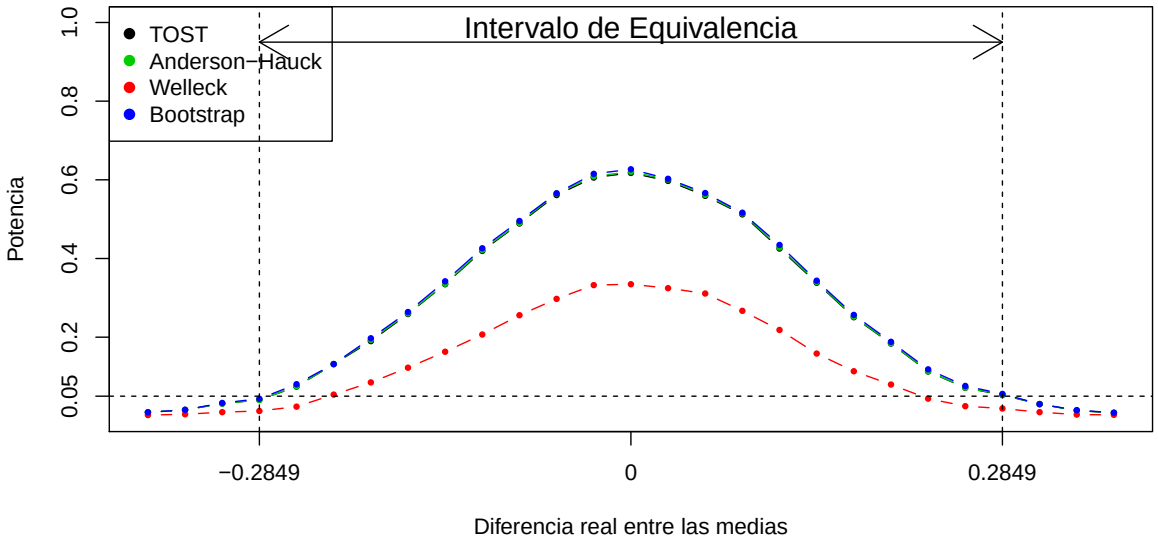

Figura 2.10: Aproximación de las curvas de potencia de los tests cuando $F_0 \sim N\left(\frac{5}{4}, \frac{9}{16}\right)$ y $G_0 \sim U\left(0, \frac{5}{2}\right)$ con $n_1 = n_2 = 100$

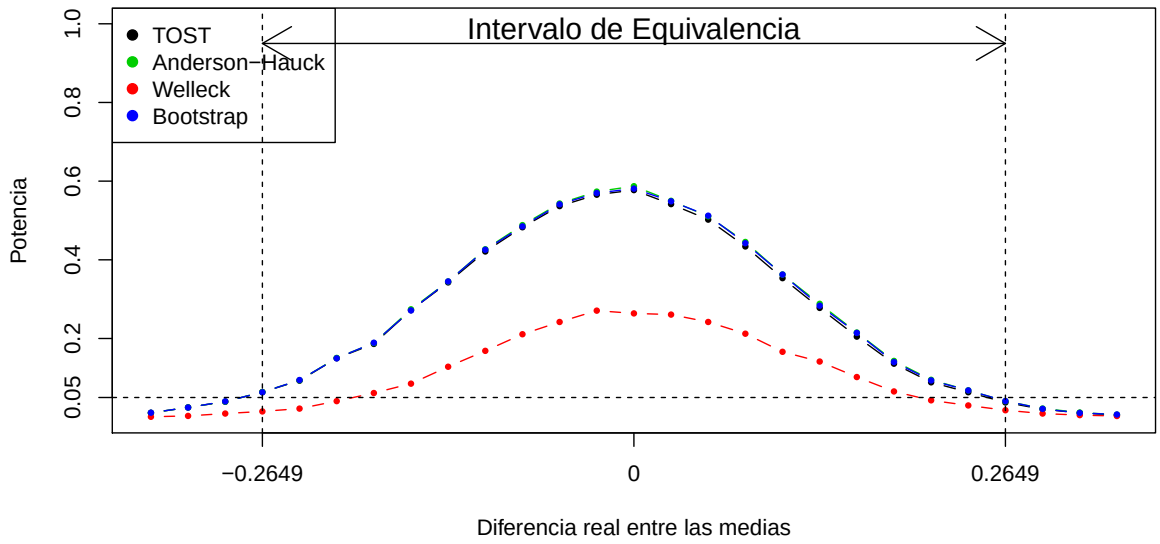

Figura 2.11: Aproximación de las curvas de potencia de los tests cuando $F_0 \sim Exp\left(\frac{10}{9}\right)$ y $G_0 \sim Weibull\left(\frac{3}{2}, 1\right)$ con $n_1 = n_2 = 100$

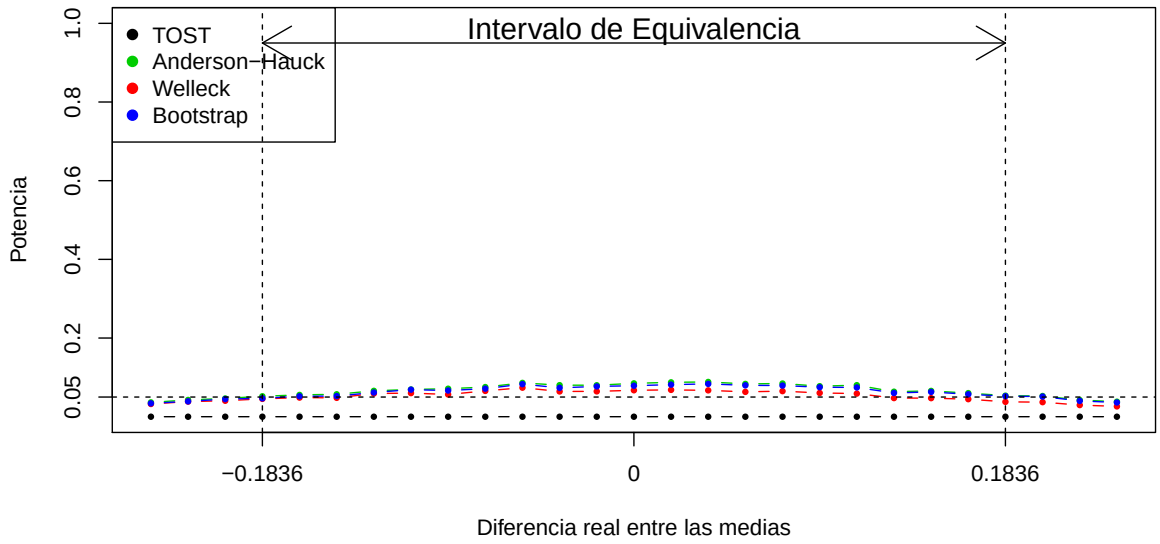

Figura 2.12: Aproximación de las curvas de potencia de los tests cuando $F_0 \sim Log - N\left(0, \frac{1}{2}\right)$ y $G_0 \sim Exp\left(\frac{1}{e^{1+\left(\frac{1}{2}\right)^2}}\right)$ con $n_1 = n_2 = 100$

2.4. Aplicación a datos reales

Aplicaremos los diferentes tests al conjunto de datos expuesto en Biesheuvel y Hothorn (2002), un estudio sobre la aplicación de un tratamiento para los hombres que sufren el síndrome de intestino irritable, el cual, ya ha sido testado en mujeres. Para ello se mide el dolor abdominal promedio en una escala de 0 (sin dolor) hasta 4 (incapacitación) expuestos a un placebo o a 4 dosis diferentes. Compararemos si la media de dolor de hombres y mujeres es equivalente aplicando la segunda dosis. Consideraremos como intervalo de equivalencia propuesto en el capítulo 1. En nuestro caso particular da como resultado el intervalo de equivalencia $(-0.42, 0.42)$. La media y desviación típica muestral para mujeres es 0.515 y 0.710, respectivamente, con 56 datos y la media y desviación típica muestral para los hombres es de 0.512 y 0.706, respectivamente, con 26 datos.

En la figura 2.13 se observa que las medianas muestrales del dolor de hombres y mujeres son prácticamente iguales. Si aplicamos el test de Shapiro-Wilk a la muestra de hombres obtenemos un p-valor de 0.5552 y si lo aplicamos sobre la muestra de mujeres obtenemos un p-valor de 0.8818. Por tanto, no tenemos evidencias significativas de que las poblaciones no sigan una distribución normal.

Los test de Anderson-Hauck, Wellek y el basado en boostrap son capaces de encontrar evidencias de que las medias de ambas poblaciones son equivalentes con un nivel de significación de $\alpha = 5\,\%$. Sin embargo, el test TOST acepta la hipótesis nula de no equivalencia. Esto concuerda con lo visto en la figura 2.3 y las simulaciones donde vimos que el test TOST necesita tener un número mínimo de datos para tener potencia.

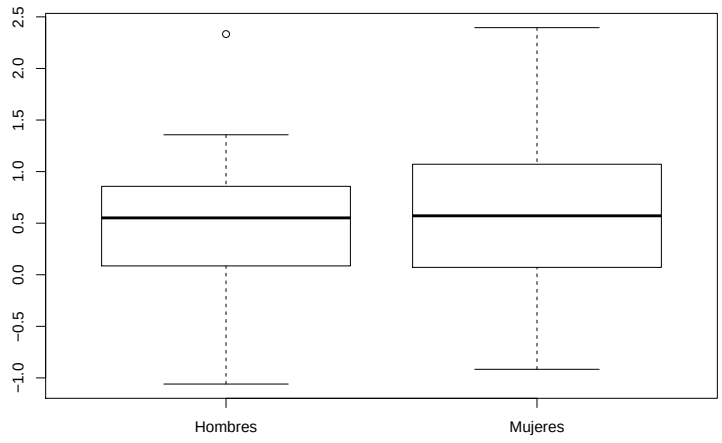

Figura 2.13: Boxplot de la media de dolor de hombres y de mujeres bajo la segunda dosis

Test	Estadístico/s	p-valor
TOST	$T_a = 0.599$ y $T_b = -0.633$	0.1094
Anderson-Hauck	$T_{AH} = -0.0167$ y $\hat{\delta} = 0.615$	0.0060
Wellek	$\hat{\gamma}^2 = 2.560 \cdot 10^{-3}$	0.0091
Bootstrap	$\hat{d^2} = 7.821 \cdot 10^{-6}$	0.0055

Tabla 2.9: Estadísticos y p-valores de los tests

Capítulo 3

Contrastes de equivalencia para dos modelos de regresión paramétricos

En el capítulo 2 introdujimos los contrastes de equivalencia para las medias de dos poblaciones. Ahora vamos a llevar este concepto a los modelos de regresión simple. En este caso se desea establecer la equivalencia entre dos modelos de regresión paramétricos que usaremos para describir la relación entre una variable respuesta numérica y una variable explicativa categórica de dos grupos diferentes. Una vez establecida esta equivalencia mostraremos varias simulaciones donde generalizaremos este tipo de contrastes para el uso de una variable explicativa numérica. Este tipo de contrastes fueron estudiados por Liu et al. (2009) y Gsteiger et al. (2011) obteniendo unos tests con baja potencia. Más adelante Dette et al. (2018) propusieron estimar directamente la distancia entre ambos modelos obteniendo así tests con una mayor potencia. En este trabajo revisaremos la metodología propuesta en Dette et al. (2018).

