Ciencia Digital

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

Verificación de supuestos en las pruebas de comparación de medias. Una revisión.

Verification of assumptions in the tests of comparison of means. A review.

Pablo Flores. [1] , Jordi Ocaña.[2] & Tania Sánchez[3]

Recibido: 13-07-2017 / Revisado: 07-09-2018 Aceptado: 05-10-2018/ Publicado: 01-11-2018

Abstract.

DOI: https://doi.org/10.33262/cienciadigital.v2i4.1..187

The Student's t test for the significance of differences between means requires the fulfillment of normality and homoscedasticity assumptions. This paper collects the results of researches showing that when these assumptions are verified by means of other traditional hypothesis tests (e.g., the F or the Levene's test for homoscedasticity; goodness of fit to normality tests…), there is a high risk of a type I error. In addition, these pretesting procedures present some theoretical difficulties. On the other hand, when these assumptions are verified through an equivalence approach, using an appropriate irrelevance interval, these risks are better controlled. As a consequence, the equivalence approach is recommended, instead of the traditional one, whose use is not advisable. Advances in this type of studies and suggestions of future research are presented.

Keywords: equivalence, pre-testing, assumptions, Type I Error, Simulation.

Resumen.

La validez de la prueba t de Student para determinar la existencia de diferencias significativas entre medias está limitada al cumplimiento de los supuestos de normalidad y homocedasticidad. El presente trabajo recopila los resultados de investigaciones que muestran que cuando estos supuestos se verifican mediante otros test de hipótesis tradicionales (por ejemplo, el test F o el test de Levene para

[1] Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Grupo de Investigación en Ciencia de Datos CIDED, Riobamba – Ecuador, p_flores@espoch.edu.ec
[2] Departament de Genética, Microbiologia i Estadística, Secció d' Estadística, Universitat de Barcelona, Facultat de Biologia, Barcelona – España, jocana@ub.edu
[3] Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Grupo de Investigación en Ciencia de Datos CIDED, Riobamba – Ecuador.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

la homoscedasticidad; tests de bondad de ajuste a una distribución normal…) existe un alto riesgo de cometer un error tipo I, además de la existencia de dificultades teóricas. Por otra parte cuando los supuestos se verifican mediante un enfoque de equivalencia con adecuados intervalos de irrelevancia, estos riesgos quedan mejor controlados, por lo que se recomienda el uso de este enfoque en lugar del tradicional, el cual se desaconseja. Se presentan los avances realizados en este tipo de estudios así como sugerencias de posibles desarrollos futuros en esta línea.

Palabras clave: equivalencia, pre-testear, supuestos, Error Tipo I, Simulación.

1. Introducción.

La prueba t de Student (Student, 1908) para probar posibles diferencias significativas entre dos medias poblacionales está sujeta a la verificación de los supuestos de normalidad y homocedasticidad. No es extraño encontrar en la mayoría de libros sobre estadística inferencial (Gutierrez, 2008; Triola, 2009; Wackerly, Mendenhall, & Scheaffer, 2010; Wallpole & Myers, 2012) la restricción del uso de esta prueba únicamente para muestras en las que se asume su cumplimiento. Comúnmente, estos supuestos son probados previamente a través de otros test estadísticos, a los cuales llamaremos pre-test. Cuando en un pre-test de normalidad (Shapiro-Wilk, Anderson Darling, Kolmogorov-Smirnov, etc… (Darling & Anderson, 1954; KOLMOGOROV, 1933; Shapiro & Wilk, 1965; Smirnov, 1948)) la hipótesis nula se rechaza, se suele utilizar un test no paramétrico (Wilcoxon, Mann-Whitney,… (Mann & Whitney, 1947; Wilcoxon, 1945)) asumiendo implícitamente robustez ante casos de no homocedasticidad, lo cual no necesariamente es un hecho que se pueda asegurar en estas pruebas. Por otra parte, cuando la mencionada hipótesis nula de perfecta normalidad no se rechaza se suele asumir que existe normalidad y a continuación se realiza sobre la misma muestra un segundo pre-test (F, Levene, Bartlet (Brown & Forsythe, 1974; Snedecor & Cochran, 1989)) para verificar el supuesto de homoscedasticidad. Cuando en este pre-test no existe evidencia para rechazar la hipótesis nula de perfecta igualdad entre varianzas, la t de Studentsuele ser utilizada, en caso contrario una modificación de esta prueba para datos heteroscedásticos (Test de Welch (Welch, 1951)) se puede considerar como la alternativa adecuada.