Estamos interesados en contrastar la hipótesis nula de que la distancia entre ambas curvas sea mayor o igual que una cantidad positiva prefijada de antemano ε frente a la alternativa de que sea menor. Es decir, siendo $d(m_1, m_2)$ la distancia entre las curvas de regresión m_1 y m_2 se desea contrastar la hipótesis nula

$$H_0 : d(m_1, m_2) \geq \varepsilon$$

frente a la alternativa

$$H_1 : d(m_1, m_2) < \varepsilon.$$

Como distancias consideraremos la distancia del supremo L^∞ y la distancia L^2.

Siguiendo el trabajo de Dette et al. (2018), consideramos dos posibles modelos $m_1(x, \beta_1)$ y $m_2(x, \beta_2)$ paramétricos para explicar la variable respuesta Y frente a una covariable de dos grupos diferentes $l = 1, 2$:

$$Y_{l,i,j} = m_l(x_{l,i}, \beta_l) + \eta_{l,i,j},\ j = 1, \ldots, n_{l,i},\ i = 1, \ldots, k_l,$$

donde $x_{l,i}$ denota el i-ésimo nivel a correspondiente al grupo l, β_l es el vector de los ρ_l parámetros de cada grupo, $n_{l,i}$ el número de elementos en el nivel i del grupo l y k_l el número de niveles en cada grupo. Además, denotaremos $\mathcal{X}$ como el intervalo donde se desea hacer la comparativa, n_l el número de elementos del grupo l y n como el total de elementos en ambos grupos. Por último consideraremos que los errores $\eta_{l,i,j}$ son independientes y cuya distribución es normal de media cero y varianza σ_l^2.

Antes de nada introduciremos algunas notaciones previas necesarias.

Sean $\varsigma_{l,i} > 0$ y $\lambda \in (1, \infty)$, entonces siendo $n = n_1 + n_2$ definimos

$$\varsigma_{l,i} = \lim_{n_l \longrightarrow \infty} \frac{n_{l,i}}{n_l},\ i = 1, \ldots, k_l,\ l = 1, 2.$$

y

$$\lambda = \lim_{n_1, n_2 \longrightarrow \infty} \frac{n}{n_1}.$$

Sea $\triangle(x, \beta_1, \beta_2)$ la verdadera diferencia entre m_1 y m_2, es decir,

$$\triangle(x, \beta_1, \beta_2) = m_1(x, \beta_1) - m_2(x, \beta_2).$$

Sea Σ_l una matriz de dimensión $\rho_1 \times \rho_2$ definida por

$$\Sigma_l = \frac{1}{\sigma_l^2} \sum_{i=1}^{k_l} \varsigma_{l,i} \left(\frac{\partial}{\partial b_l} m_l(x_{l,i}, b_l)|_{b_l=\beta_l} \right) \left(\frac{\partial}{\partial b_l} m_l(x_{l,i}, b_l)|_{b_l=\beta_l} \right)^t.$$

Además, supondremos que Σ_l es no singular.

El núcleo $k(x, y)$ se define como

$$\begin{aligned} k(x,y) \quad := \quad & \lambda \left(\frac{\partial}{\partial b_l} m_1(x, b_l)|_{b_l=\beta_1} \right)^t \Sigma_1^{-1} \left(\frac{\partial}{\partial b_l} m_1(y, b_l)|_{b_l=\beta_1} \right) \\ & + \frac{\lambda}{\lambda - 1} \left(\frac{\partial}{\partial b_l} m_2(x, b_l)|_{b_l=\beta_2} \right)^t \Sigma_2^{-1} \left(\frac{\partial}{\partial b_l} m_2(y, b_l)|_{b_l=\beta_2} \right). \end{aligned}$$

3.1. Tests basados en la distancia L^2

Sea d_2 la distancia en L^2 de las curvas m_1 y m_2 definida por

$$d_2 = \int_{\mathcal{X}} (m_1(x, \beta_1) - m_2(x, \beta_2))^2 dx.$$

Estamos interesados en contrastar

$$H_0 : d_2 \geq \varepsilon_2$$

frente a la alternativa

$$H_1 : d_2 < \varepsilon_2.$$

Un estimador natural de esta cantidad es

$$\hat{d}_2 := \int_{\mathcal{X}} (m_1(x, \hat{\beta}_1) - m_2(x, \hat{\beta}_2))^2 dx,$$

donde $\hat{\beta}_1$ y $\hat{\beta}_2$ son los estimadores por mínimos cuadrados ordinarios de β_1 y β_2, respectivamente. Por tanto, rechazaremos la hipótesis nula de que ambos modelos no son equivalentes cuando $\hat{d}_2 < c$, siendo c una constante fijada de antemano por el nivel de significación α, es decir, $\mathbb{P}_{H_0}\left(\hat{d}_2 < c\right) \approx \alpha$.

3.1.1. Test basado en la distribución asintótica

Para determinar esta constante haremos uso del teorema 1 expuesto en Dette et al. (2018). Este nos dice que si $d_2 \neq 0$ la distribución asintótica del estadístico es dada por

$$\sqrt{n} \frac{\hat{d}_2 - d_2}{\sigma_{d_2}} \xrightarrow{d} N(0, 1), \tag{3.1}$$

con

$$\sigma_{d_2}^2 = 4 \int_{\mathcal{X} \times \mathcal{X}} \triangle(x, \beta_1, \beta_2) \triangle (y, \beta_1, \beta_2) k(x, y) dx dy.$$

La varianza $\sigma_{d_2}^2$ se puede estimar por $\hat{\sigma}_{d_2}^2$ remplazando β_1, β_2 y $k(x, y)$ por sus estimadores. Como $\hat{\beta}_1$ y $\hat{\beta}_2$ son consistentes para los parámetros β_1 y β_2 se tiene que, por el teorema de Mann-Wald (Mann y Wald, 1943) que

$$\hat{\sigma}_{d_2}^2 \xrightarrow{\mathbb{P}} \sigma_{d_2}^2.$$

Ahora, utilizando 3.1 podemos obtener un test asintóticamente de nivel α. Por tanto, si $d = \varepsilon_2$

$$\alpha \cong \mathbb{P}\left(\sqrt{n} \frac{\hat{d}_2 - \varepsilon_2}{\sigma_{d_2}} < z_\alpha \right) = \mathbb{P}\left(\hat{d}_2 < \varepsilon_2 + z_\alpha \frac{\sigma_{d_2}}{\sqrt{n}} \right)$$

siendo z_α el cuantil α de una ditribución normal estándar.

Por último, teniendo en cuenta que $\hat{\sigma}_{d_2}$ es un estimador consistente de σ_{d_2} obtenemos un test asintóticamente de nivel α, por tanto, rechazaremos la hipótesis nula si

$$\hat{d}_2 < \varepsilon_2 + \frac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha, \tag{3.2}$$

siendo z_α el cuantil α de una ditribución normal estándar. El siguiente corolario del teorema 1 de Dette et al. (2018) expone que el test presentado es consistente y asintóticamente de nivel α.

Corolario 3.1. *El test expuesto en* (3.2) *es consistente y asintóticamente de nivel* α*. De manera más precisa se tiene que*

$$\mathbb{P}\left(\hat{d}_2 < \varepsilon_2 + \frac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha\right) \xrightarrow{n_1, n_2 \longrightarrow \infty} \begin{cases} 0 & si \quad d_2 > \varepsilon_2, \\ \alpha & si \quad d_2 = \varepsilon_2, \\ 1 & si \quad d_2 < \varepsilon_2. \end{cases}$$

Demostración.

Teniendo en cuenta que $\hat{\sigma}_{d_2}$ es un estimador consistente de σ_{d_2} y la convergencia dada en (3.1), el teorema de Slutsky garantiza que

$$\sqrt{n} \frac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} \xrightarrow{d} N(0,1).$$

Consideraremos los siguientes casos:

- Si $d_2 > \varepsilon_2$ tenemos que $\varepsilon_2 - d_2 < 0$ y por tanto

  $$\mathbb{P}\left(\hat{d}_2 < \varepsilon_2 + \tfrac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha\right) = \mathbb{P}\left(\sqrt{n}\tfrac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} < \sqrt{n}\tfrac{\varepsilon_2 + \frac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha - d}{\hat{\sigma}_{d_2}}\right) = \mathbb{P}\left(\sqrt{n}\tfrac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} < \sqrt{n}\tfrac{\varepsilon_2 - d_2}{\hat{\sigma}_{d_2}} + z_\alpha\right) \xrightarrow{n \to \infty} 0$$

  ya que $\lim\limits_{n \to \infty} \sqrt{n}\frac{\varepsilon_2 - d_2}{\hat{\sigma}_{d_2}}$ es divergente hacia $-\infty$.

- Si $d_2 = \varepsilon_2$ tenemos que

  $$\mathbb{P}\left(\hat{d}_2 < \varepsilon_2 + z_\alpha \frac{\hat{\sigma}_{d_2}}{\sqrt{n}}\right) = \mathbb{P}\left(\sqrt{n}\frac{\hat{d}_2 - \varepsilon_2}{\hat{\sigma}_{d_2}} < z_\alpha\right) = \mathbb{P}\left(\sqrt{n}\frac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} < z_\alpha\right) \xrightarrow{n \to \infty} \alpha.$$

- Si $d_2 < \varepsilon_2$ tenemos que $\varepsilon_2 - d_2 > 0$ y por tanto

  $$\mathbb{P}\left(\hat{d}_2 < \varepsilon_2 + \tfrac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha\right) = \mathbb{P}\left(\sqrt{n}\tfrac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} < \sqrt{n}\tfrac{\varepsilon_2 + \frac{\hat{\sigma}_{d_2}}{\sqrt{n}} z_\alpha - d}{\hat{\sigma}_{d_2}}\right) = \mathbb{P}\left(\sqrt{n}\tfrac{\hat{d}_2 - d_2}{\hat{\sigma}_{d_2}} < \sqrt{n}\tfrac{\varepsilon_2 - d_2}{\hat{\sigma}_{d_2}} + z_\alpha\right) \xrightarrow{n \to \infty} 1$$

  ya que $\lim\limits_{n \to \infty} \sqrt{n}\frac{\varepsilon_2 - d_2}{\hat{\sigma}_{d_2}}$ es divergente hacia $+\infty$.