A pesar de que el proceso descrito (o alguna variante del mismo) se aplica comúnmente en el ámbito de la estadística inferencial, existen estudios (Albers, Boon, & Kallenberg, 1998; Hsu, 1938; Overall, John E and Atlas, Robert S and Gibson, 1995; Scheffé, 1970) que muestran que este proceso altera la probabilidad global de cometer un error de Tipo I (TIEP, por sus siglas en inglés), manteniéndola muy alejada del nivel de significancia planteado para la prueba de comparación de medias, especialmente para casos de muestras pequeñas y desbalanceadas. En ocasiones, en estos estudios los autores concluyen advirtiendo de los peligros de pre-testear y recomiendan evitar este proceso y

substituirlo por el uso directo del test de Welch (sin pre-testear), ya que parece ser que esta prueba mantiene estable la TIEP alrededor del nivel de significancia planteado en la prueba de comparación de medias. Al respecto, un estudio (Moder, Rasch, & Kubinger, 2009; Rasch, Kubinger, & Moder, 2011), que estima la TIEP global cuando ambos supuestos son pre-testeados a través de de Kolmogorov – Smirnov y Levene (que es un test que permanece robusto ante desviaciones de la normalidad (Rasch & Guiard, 2004)), usando muestras balanceadas, desbalanceadas y con diferentes niveles de no normalidad y heterocedasticidad, muestra que pre-testear conduce a graves riesgos desconocidos en las probabilidades de cometer un error de tipo I y tipo II, por lo que hacerlo no vale la pena. Aplicar directamente el t de Studento la prueba no paramétrica de Wilcoxon conduce incluso a peores alteraciones de la TIEP, en su lugar la aplicación directa del test de Welch parece ser la mejor opción, aunque se observa que esta deja de ser efectiva para muestras pequeñas y con un grado severo de no normalidad. Otro estudio (Zimmerman, 2004), donde únicamente se pre-testeó el supuesto de homocedasticidad mediante la prueba de Levene tuvo resultados similares, nuevamente el uso directo de Welch resultó ser una alternativa que mantiene controlada la TIEP independientemente del tamaño muestral y el grado de heteroscedasticidad de las muestras. Además se concluyó que, pre-testear, especialmente para muestras pequeñas y desbalanceadas altera significativamente la TIEP, aunque esta empieza a decrecer conforme el nivel de significancia del pre-test es mayor, pero esto a partir del valor poco práctico del nivel de significancia $\propto = 0.20$.

Aunque estos estudios preliminares muestran que aplicar directamente (sin pre-testear) el test de Welch es una buena opción y además más simple de implementar y aplicar que el proceso de pre-test (Rasch & Schott, 2018), un estudio realizado para comparar más de dos medias (Flores M, 2018) mediante la versión generalizada del test de Welch como alternativa del ANOVA de un factor para muestras heterocedásticas, revela que cuando se verifica el supuesto de homocedasticidad, el procedimiento deja de ser efectivo. Además se confirma que el proceso de pre-testear altera significativamente la TIEP de manera similar a lo que ocurría en el proceso de comparación de dos medias.

Parece ser que existe suficiente evidencia para pensar que pre-testear mediante el enfoque tradicional no es un buena idea. En su libro “Testing statistical hypotheses of equivalence and noninferiority”, (Wellek, 2010) plantea la posibilidad de aplicar pruebas de equivalencia como alternativa a los pre-tests antes citados. En ellas se contrastaría una hipótesis nula que afirma la existencia de un elevado grado de incumplimiento de la condición de validez bajo estudio (por ejemplo, homoscedasticidad) frente a una alternativa que afirma que este grado de incumplimiento, quizás no sea nulo pero es suficientemente pequeño, irrelevante, como para alterar de forma importante la validez de la prueba t de Student. (La idea de Wellek es de alcance más general, a otras condiciones de validez de técnicas estadísticas, pero aquí nos centramos en el problema de la comparación de medias.)