□

3.1.2. Test basado en bootstrap

Cuando el tamaño muestral es pequeño la teoría asintótica podría dar lugar a un test demasiado conservador y una baja potencia. Por ello, vamos a presentar un test basado en el siguiente algoritmo bootstrap.

Algoritmo 3.2.

1. Calculamos los estimadores por mínimos cuadrados $\hat{\beta}_1$ y $\hat{\beta}_2$, los correspondiantes estimadores de las varianzas
$$\hat{\sigma}_l^2 = \frac{1}{n_l}\sum_{i=1}^{k_l}\sum_{j=1}^{n_{l,i}}(y_{l,i,j} - m_l(x_{l,i}, \hat{\beta}_l))^2,\ l = 1, 2,$$
y el estadístico
$$\hat{d}_2 = \int_{\mathcal{X}} (m_1(x, \hat{\beta}_1) - m_2(x, \hat{\beta}_2))^2 dx.$$

2. Definimos los estimadores de los parámetros β_1 y β_2 bajo la hipótesis nula como
$$\hat{\hat{\beta}}_l = \begin{cases} \hat{\beta}_l & si \quad \hat{d}_2 \geq \varepsilon_2, \\ \check{\beta}_l & si \quad \hat{d}_2 < \varepsilon_2. \end{cases} \quad l = 1, 2,$$
donde $\check{\beta}_1$ y $\check{\beta}_2$ denota los estimadores por mínimos cuadrados de los parámetros β_1 y β_2 bajo la siguiente restricción:
$$\int_{\mathcal{X}} (m_1(x, \beta_1) - m_2(x, \beta_2))^2 dx = \varepsilon_2.$$
Ahora estamos en condiciones de presentar el algoritmo bootstrap.

Algoritmo bootstrap

(i) Primero usaremos los parámetros calculados anteriormente para generar remuestras bootstrap bajo la hipótesis nula de la siguiente forma:
$$y^*_{l,i,j} = m_l(x_{l,i}, \hat{\hat{\beta}}_l) + \eta^*_{l,i,j},\ l = 1, 2$$
donde los términos de error $\eta^*_{1,i,j}$ y $\eta^*_{2,i,j}$ se simulan de manera independiente de distribuciones $N(0, \hat{\sigma}_1)$ y $N(0, \hat{\sigma}_2)$ respectivamente.

(ii) Calculamos por mínimos cuadrados los estimadores $\hat{\beta}^*_1$ y $\hat{\beta}^*_2$ usando las remuestras anteriores y calculamos el estadístico bootstrap
$$\hat{d}^*_2 = \int_{\mathcal{X}} (m_1(x, \hat{\beta}^*_1) - m_2(x, \hat{\beta}^*_2))^2 dx$$

Ahora, repetimos B veces los pasos (i) y (ii) para generar $\hat{d}^*_{2,1}, \ldots, \hat{d}^*_{2,B}$ estadísticos bootstrap. Por último, siendo $\hat{d}^*_{2,(1)} \leq \ldots \leq \hat{d}^*_{2,(B)}$ los estadísticos bootstrap ordenados, rechazaremos la hipótesis nula si $\hat{d}_2$ es menor que el elemento $\lfloor B\alpha \rfloor$ de los estadísticos bootstrap ordenados. Es decir, rechazaremos la hipótesis nula si
$$\hat{d}_2 < \hat{d}^*_{2,(\lfloor B\alpha \rfloor)}.$$

Siguiendo el teorema 2 expuesto en Dette et al. (2018) tenemos que el test bootstrap es consistente y asintóticamente de nivel α.

3.2. Tests basados en la distancia del supremo

Sea ahora d_∞ la distancia del supremo para las curvas m_1 y m_2 definida por
$$d_\infty = \max_{x \in \mathcal{X}} |m_1(x, \beta_1) - m_2(x, \beta_2)|.$$

Estamos interesados en contrastar la hipótesis nula
$$H_0 : d_\infty \geq \varepsilon_\infty \tag{3.3}$$

frente a la alternativa

$$H_1 : d_\infty < \varepsilon_\infty.$$

Un estimador natural de esta cantidad es

$$\hat{d}_\infty := \max_{x\in\mathcal{X}} |m_1(x,\hat{\beta}_1) - m_2(x,\hat{\beta}_2)|,$$

donde $\hat{\beta}_1$ y $\hat{\beta}_2$ son los estimadores por mínimos cuadrados ordinarios de β_1 y β_2 respectivamente. Por tanto, al igual que ocurría en el apartado 3.1, rechazaremos la hipótesis nula de que ambos modelos no son equivalentes cuando $\hat{d}_\infty < c$, siendo c es una constante fijada de antemano por el nivel de significación α, es decir,

$$\mathbb{P}_{H_0}\left(\hat{d}_\infty < c\right) \approx \alpha. \tag{3.4}$$

3.2.1. Test basado en la distribución asintótica

Antes de dar la distribución asintótica del estadístico $\hat{d}_\infty$ definimos el conjunto de puntos extremos

$$\mathcal{E} = \{x \in \mathcal{X} : \max_{x\in\mathcal{X}} |m_1(x,\beta_1) - m_2(x,\beta_2)| = d_\infty\}$$

y lo descomponemos en $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$, donde,

$$\mathcal{E}^- = \{x \in \mathcal{X} : \max_{x\in\mathcal{X}}(m_1(x,\beta_1) - m_2(x,\beta_2)) = -d_\infty\},$$
$$\mathcal{E}^+ = \{x \in \mathcal{X} : \max_{x\in\mathcal{X}}(m_1(x,\beta_1) - m_2(x,\beta_2)) = d_\infty\}.$$

Siguiendo el teorema 3 de Dette et al. (2018) se tiene que

$$\sqrt{n}\left(\hat{d}_\infty - d_\infty\right) \xrightarrow{d} \mathcal{Z} := \max\left\{\max_{x\in\mathcal{E}^+} G(x), \max_{x\in\mathcal{E}^-} -G(x)\right\},$$

donde $\{G(x)\}_{x\in\mathcal{X}}$ denota el proceso gaussiano definido por

$$G(x) = \left(\frac{\partial}{\partial b_l} m_1(x,b_1)|_{b_1=\beta_1}\right)^t \sqrt{\lambda}\Sigma_1^{-\frac{1}{2}} Z_1 - \left(\frac{\partial}{\partial b_2} m_2(x,b_l)|_{b_2=\beta_2}\right)^t \Sigma_2^{-\frac{1}{2}} \sqrt{\frac{\lambda}{\lambda-1}} Z_2,$$

donde Z_1 y Z_2 son normales estándar multivariantes independientes de dimensión ρ_1 y ρ_2 respectivamente. Por tanto, rechazaremos la hipótesis nula (3.3) si

$$\hat{d}_\infty < q_{\alpha,\infty}, \tag{3.5}$$

siendo $q_{\alpha,\infty}$ el cuantil α de la distribución $\mathcal{Z}$.

Sin embargo, dicha distribución tiene una estructura compleja. Por ejemplo, si $\mathcal{E} = \{x_0\}$ la distribución $\mathcal{Z}$ es una normal de media 0 y varianza

$$\begin{aligned}\sigma^2_{d_\infty} &= \lambda\left(\frac{\partial}{\partial b_l} m_1(x_0,b_l)|_{b_l=\beta_1}\right)^t \Sigma_1^{-1}\left(\frac{\partial}{\partial b_l} m_1(x_0,b_l)|_{b_l=\beta_1}\right)\\ &\quad + \frac{\lambda}{\lambda-1}\left(\frac{\partial}{\partial b_l} m_2(x_0,b_l)|_{b_l=\beta_2}\right)^t \Sigma_2^{-1}\left(\frac{\partial}{\partial b_l} m_2(x_0,b_l)|_{b_l=\beta_2}\right),\end{aligned}$$

la cual depende de la localización del punto extremo x_0.

En consecuencia, si el conjunto de puntos extremos $\mathcal{E}$ consiste en un único punto y siendo $\hat{\sigma}_{d_\infty}$ un estimador de $\sigma^2_{d_\infty}$ rechazaremos la hipótesis nula de no equivalencia entre las curvas de regresión si

$$\hat{d}_\infty < \varepsilon_\infty + \frac{\hat{\sigma}_{d_\infty}}{\sqrt{n}} z_\alpha.$$

Esto da lugar a un test consistente y asintóticamente de nivel α.

3.2.2. Test basado en bootstrap

Igual que ocurría con el contraste para la distancia d_2, cuando el número de datos sea reducido la teoría asintótica podría dar un test demasiado conservativo. Además, hemos visto la dificultad que tiene la aplicación de la teoría asintótica para aproximar la constante c de la ecuación (3.4). Por ello, igual que hacíamos con la distancia d_2, presentaremos un algoritmo bootstrap.

Algoritmo 3.3.

1. Calculamos los estimadores por mínimos cuadrados $\hat{\beta}_1$ y $\hat{\beta}_2$, los correspondientes estimadores de las varianzas
$$\hat{\sigma}_l^2 = \frac{1}{n_l}\sum_{i=1}^{k_l}\sum_{j=1}^{n_{l,i}}(y_{l,i,j} - m_l(x_{l,i}, \hat{\beta}_l))^2,\ l = 1, 2,$$
y el estadístico
$$\hat{d}_\infty = \max_{x\in\mathcal{X}} |m_1(x, \hat{\beta}_1) - m_2(x, \hat{\beta}_2)|$$

2. Definimos los estimadores de los parámetros β_1 y β_2 bajo la hipótesis nula como
$$\hat{\check{\beta}}_l = \begin{cases} \hat{\beta}_l & si \quad \hat{d}_\infty \geq \varepsilon_\infty, \\ \check{\beta}_l & si \quad \hat{d}_\infty < \varepsilon_\infty. \end{cases} \quad l = 1, 2,$$
donde $\check{\beta}_1$ y $\check{\beta}_2$ denota los estimadores por mínimos cuadrados de los parámetros β_1 y β_2 bajo la siguiente restricción:
$$\max_{x\in\mathcal{X}} |m_1(x, \beta_1) - m_2(x, \beta_2)| = \varepsilon_\infty. \tag{3.6}$$
Sin embargo, al contrario que ocurría en el algoritmo 3.2, el máximo no es una función diferenciable, por tanto, es necesario utilizar una aproximación derivable del máximo. Siendo
$$f_\varepsilon(x_1, \ldots, x_r) = \varepsilon \log\left(\sum_{i=1}^{r} \exp\left(\frac{x_i}{\varepsilon}\right)\right),\ \varepsilon > 0$$
se tiene que
$$\lim_{\varepsilon\to 0} f_\varepsilon(x_1, \ldots, x_r) = \max_{i=1,\ldots,r} x_i.$$
Ahora, escogemos una partición de la región de covariables $\mathcal{X}$ fijando r nodos $x_1, \ldots, x_r$, los cuales usaremos para aproximar el máximo de la diferencia en valor absoluto entre ambos modelos, es decir,
$$\max_{x\in\mathcal{X}} |m_1(x, \beta_1) - m_2(x, \beta_2)| \approx \max_{i\in\{1,\ldots r\}} |m_1(x_i, \beta_1) - m_2(x_i, \beta_2)| =$$
$$\lim_{\varepsilon\to 0} f_\varepsilon(m_1(x_1, \ldots, x_r, \beta_1) - m_2(x_1, \ldots, x_r, \beta_2)) = \lim_{\varepsilon\to 0} \varepsilon \log\left(\sum_{i=1}^{r} \exp\left(\frac{m_1(x_i,\beta_1) - m_2(x_i,\beta_2)}{\varepsilon}\right)\right).$$
En conclusión, substituiremos la restricción (3.6) por
$$\varepsilon \log\left(\sum_{i=1}^{r} \exp\left(\frac{m_1(x_i, \beta_1) - m_2(x_i, \beta_2)}{\varepsilon}\right)\right) = \varepsilon_\infty,$$
siendo ε un número lo suficientemente pequeño.