Parece ser que existe suficiente evidencia para pensar que pre-testear mediante el enfoque tradicional no es un buena idea. El presente artículo de revisión pretende recopilar los

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

fundamentos teóricos que sustentan la validez de un test tanto tradicional como de equivalencia en el proceso de verificación de supuestos, así como los resultados de investigaciones realizadas con el objetivo de determinar la efectividad de estos enfoques en el proceso de pre-testeo para pruebas de comparación de dos medias en el sentido de la medición de la probabilidad de cometer errores. En este sentido, la siguiente sección muestra una argumentación teórica del enfoque tradicional usado para pre-testear los supuestos de normalidad y homocedasticidad. La sección tres muestra un análisis similar pero para el enfoque de equivalencia. La sección cuatro muestra un resumen de los principales resultados de la estimación de la TIEP cuando se pre-testea los supuestos usando ambos enfoques y finalmente la sección cinco muestra algunas conclusiones generales y discusión sobre posibles investigaciones futuras.

1. Verificación de supuestos mediante el enfoque tradicional.

1.1.Respecto al supuesto normalidad

Refiriéndose a la normalidad, (G. E. Box, 1979) mencionó: "*en la vida real no existe una distribución perfectamente normal, sin embargo, con modelos, que se sabe que son falsos, a menudo se puede derivar resultados que coinciden, con una aproximación útil a los que se encuentran en el mundo real*". Entonces, de acuerdo a lo mencionado por Box, es evidente que los test orientados a probar una perfecta normalidad (como lo hacen los pre-test tradicionales) no tienen sentido, a lo mejor tienen algún tipo de utilidad pero no prueban algo real. En este sentido podríamos mencionar lo que el mismo autor concluyó de forma más general: "*Todos los modelos son erróneos pero algunos son útiles*" (G. Box & Drapper, 1987).

Siguiendo el enunciado de (G. E. Box, 1979), podemos deducir que lo importante no es determinar si una muestra proviene de una perfecta distribución normal –ya sabemos que no. En su lugar lo que verdaderamente importa es saber si la aproximación del modelo para verificar el supuesto es lo suficientemente buena como para ser considerada útil. En este sentido, el criterio de Cochran sugiere que un modelo de prueba de hipótesis puede ser considerado útil o preciso si la TIEP tiene una desviación respecto al nivel de significancia de máximo del 20% de su valor, esto es la TIEP de la prueba tiene que ubicarse dentro del intervalo $[\propto \pm 0.2 \propto]$ (Cochran, 1942). En esencia este mismo criterio es utilizado para definir la robustez de una prueba de hipótesis (Rasch & Guiard, 2004).

1.2.Respecto al supuesto de homocedasticidad.

En el enfoque tradicional, la hipótesis nula de una prueba de homocedasticidad establece perfecta igualdad de los parámetros comparados (H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$). Trasladando la idea de Box a este pre-test, podríamos también concluir que verificar perfecta igualdad de varianzas carece de sentido, ya que en la realidad esto tampoco existe. Sin embargo estableceremos un análisis adicional respecto a los test tradicionales para verificar este supuesto.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

El hecho es que cuando un investigador está probando homocedasticidad bajo el enfoque tradicional se enfrenta a la dificultad lógica de que no rechazar la hipótesis nula de perfecta igualdad de varianzas solo significa ausencia de evidencia para concluir diferencias significativas entre los parámetros comparados, lo cual no es necesariamente una prueba de la existencia de homocedasticidad, como en este mismo sentido (Altman, Douglas G and Bland, 1995) mencionó "Ausencia de evidencia no es evidencia de ausencia" y (Wellek, 2010) también al respecto dijo "*Una diferencia no significativa no debe ser confundida con una significativa homogeneidad*". La parte izquierda del Figura 1 ayuda a visualizar de mejor manera esta idea. Por otra parte, rechazar esta hipótesis nula podría estar sugiriendo tan solo una irrelevante o despreciable heterocedasticidad. En ambos casos basándonos en la decisión de rechazo o no rechazo, existe la duda de saber si las pruebas tradicionales están realmente probando lo que necesitamos saber previo a un test de comparación de medias.

Ya podemos notar con más claridad que además de los estudios técnicos realizados por simulación, existen argumentos teóricos que muestran las dificultades conceptuales y peligros que tenemos al pre-testear los supuestos de normalidad y homocedasticidad usando el enfoque tradicional. Es interesante a continuación analizar lo que sucede con el enfoque de equivalencia

2. Verificación de supuestos mediante el enfoque de equivalencia.

(Wellek, 2010) en su libro, usa el término de equivalencia como una forma dilatada de una relación de identidad entre los parámetros analizados y considera que esta dilatación en la hipótesis de equivalencia se induce al añadir en la hipótesis tradicional una zona de irrelevancia alrededor de la correspondiente región o punto en el espacio paramétrico que denota la igualdad perfecta de varianzas (o el perfecto ajuste a la normalidad). Esta zona de irrelevancia está limitada por los llamados límites de equivalencia cuyos valores son constantes positivas que deben ser asignadas a priori y sin mayor conocimiento de la muestra a analizar. Inverso al test de hipótesis tradicional donde en la hipótesis nula se especifica la igualdad de los parámetros comparados, este tipo de pruebas se plantean de tal forma que la hipótesis nula establece la no equivalencia, mientras que la alternativa establece la equivalencia. Este cambio de interés en la investigación conduce a diseñar un estudio que pretende demostrar ausencia de una diferencia relevante entre los efectos de dos o más tratamientos, es decir equivalencia (Flores, 2017). Además es muy importante notar que las pruebas de equivalencia no buscan probar normalidad u homocedasticidad perfecta, en lugar de esto, la intención es declarar el cumplimiento de los supuestos incluso para desviaciones que pueden ser consideradas como irrelevantes o despreciables. Parece ser que con estas precisiones, este enfoque supera las dificultades lógicas analizadas en la Sección 2.

Con el fin de ilustrar de forma adecuada el planteamiento de una prueba de irrelevancia, presentamos a continuación el correspondiente test para determinar una posible equivalencia entre las varianzas σ_1^2 y σ_2^2, lo cual se traduciría en homocedasticidad:

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

Para las hipótesis:

$H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq \omega_1^2 \, \Lambda \, \frac{\sigma_1^2}{\sigma_2^2} \geq \omega_2^2$ → No equivalencia (Diferencia relevantes de varianzas)

$H_1: \omega_1^2 < \frac{\sigma_1^2}{\sigma_2^2} < \omega_2^2$ → Equivalencia (Diferencias irrelevantes de varianzas)

Con límites de irrelevancia $(\omega_1^2 < 1 < \omega_2^2)$. Un test invariante uniformemente más potente (UMPI), es aquel cuya región crítica está dada por:

$$\{\tilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2, \omega_2^2) < Q < \tilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2, \omega_2^2)\}$$

Donde el estadístico de prueba Q viene dado por

$$Q = \frac{S_X^2}{S_Y^2} = \frac{(n_2-1)}{(n_1-1)} \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2}{\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}$$

Y los valores críticos $(\tilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2, \omega_2^2),\ \tilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2, \omega_2^2))$ están determinados al resolver el sistema de ecuaciones:

$$F_{n_1-1,n_2-1}\left(\frac{\tilde{C}_2}{\omega_1^2}\right) - F_{n_1-1,n_2-1}\left(\frac{\tilde{C}_1}{\omega_1^2}\right) = \alpha = F_{n_1-1,n_2-1}\left(\frac{\tilde{C}_2}{\omega_2^2}\right) - F_{n_1-1,n_2-1}\left(\frac{\tilde{C}_1}{\omega_2^2}\right)$$

Donde $F_{n_1-1,n_2-1}()$ corresponde a la función de distribución acumulada de una distribución F centrada, con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

Notemos que en esta ocasión, cuando se rechaza la hipótesis nula de no equivalencia podemos concluir afirmando la existencia de evidencia para inferir diferencias irrelevantes entre las varianzas comparadas, lo cual sí se puede considerar como homocedasticidad (desde el punto de vista de la equivalencia). Esto, desde el punto de vista teórico supera las dificultades lógicas tratadas en la Sección 2.2. La Figura 1 muestra claramente estas diferencias entre el enfoque tradicional y de equivalencia cuando se pretende verificar el cumplimiento del supuesto de homocedasticidad.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