Ahora estamos en condiciones de presentar el algoritmo bootstrap.

Algoritmo bootstrap

(i) Primero usaremos los parámetros calculados anteriormente para generar remuestras bootstrap bajo la hipótesis nula de la siguiente forma:

$$y^*_{l,i,j} = m_l(x_{l,i}, \hat{\hat{\beta}}_l) + \eta^*_{l,i,j},\ l = 1, 2$$

donde los términos de error $\eta^*_{1,i,j}$ y $\eta^*_{2,i,j}$ se simulan de manera independiente de distribuciones $N(0, \hat{\sigma}_1)$ y $N(0, \hat{\sigma}_2)$ respectivamente.

(ii) Calculamos por mínimos cuadrados los estimadores $\hat{\beta}^*_1$ y $\hat{\beta}^*_2$ usando las remuestras anteriores y calculamos el estadístico bootstrap

$$\hat{d}^*_\infty = \max_{x \in \mathcal{X}} |m_1(x, \hat{\beta}^*_1) - m_2(x, \hat{\beta}^*_2)|.$$

Ahora, repetimos B veces los pasos (i) y (ii) para generar $\hat{d}^*_{\infty,1}, \ldots, \hat{d}^*_{\infty,B}$ estadísticos bootstrap. Por último, siendo $\hat{d}^{*1}_\infty \leq \ldots \leq \hat{d}^{*B}_\infty$ los estadísticos bootstrap ordenados, rechazaremos la hipótesis nula si $\hat{d}_\infty$ es menor que el elemento $\lfloor B\alpha \rfloor$ del estadístico bootstrap ordenado. Es decir, rechazaremos la hipótesis nula si

$$\hat{d}_\infty < \hat{d}^{*\lfloor B\alpha \rfloor}_\infty.$$

El teorema 4 expuesto en Dette et al. (2018) muestra que, siendo $\hat{q}_{\alpha,\infty}$ el cuantil α teórico de la distribución bootstrap, si el conjunto de puntos extremos $\mathcal{E}$ consiste en un único punto el test es consistente y de nivel α, cuando el cardinal de dicho conjunto sea mayor que uno el test controla el nivel α aunque se convierte en un test más conservador. De manera más concreta, se tiene que:

1. Si se cumple la hipótesis nula en (3.3) y el conjunto de puntos extremos $\mathcal{E}$ es un único punto, entonces para cualquier $\alpha \in (0, 0.5)$ se satisface que

$$\lim_{n_1, n_2 \longrightarrow \infty} \mathbb{P}\left(\hat{d}_\infty < \hat{q}_{\alpha,\infty}\right) = \begin{cases} 0 & si \quad d_\infty > \varepsilon_\infty, \\ \alpha & si \quad d_\infty = \varepsilon_\infty. \end{cases}$$

2. Sean $F_{\mathcal{Z}}$ y $q_{\mathcal{Z},\alpha}$ la función de distribución y el cuantil α de la variable $\mathcal{Z}$ respectivamente. Asumiendo que $F_{\mathcal{Z}}$ es continua en $q_{\mathcal{Z},\alpha}$ y que $q_{\mathcal{Z},\alpha} < 0$, si se satisface la hipótesis nula en (3.3) se tiene que

$$\lim \sup_{n_1, n_2 \longrightarrow \infty} \mathbb{P}\left(\hat{d}_\infty < \hat{q}_{\alpha,\infty}\right) \leq \alpha$$

3. Si se satisface la hipótesis alternativa en (3.3) entonces para cualquier $\alpha \in (0, 0.5)$

$$\lim_{n_1, n_2 \longrightarrow \infty} \mathbb{P}\left(\hat{d}_\infty < \hat{q}_{\alpha,\infty}\right) = 1.$$

Este cuantil $\hat{q}_{\alpha,\infty}$ teórico bootstrap lo aproximaremos por el cuantil empírico boostrap.

3.3. Simulación

En este apartado vamos a comparar por Monte Carlo los tests basados en la teoría asintótica y los basados en el bootstrap para las distancias d_2 y d_∞. Para ello se simularán muestras de modelos lineales y cuadráticos bajo las hipótesis nula y alternativa lo cual nos permitirá comparar la significación y la potencia con un nivel de significación de $\alpha = 0.05$. La variable explicativa X tomará valores en el conjunto $\{0, 0.25, 0.5, 0.75, 1\}$.

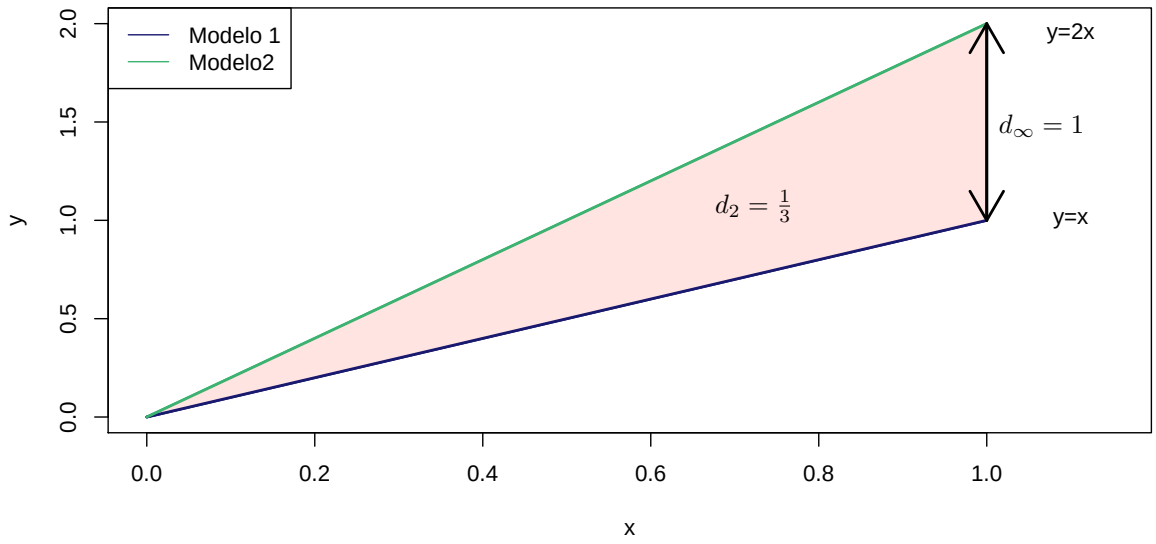

Figura 3.1: Dos modelos lineales

3.3.1. Dos modelos lineales

Comenzaremos con dos modelos lineales tomando como puntos de equivalencia $\varepsilon_2 = \frac{1}{3}$ y $\varepsilon_\infty = 1$.

$$\text{Modelo 1: } Y_{1i} = \beta_{10} + \beta_{11} X_{1i} + \xi_{1i} \quad i = 1, \ldots, n_1$$

$$\text{Modelo 2: } Y_{1j} = \beta_{20} + \beta_{21} X_{2j} + \xi_{2j} \quad j = 1, \ldots, n_2$$

Significación

Comenzaremos aproximando la significación, es decir, aceptar la hipótesis nula cuando esta es verdadera. Para realizar esta tarea simularemos 1000 muestras de dos modelos lineales con tamaños muestrales 15, 30 y 50.

En el primer modelo tomaremos $\beta_{10} = 0$, $\beta_{11} = 1$ y $\xi_1 \sim N(0, 1)$, en el segundo modelo tomaremos $\beta_{20} = 0$, $\beta_{21} = 2$ y $\xi_2 \sim N(0, 1)$. Por tanto los modelos quedarían de la siguiente forma.

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} \qquad i = 1, \ldots, n_1 \\ &\text{Modelo 2: } Y_{1j} = 2X_{2j} + \xi_{2j} \quad j = 1, \ldots, n_2 \end{aligned} \tag{3.7}$$

En la figura 3.1 podemos ver el primer modelo en color azul y el segundo modelo en color verde. En color rosa podemos ver el área entre ambos modelos de regresión y en negro su distancia máxima.

En la tabla 3.1 podemos observar que para la distancia d_2 el test asintótico tiene una proporción de rechazos superior al nivel de significación, por tanto en este caso es un test liberal. El test basado en el bootstrap tiene una proporción de rechazos similar al nivel de significación.

En la tabla 3.1 también podemos observar que para la distancia d_∞ el test asintótico tieneuna proporción de rechazos no alcanza el nivel de significación, por tanto en este caso es un test conservador. Sin embargo, el test basado el bootstrap igual que ocurría para la distancia d_2 tiene una proporción de rechazos similar al nivel de significación.

	$d_2 = \frac{1}{3}$			$d_\infty = 1$	
(n_1, n_2)	Asintótico	Bootstrap	(n_1, n_2)	Asintótico	Bootstrap
(15,15)	0.072	0.057	(15,15)	0.005	0.053
(15,30)	0.0.88	0.043	(15,30)	0.008	0.040
(30,15)	0.077	0.056	(30,15)	0.003	0.047
(15,50)	0.096	0.054	(15,50)	0.018	0.057
(50,15)	0.084	0.055	(50,15)	0.023	0.053
(30,30)	0.089	0.054	(30,30)	0.16	0.049
(30,50)	0.079	0.050	(30,50)	0.021	0.052
(50,30)	0.094	0.061	(50,30)	0.038	0.068
(50,50)	0.072	0.045	(50,50)	0.043	0.078

Tabla 3.1: Proporción de rechazos para los modelos expuestos en (3.7)

Potencia

Ahora aproximaremos la curva de potencia de los tests expuestos para las distancias d_2 y d_∞ . Para ello simularemos 5000 muestras con tamaños muestrales $n_1 = n_2 = 50$ bajo la hipótesis nula y diferentes alternativas. En el primer modelo tomaremos $\beta_{10} = 0$, $\beta_{11} = 1$ y $\xi_1 \sim N(0, 1)$, en el segundo modelo tomaremos $\beta_{20} = 0$, $\beta_{21} \in [0, 2]$ y $\xi_2 \sim N(0, 1)$. Es decir, simularemos muestras de los siguientes modelos

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} \qquad i = 1, \ldots, n_1, \\ &\text{Modelo 2: } Y_{1j} = \delta X_{2j} + \xi_{2j} \quad j = 1, \ldots, n_2, \end{aligned} \tag{3.8}$$

donde $\delta \in [0, 2]$.