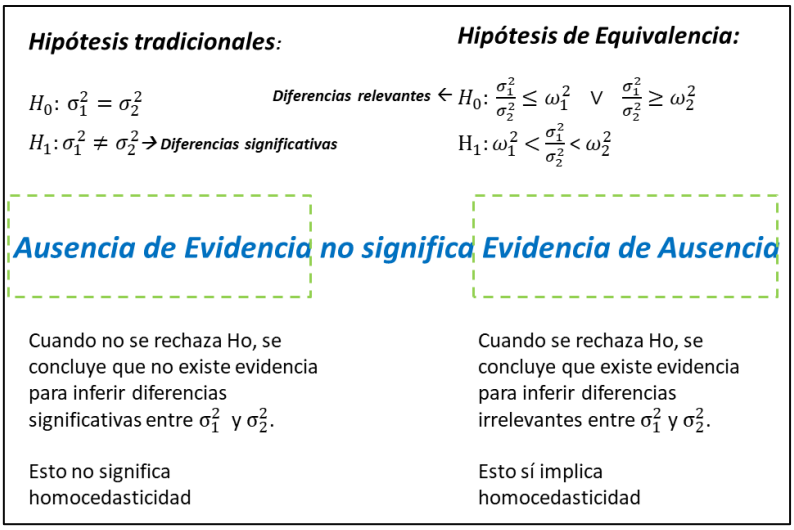

Figura 1. Comparación entre el enfoque tradicional y de equivalencia en las pruebas para verificar homocedasticidad

2.1.Los límites de equivalencia $(\omega_1^2,\ \omega_2^2)$

Con lo expuesto hasta ahora sobre pruebas de equivalencia para verificar el supuesto de homocedasticidad, estaremos conscientes de que el planteamiento de una adecuada zona de irrelevancia es fundamental para establecer una correcta equivalencia entre los parámetros a compararse. Sin embargo, Wellek no plantea un criterio técnico para determinar los límites de esta zona y en su lugar propone un criterio que nos parece subjetivo basado en la experiencia del investigador o en lo que él denomina como un sentido estadístico común.

(Flores M & Ocana, 2018), plantean un algoritmo computacional para determinar los límites de equivalencia en función del tamaño muestral, nivel de significancia y nivel de permisividad, el cual se muestra como un criterio más técnico al momento de establecer una adecuada zona de irrelevancia. El algoritmo se encuentra implementado en la función "limvarRatio" creada con el paquete base del software R (Team, 2016).

Para explicar la lógica de esta función es necesario definir el llamado nivel de permisividad "δ" como la distancia máxima por encima y por debajo del nivel de significancia "α" que se puede tolerar en una prueba de hipótesis para considerar que el modelo es lo suficientemente bueno como para ser considerado útil --lógicamente el mejor valor para esta permisividad basado en el criterio de Cochran será $\delta = 0.20\alpha$. Luego, debemos tener en cuenta que para una prueba t - Student de comparación de medias en muestras donde se garantice el cumplimiento de la normalidad y homocedasticidad, la TIEP coincide con el nivel de significancia α planteado y que cuando empiezan a existir desviaciones de la homocedasticidad, la TIEP de la prueba puede ser estimada mediante Simulación de Montecarlo como la proporción de rechazos de la hipótesis nula de igualdad de medias cuando esta es verdadera.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

Con base a estas precisiones, la función "limvarRatios" para determinar una adecuada zona de irrelevancia en las pruebas de equivalencia para verificar homocedasticidad, consiste en un algoritmo que busca de manera iterativa los valores más alejados entre σ_1 y σ_2 que producen muestras que al ser sometidas mediante un proceso de simulación a una t de Studentpara comparar medias provocan una estimación (por intervalos de confianza) de la TIEP que se encuentra dentro del intervalo de permisividad $(\alpha \pm \delta)$, pero no solo es necesario que se ubique dentro de este intervalo, además debe verificarse que se encuentre lo más cerca posible a uno de sus límites, esto debido a que para un δ fijo (lo mejor siempre será $\delta = 0.20\alpha$), a medida que aumenta el tamaño muestral, la zona de indiferencia debe ser más amplia, para que de esta forma la hipótesis de equivalencia tenga una región paramétrica lo más ancha posible, lo cual garantiza que la prueba de equivalencia no sea menos potente de lo que podría llegar a ser. Finalmente la razón al cuadrado entre los valores σ_1 y σ_2 hallados serán considerados los límites del intervalo de irrelevancia $(\omega_1^2,\ \omega_2^2)$. Ya que para casos balanceados existe una completa simetría entre los límites de equivalencia, es suficiente obtener una sola razón entre estos valores que determine un $\omega_1^2 = 1/\omega_2^2$, esto no se cumple en casos desbalanceados, donde se debe determinar dos razones ω_1^2 y ω_2^2 por encima y por debajo de la perfecta igualdad entre varianzas, las cuales no necesariamente son simétricas.