En la figura 3.2 podemos ver el primer modelo en color negro y el segundo modelo en diferentes tonalidades de azul, según nos acerquemos a los extremos de la hipótesis nula.

En la figura 3.3 tenemos las curvas de potencia de los test expuestos para la distancia d_2 y en la figura 3.4 tenemos las curvas de potencia de los test para la distancia d_∞. En color negro se muestra la curva de potencia del test basado en la teoría asintótica y en color rojo la curva de potencia del test basado en el bootstrap.

En la figura 3.3 observamos el test asintótico tiene más potencia que el test boostrap. Sin embargo, ya hemos visto que el test basado en la teoría asintótica no respeta el nivel de significación de $\alpha = 0.05$.

En la figura 3.4 observamos que al contrario que en la figura 3.3 la curva de potencia del test basado en el bootstrap está claramente por encima de la curva de potencia de test basado en la teoría asintótica. Como era de suponer cuando el tamaño muestral es pequeño la teoría asintótica tiene menos potencia.

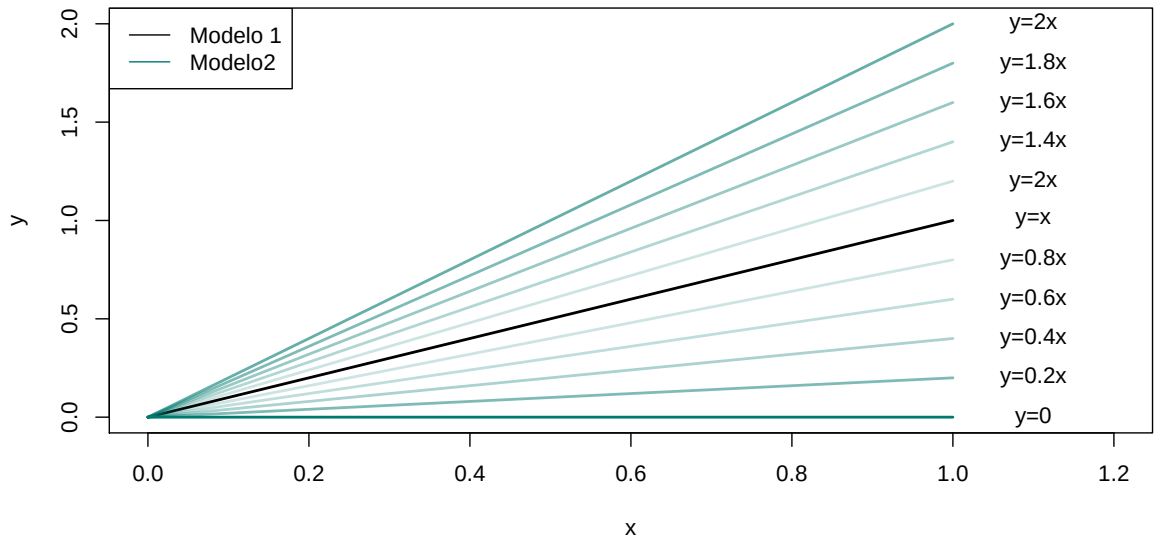

Figura 3.2: Modelos lineales utilizados en el estudio de la potencia

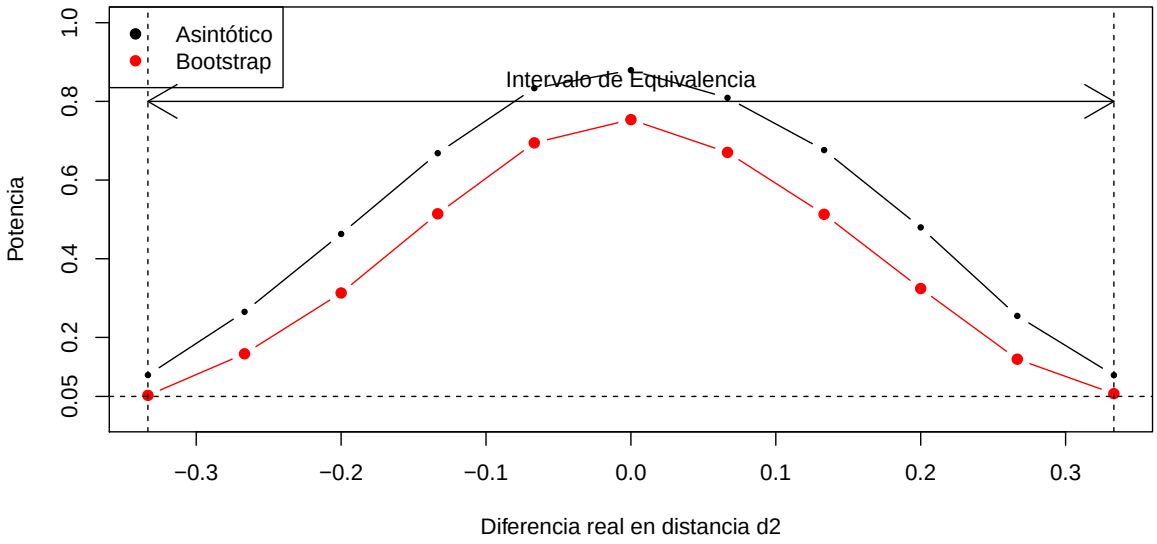

Figura 3.3: Aproximación de la curva de potencia de los modelos expuestos en (3.8) con la distancia d_2

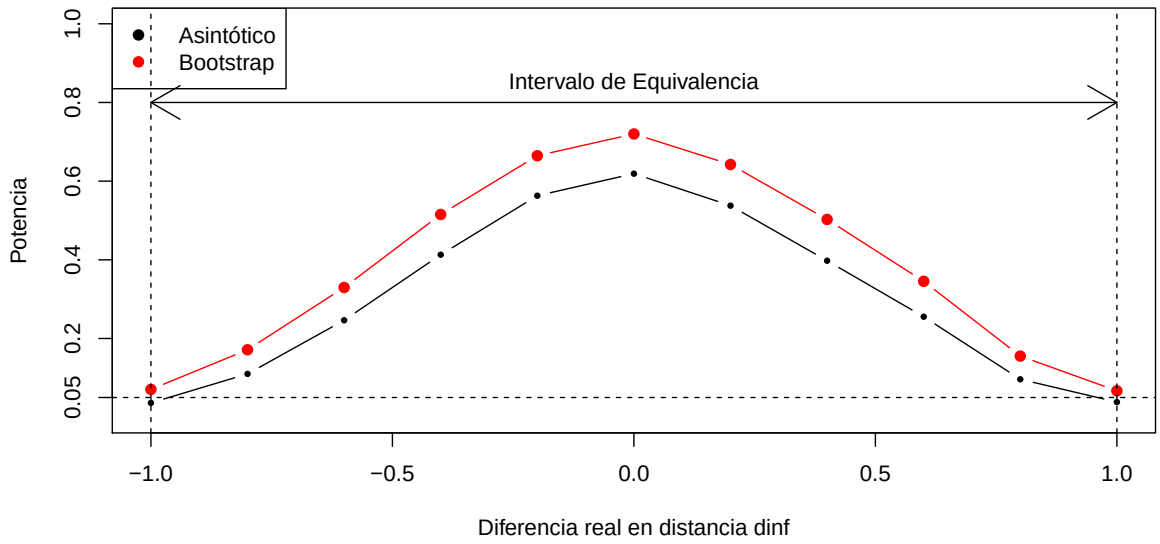

Figura 3.4: Aproximación de la curva de potencia de los modelos expuestos en (3.8) con la distancia d_∞

3.3.2. Un modelo lineal y uno cuadrático

A continuación simularemos un modelo lineal y uno cuadrático y tomaremos como puntos de equivalencia $\varepsilon_2 = \frac{1}{5}$ y $\varepsilon_\infty = 1$.

$$\text{Modelo 1: } Y_{1i} = \beta_{10} + \beta_{11} X_{1i} + \xi_{1i} \qquad i = 1, \ldots, n_1$$

$$\text{Modelo 2: } Y_{1j} = \beta_{20} + \beta_{21} X_{2j} + \beta_{22} X_{2j}^2 + \xi_{2j} \quad j = 1, \ldots, n_2$$

Significación

En el primer modelo tomaremos $\beta_{10} = 0$, $\beta_{11} = 1$ y $\xi_1 \sim N(0,1)$, en el segundo modelo tomaremos $\beta_{20} = 0$, $\beta_{21} = 1$, $\beta_{22} = 1$ y $\xi_2 \sim N(0,1)$. Es decir,

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} && i = 1, \ldots, n_1, \\ &\text{Modelo 2: } Y_{1j} = X_{2j} + X_{2j}^2 + \xi_{2j} && j = 1, \ldots, n_2. \end{aligned} \tag{3.9}$$

En la figura 3.5 podemos ver el modelo lineal en color azul y el cuadrático en color verde. En color rosa podemos ver el area entre ambos modelos de regresión y en negro su distancia máxima.

En la tabla 3.2 se puede ver que para la distancia d_2 el test asintótico es conservador cuando el tamaño muestral es pequeño, sin embargo, en el momento en que los tamaños muestrales aumentan, es decir, $n_1 = n_2 = 50$ el test aproxima bien el nivel de significación. El test basado en el bootstrap tiene una proporción de rechazos similar al nivel de significación de $\alpha = 0.05$.

En la tabla 3.2 también podemos observar que para la distancia d_∞ el test asintótico es conservador para tamaños muestrales pequeños y aproxima mejor el nivel de significación conforme los tamaños muestrales aumentan. Por otro lado tenemos que el test basado en el bootstrap presenta una proporción de rechazos superior al nivel de significación.

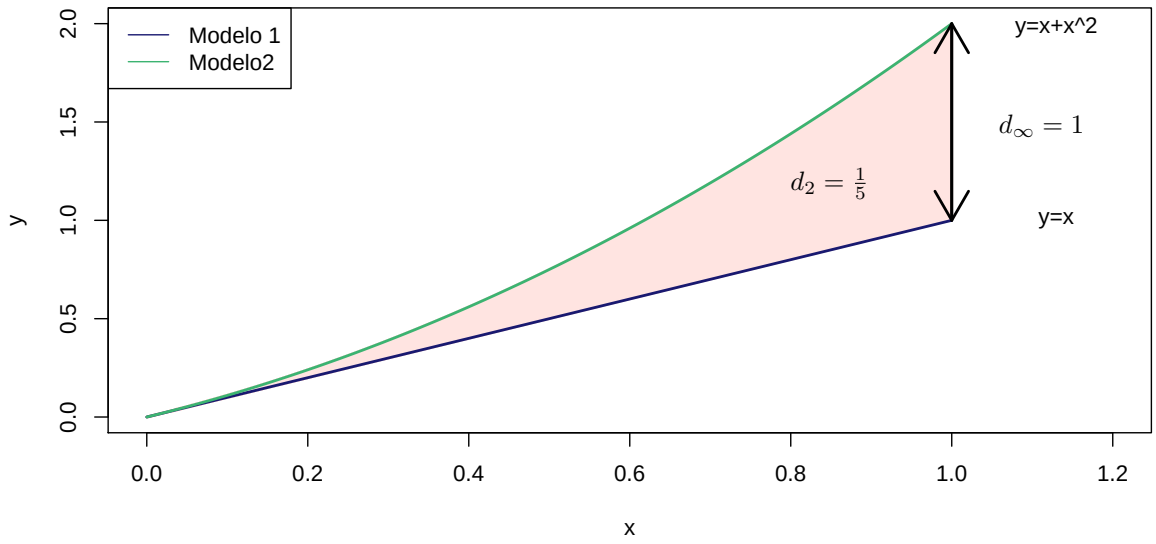

Figura 3.5: Un modelo lineal y otro cuadrático

	$d_2 = \frac{1}{5}$			$d_\infty = 1$	
(n_1, n_2)	Asintótico	Bootstrap	(n_1, n_2)	Asintótico	Bootstrap
(15,15)	0.006	0.066	(15,15)	0.004	0.083
(15,30)	0.017	0.052	(15,30)	0.011	0.068
(30,15)	0.012	0.053	(30,15)	0.007	0.073
(15,50)	0.014	0.051	(15,50)	0.008	0.069
(50,15)	0.017	0.054	(50,15)	0.010	0.069
(30,30)	0.003	0.052	(30,30)	0.018	0.071
(30,50)	0.032	0.053	(30,50)	0.037	0.081
(50,30)	0.037	0.052	(50,30)	0.023	0.073
(50,50)	0.053	0.064	(50,50)	0.039	0.067

Tabla 3.2: Proporción de rechazos para los modelos expuestos en (3.9)

Potencia

Como hemos hecho para el caso de dos modelos lineales aproximaremos las curvas de potencia mediante la simulación de 5000 muestras con tamaños muestrales $n_1 = n_2 = 50$. En el primer modelo tomaremos $\beta_{10} = 0$, $\beta_{11} = 1$ y $\xi_1 \sim N(0,1)$, en el segundo modelo tomaremos $\beta_{20} = 0$, $\beta_{21} = 1$, $\beta_{22} \in [-1.2, 1.2]$ y $\xi_2 \sim N(0,1)$. Es decir, muestras de los siguientes modelos

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} && i = 1, \ldots, n_1, \\ &\text{Modelo 2: } Y_{1j} = X_{2j} + \delta X_{2j}^2 + \xi_{2j} && j = 1, \ldots, n_2, \end{aligned} \tag{3.10}$$

donde $\delta \in [-1, 1]$.

En la figura 3.6 podemos ver el modelo lineal en color negro y el cuadrático en diferentes tonalidades de azul, según nos acerquemos a los extremos de la hipótesis nula.

En la figura 3.7 tenemos las curvas de potencia de los test expuestos para la distancia d_2 y en la figura 3.8 tenemos las curvas de potencia de los test para la distancia d_∞. En color negro se muestra la curva de potencia del test basado en la teoría asintótica y en color rojo la curva de potencia del test basado en el bootstrap.

En la figura 3.7 observamos que las curvas de potencia son prácticamente iguales, por tanto para la distancia d_2 en este caso concreto con muestras de $n_1 = n_2 = 40$ se cumple la teoría asintótica.

En la figura 3.8 en contra de lo que ocurre en la figura 3.7 la curva de potencia del test basado en la aproximación bootstrap es superior a la curva de potencia del test basado en la teoría asintótica.

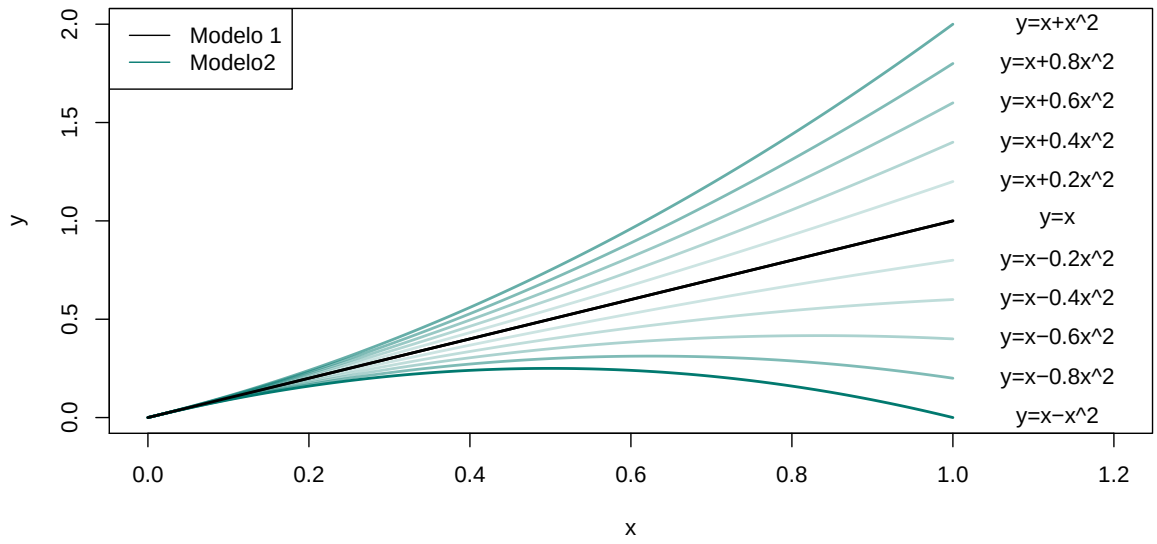

Figura 3.6: Un modelo lineal y varios cuadráticos utilizados en el estudio de la potencia

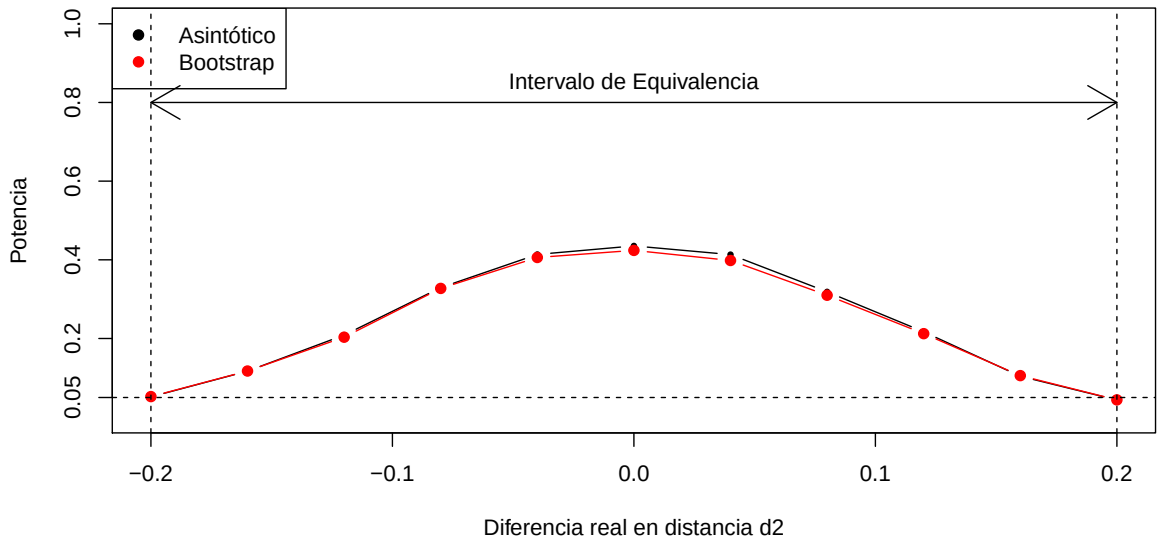

Figura 3.7: Aproximación de la curva de potencia de los modelos expuestos en (3.10) con la distancia d_2

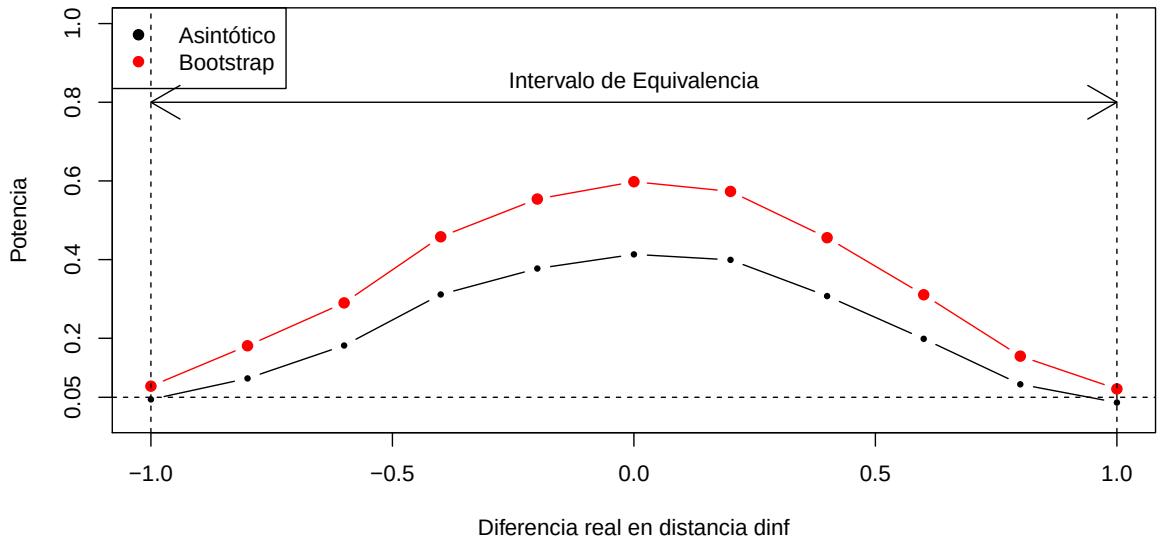

Figura 3.8: Aproximación de la curva de potencia de los modelos expuestos en (3.10) con la distancia d_∞

3.4. Diseño aleatorio

En los anteriores apartados hemos supuesto diseño fijo para la covariable y una proporción de observaciones balanceado entre los distintos valores de la variable explicativa. Esto resulta ser una imposición demasiado restrictiva y reduce la aplicación de los tests a casos reales. Por tanto, estudiaremos mediante simulaciones si es razonable eliminar esta restricción. Para esto, compararemos la significación cuando el tamaño muestral sea elevado y aproximaremos la curva de potencia de los modelos presentados en la sección 3.3 suponiendo que la variable X se distribuye de manera uniforme en el intervalo $[0, 1]$.

3.4.1. Dos modelos lineales

En este apartado consideraremos los mismos modelos y puntos de equivalencia expuestos en 3.3.1 para la significación y la potencia. Esto nos permitirá poder comparar la situación donde la variable explicativa X toma valores en el conjunto $\{0, 0.25, 0.5, 0.75, 1\}$ on en el intervalo $[0, 1]$.

Significación

Como hemos dicho antes consideramos los modelos

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} \qquad i = 1, \ldots, n_1, \\ &\text{Modelo 2: } Y_{1j} = 2X_{2j} + \xi_{2j} \quad j = 1, \ldots, n_2, \end{aligned} \tag{3.11}$$

e intentaremos imitar la situación $n_1, n_2 \to \infty$. Por tanto consideramos los tamaños muestrales 100, 1000, 10000 y 100000.

En la tabla 3.3 se observa que el test asintótico para la distancia d_2 presenta una proporción de rechazos superior al nivel de significación y el test asintótico para la distancia d_∞ inferior cuando $n_1 = n_2 = 100$. En los restantes casos todos los tests mantienen una proporción de rechazos similar al nivel de significación. Además, observamos que cuando aumenta el tamaño muestral la proporción de rechazos se mantiene entorno al nivel de significación de $\alpha = 0.05$. Por tanto, parece factible usar estos tests cuando el soporte de la variable explicativa sea no finito.

Potencia

Ahora consideramos los mismos escenarios para la potencia que en la sección 3.3 con los mismos tamaños muestrales.

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} \qquad i = 1, \ldots, n_1, \\ &\text{Modelo 2: } Y_{1j} = \delta X_{2j} + \xi_{2j} \quad j = 1, \ldots, n_2, \end{aligned} \tag{3.12}$$

donde $\delta \in [0, 2]$Además, compararemos la potencia de los tests cuando el soporte de la variable explicativa sea $\{0, 0.25, 0.5, 0.75, 1\}$ o el intervalo $[0, 1]$

En la figura 3.9 se muestras las curvas de potencia de los tests asintótico y bootstrap para la distancia d_2 de dos modelos lineales con tamaños muestrales $n_1 = n_2 = 50$. Observamos que las curvas de potencia son prácticamente iguales a las mostradas en la figura 3.3, la cual se corresponde con las curvas de potencia para la distancia d_2 cuando la variable explicativa toma valores en el conjunto $\{0, 0.25, 0.5, 0.75, 1\}$. Por tanto podemos sacar idénticas conclusiones, es decir, aunque la potencia del test asintótico sea superior a la del test bootstrap es preferible el test bootstrap ya que el test asintótico no respeta el nivel de significación.

En la figura 3.10 se muestran la curvas de potencia de los tests asintótico y bootstrap para la distancia d_∞ de dos modelos lineales con tamaños muestrales $n_1 = n_2 = 50$. En este caso, a diferencia de antes, observamos que la potencia de los tests cuando estamos bajo diseño aleatorio es menor para la distancia d_∞. Observamos que la potencia del test bootstrap es superior a la del test basado en la teoría asintótica.

	$d_2 = \frac{1}{3}$			$d_\infty = 1$	
(n_1, n_2)	Asintótico	Bootstrap	(n_1, n_2)	Asintótico	Bootstrap
(100,100)	0.76	0.052	(100,100)	0.39	0.053
(1000,1000)	0.056	0.053	(1000,1000)	0.52	0.051
(10000,10000)	0.055	0.056	(10000,10000)	0.56	0.052
(100000,100000)	0.048	0.053	(100000,100000)	0.51	0.049

Tabla 3.3: Proporción de rechazos para los modelos expuestos en (3.11)

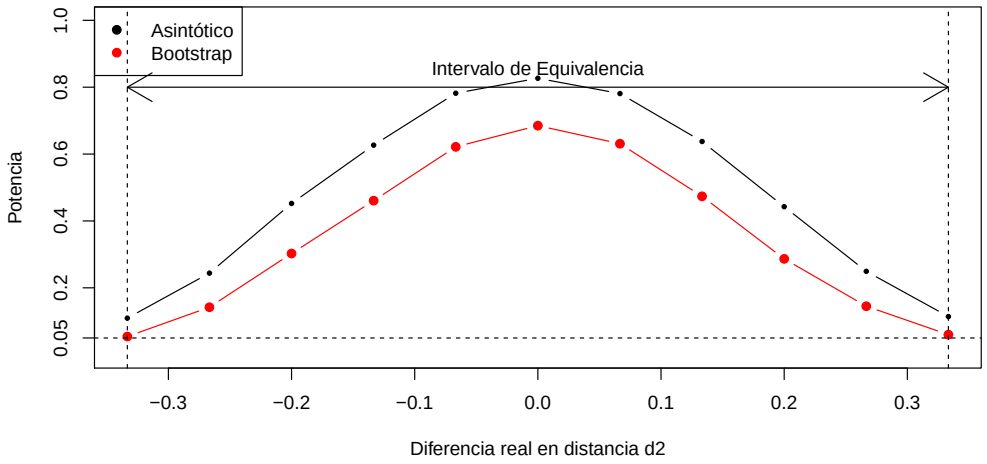

Figura 3.9: Aproximación de la curva de potencia de los modelos expuestos en (3.12) con la distancia d_2

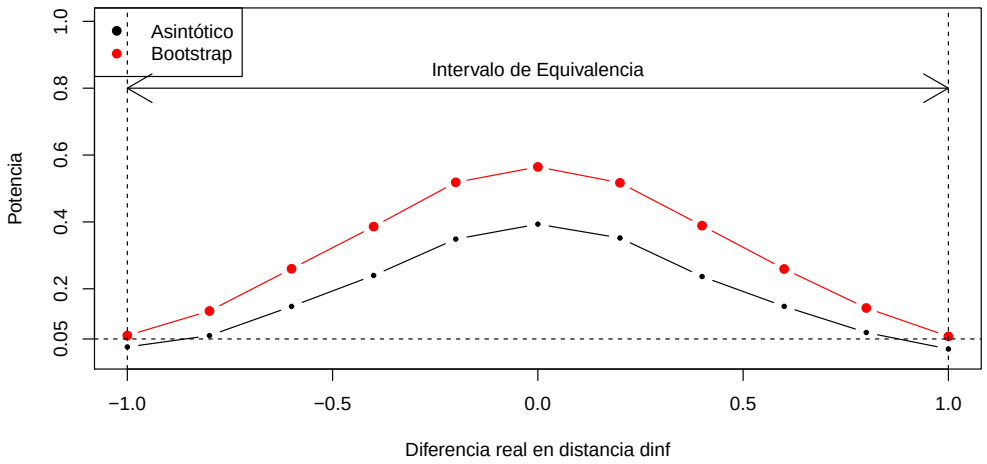

Figura 3.10: Aproximación de la curva de potencia de los modelos expuestos en (3.12) con la distancia d_∞

3.4.2. Un modelo lineal y uno cuadrático

Ahora consideraremos los modelos y puntos de equivalencia expuestos en 3.3.2 para la significación y la potencia.

Significación

Como ya hemos dicho antes consideraremos los siguientes modelos para la significación

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} && i = 1, \dots, n_1, \\ &\text{Modelo 2: } Y_{1j} = X_{2j} + X_{2j}^2 + \xi_{2j} && j = 1, \dots, n_2. \end{aligned} \tag{3.13}$$

Igual que en el caso de dos modelos lineales intentaremos emular la situación donde $n_1, n_2 \to \infty$ tomando tamaños muestrales de 100, 1000, 10000 y 100000.

En la tabla 3.4 se muestra la significación de un modelo lineal y uno cuadrático. En ella podemos ver que para el caso de $n_1 = n_2 = 100$ el test basado en la teoría asintótica para la distancia d_2 presenta una proporción de rechazos superior al nivel de significación y para la distancia d_∞ una proporción de rechazos inferior. En el resto de casos se respeta el nivel de significación y por tanto podemos decir que estos test son validos cuando el soporte de la variable sea no finito.

Potencia

Por último aproximaremos la curva de potencia de los tests de los siguientes modelos

$$\begin{aligned} &\text{Modelo 1: } Y_{1i} = X_{1i} + \xi_{1i} && i = 1, \dots, n_1, \\ &\text{Modelo 2: } Y_{1j} = X_{2j} + \delta X_{2j}^2 + \xi_{2j} && j = 1, \dots, n_2, \end{aligned} \tag{3.14}$$

donde $\delta \in [-1, 1]$.

En la figura 3.11 se muestran las curvas de potencia de los tests para la distancia d_2 con $n_1 = n_2 = 50$. Se observa que la curva de potencia del test basado en el bootstrap es ligeramente superior. Igual que ocurría en el caso de dos modelos lineales observamos que la potencia de los tests cuando estamos bajo diseño aleatorio es menor para la distancia d_∞.

En la figura 3.12 se muestran las curvas de potencia de los tests para la distancia d_∞ con $n_1 = n_2 = 50$. Se observa que la curva de potencia del test basado en el bootstrap es bastante superior. Igual que ocurría en el caso de dos modelos lineales, estas curvas de potencia son muy similares a las de la figura 3.8.

	$d_2 = \frac{1}{5}$			$d_\infty = 1$	
(n_1, n_2)	Asintótico	Bootstrap	(n_1, n_2)	Asintótico	Bootstrap
(100,100)	0.051	0.053	((100,100)	0.048	0.051
(1000,1000)	0.054	0.053	(1000,1000)	0.053	0.052
(10000,10000)	0.052	0.053	(10000,10000)	0.054	0.052
(100000,100000)	0.048	0.051	(100000,100000)	0.050	0.049

Tabla 3.4: Proporción de rechazos para los modelos expuestos en (3.13)

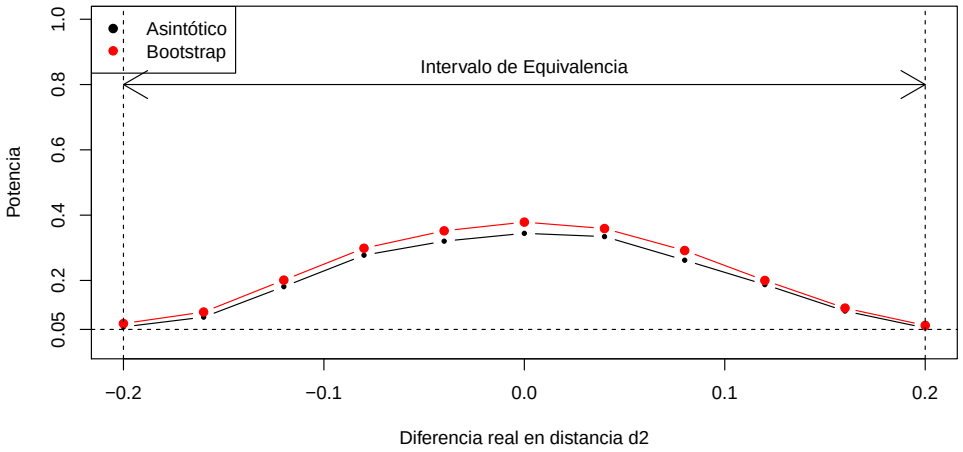

Figura 3.11: Aproximación de la curva de potencia de los modelos expuestos en (3.14) con la distancia d_2

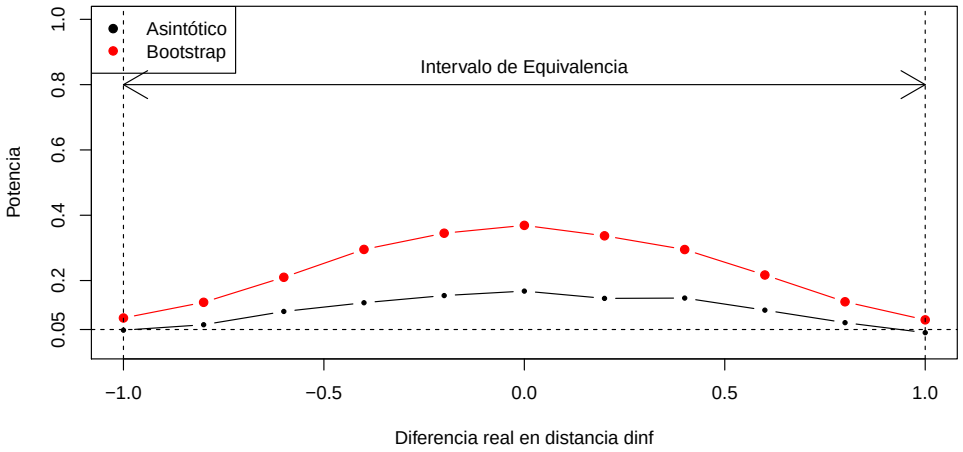

Figura 3.12: Aproximación de la curva de potencia de los modelos expuestos en (3.14) con la distancia d_∞

3.5. Aplicación a datos reales

En este apartado aplicaremos los diferentes tests para el contraste de equivalencia de dos modelos de regresión paramétricos al conjunto de datos expuesto en Pardo-Fernández et al. (2007), un estudio sobre los gastos mensuales de los hogares holandeses. Los datos están registrados en florines holandeses y corresponden al período comprendido entre octubre de 1986 y septiembre de 1987. Tomaremos como variable explicativa X el logaritmo del gasto mensual total y como variable respuesta Y el logaritmo del gasto mensual en alimentación.

Comparamos las curvas de regresión para tres grupos de hogares: hogares compuestos por dos miembros (159 en total), tres miembros (45 en total) y cuatro miembros (73 en total). Ajustaremos un modelo lineal para cada grupo y tomaremos como puntos de equivalencia $\varepsilon_2 = \frac{1}{8}$ y $\varepsilon_\infty = \frac{1}{4}$.

En la figura 3.13 se muestra el diagrama de dispersión del logaritmo de los gastos totales frente al logaritmo de los gatos en alimentación junto con las correspondientes rectas ajustadas. Los puntos en color negro representan a las familias formadas por 2 miembros, en color rojo las compuestas por 3 miembros y en color verde las formadas por 4 miembros.

En la tabla 3.5 se muestra la distancia d_2 entre los modelos lineales de 2, 3 y 4 miembros y los p-valores asociados a los tests asintótico y bootstrap para la distancia d_2. Observamos que con un nivel de significación de $\alpha = 0.05$ el test asintótico rechaza la hipótesis nula de no equivalencia para todos los modelos lineales. Sin embargo, el test asintótico solamente encuentra evidencias de que los modelos de 3 y 4 miembros son equivalentes. Esto puede deberse a que el test asintótico puede comportarse de manera liberal como ya hemos visto en 3.1 y 3.3.

En la tabla 3.6 se muestra lo mismo que en la tabla 3.5 para la distancia d_∞. Se observa que no se encuentran evidencias significativas para ninguno de los modelos con un nivel de significación de $\alpha = 0.05$.

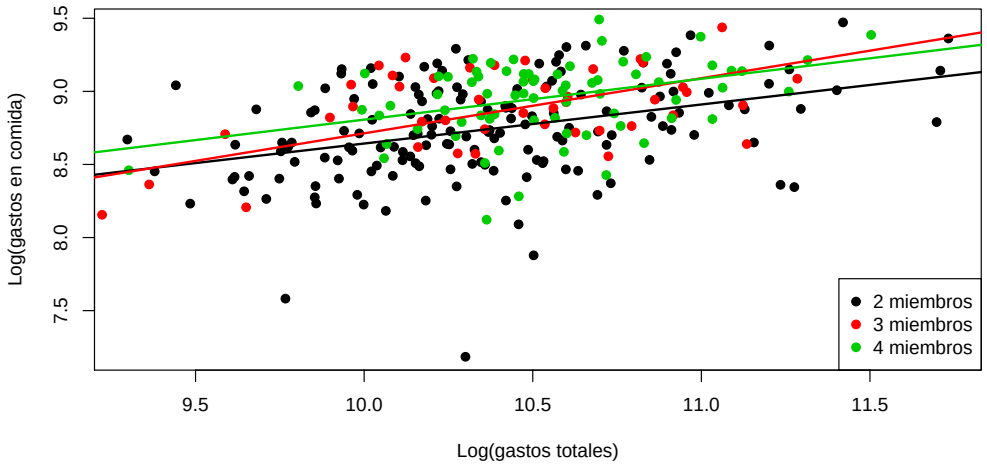

Figura 3.13: Diagrama de dispersión de los datos

Miembros	d_2	p-valor asintótico	p-valor bootstrap
2 y 3	0.053	0.010	0.092
2 y 4	0.070	0.037	0.052
3 y 4	0.018	$6 \cdot 10^{-5}$	0.025

Tabla 3.5: p-valores para la distancia d_2

Miembros	d_∞	p-valor asintótico	p-valor bootstrap
2 y 3	0.260	0.528	0.488
2 y 4	0.185	0.305	0.083
3 y 4	0.175	0.319	0.249

Tabla 3.6: p-valores para la distancia d_∞

Apéndice A

Distribuciones no centrales

A.1. Distribución t de Student no central

La distribución t de Student no central con ν grados de libertad y parámetro de no centralidad δ, que denotaremos por $t'_\nu(\delta)$, se define como el cociente de una variable normal estándar Z desplazada una cantidad δ sobre la raíz cuadrada de una variable chi-cuadrado V con ν grados de libertad entre sus grados de libertad ν. Es decir,

$$\frac{Z+\delta}{\sqrt{\frac{V}{\nu}}} \sim t'_\nu(\delta).$$

A.2. Distribución chi-cuadrado no central

La distribución chi-cuadrado no central con k grados de libertad y parámetro de no centralidad ψ, que denotaremos por ${\chi'}^2_k(\psi)$, se define como la suma de k variables aleatorias normales de media μ_i y varianza uno. Es decir, sean $X_1 \ldots, X_k$ variables aleatorias normales independientes de medias $\mu_1, \ldots, \mu_k$ respectivamente y varianza uno, entonces la variable aleatoria

$$\sum_{i=1}^{k} X_i^2$$

sigue una distribución χ^2 no central con k grados de libertad y parámetro de no centralidad

$$\psi = \sum_{i=1}^{k} \mu_i^2.$$

A.3. Distribución F de Snedecor no central

La F de Snedecor no central se define como el cociente de una chi-cuadrado no central (A.2) de parámetro de no centralidad ψ, que definiremos por $F'_{\nu_1,\nu_2}(\psi)$, entre sus grados de libertad y una chi-cuadrado entre sus grados de libertad siendo estas independientes, es decir, si $U_1 \sim {\chi'}^2_{\nu_1}(\psi)$ y $U_2 \sim \chi^2_{\nu_2}$, entonces

$$\frac{\frac{U_1}{\nu_1}}{\frac{U_2}{\nu_2}} \sim F'_{\nu_1,\nu_2}(\psi).$$

Bibliografía

Anderson, S. y Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, 12:2663–2692.

Biesheuvel, E. y Hothorn, L. (2002). Many-to-one comparisons in stratified designs. *Biometrical Journal*, 44:101–116.

Center for Drug Evaluation and Research (2001). *Statistical approaches to establishing bioequivalence.*

Dannenberg, O., Dette, H., y Munk, A. (1994). An extension of Welch's approximate t-solution to comparative bioequivalence trials. *Biometrika*, 81:91–101.

Dette, H., Möllenhoff, K., Volgushev, S., y Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, 113:711–729.

Food and Drug Administration (2016). *Statistical review of BLA761054.*

Gsteiger, S., Bretz, F., y Liu, W. (2011). Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. *Journal of Biopharmaceutical Statistics*, 21:708–725.

Hollander, M., A. Wolfe, D., y Eric, C. (1999). *Nonparametric statistical methods.* John Wiley & Sons.

Liu, W., Bretz, F., Hayter, A. J., y Wynn, H. (2009). Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. *Biometrics*, 65:1279–87.

Luzar-Stiffler, V. y Stiffler, C. (2002). Equivalence testing the easy way. *Journal of Computing and Information Technology*, 10:103–108.

Mann, H. B. y Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14:217–226.

Mann, H. B. y Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60.

Martín Andrés, A. (1990). On testing for bioequivalence. *Biometrical Journal*, 32:125–126.

Ocaña Rebull, J., Sánchez Olavarría, M. P., Sánchez, A., y Carrasco Jordan, J. L. (2008). On equivalence and bioequivalence testing. *SORT: Statistics and Operations Research Transactions,*, 32:151–176.

Pardo-Fernández, J. C., Keilegom, I., y González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, 17:1115–1137.

Randles, R. H. y Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* John Wiley & Sons.

Rocke, D. M. (1984). On testing for bioequivalence. *Biometrics*, 40:225–230.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, 37:617.

Shen, M., Russek-Cohen, E., y Slud, E. (2015). Exact calculation of power and sample size in bioequivalence studies using two one-sided tests. *Pharmaceutical Statistics*, 14:95–101.

Student (1908). The probable error of a mean. *Biometrika*, 1:1–25.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Wellek, S. (1993). A log-rank test for equivalence of two survivor functions. *Biometrics*, 49:877–881.

Wellek, S. (1996). A new approach to equivalence assessment in standard comparative bioavailability trials by means of the Mann-Whitney statistic. *Biometrical Journal*, 38:695–710.

Wellek, S. (2002). *Testing statistical hypotheses of equivalence and noninferiority.* Chapman and Hall/CRC.

Westlake, W. J. (1981). Response to T.B.L. Kirkwood on bioequivalence testing - a need to rethink. *Biometrics*, 37:589–594.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 32:741–744.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.