Utilizando un nivel de permisividad basado en el criterio de Cochran, la Tabla 1 muestra límites de irrelevancia calculados con la función "limvarRatios", a partir de 100000 réplicas de simulación, usando los niveles de significancia más comunes y para distintos tamaños muestrales tanto balanceados como desbalanceados. Los resultados muestran que el área de irrelevancia es más ancha para casos balanceados que para desbalanceados y además que para casos balanceados los tamaños muestrales más grandes corresponden a los intervalos de equivalencia más anchos, algo que no ocurre con los casos desbalanceados donde independientemente del tamaño muestral los límites de equivalencia son muy parecidos. Este comportamiento se mantiene para los distintos niveles de significancia utilizados.

Tabla 1: Zona de Indiferencia $(\omega_1^2,\ \omega_2^2)$ con $\delta = 0.20\alpha$

	$\boldsymbol{\alpha = 0.1}$	$\boldsymbol{\alpha = 0.05}$	$\boldsymbol{\alpha = 0.01}$
n = (5,5)	(0.130 - 7.691)	(0.225 - 4.428)	(0.397 - 2.519)
n = (3,7)	(0.709 - 1.412)	(0.779 - 1.289)	(0.819 - 1.163)
n = (7,3)	(0.711 - 1.410)	(0.776 - 1.325)	(0.832 - 1.166)
n = (10,10)	(0.002 - 501.0)	(0.097 - 10.325)	(0.282 - 3.542)
n = (6,14)	(0.727 - 1.408)	(0.783 - 1.292)	(0.846 - 1.157)
n = (14,6)	(0.716 - 1.362)	(0.787 - 1.264)	(0.859 - 1.148)
n = (5,10)	(0.679 - 1.387)	(0.741 - 1.286)	(0.819 - 1.196)
n = (10,5)	(0.716 - 1.452)	(0.786 - 1.331)	(0.862 - 1.256

3. Estimación de la tiep en el proceso de pre – testeo.

3.1.Para muestras normales cuando se pre-testea homocedasticidad

Un primer estudio de simulación usando ambos enfoques, cuando se asegura la normalidad y se pre-testea homocedasticidad (Flores M & Ocana, 2018) confirma los resultados de las investigaciones preliminares presentados en la introducción, en el sentido de obtener valores muy inflados de la TIEP estimada cuando el proceso de pre-testeo se lleva a cabo con el enfoque tradicional, especialmente para muestras pequeñas y desbalanceadas.

Por otro lado, cuando el enfoque de equivalencia con los adecuados límites de irrelevancia proporcionados por "limvarRatios" es usado para pre-testear homocedasticidad, la TIEP de una prueba de comparación de medias queda controlada en todos los casos alrededor del nivel de significancia α, siempre y cuando se asegure el cumplimiento de la normalidad e independientemente del nivel de significancia, el tamaño de la muestra y el nivel de heterocedasticidad. Más aún, en todos estos casos la TIEP estimada siempre se encuentra dentro del intervalo $(\alpha \pm \delta)$, establecido por el Criterio de Cochran, lo cual indica que todos estos procedimientos son los suficientemente buenos y seguros para ser utilizados.

Esta TIEP estimada siempre presenta valores muy similares (o casi idénticos) a los calculados cuando se usa directamente el test de Welch, esto resulta ser favorable, puesto que lo que está ocurriendo es que la prueba de homocedasticidad funciona de tal forma que declara heterocedasticidad en todos los casos que debe hacerlo y es por eso que siempre se aplica el test que corresponde para muestras heterocedásticas, es decir Welch.

Es importante mencionar que en todos las estimaciones realizadas se utilizó la técnica de reducción de la varianza denominada "Variables de control" (Ocaña & Vegas, 1995; Vegas & Ocana, 2000), la cual permitió reducir la variabilidad y tener una mejor precisión de la simulación sin necesidad de incrementar exageradamente las réplicas de simulación que para todos los casos fueron 100000. La variable de control utilizada es el número de rechazos de la hipótesis nula de una prueba t de Studentpara comparar dos medias cuando se asegura normalidad y homocedasticidad, la cual se encuentra altamente correlacionada con la variable de interés y además se conoce su esperanza matemática ya que coincide con el nivel de significancia α. Los resultados de este trabajo se muestran en conjunto con los resultados del siguiente, el cual analiza la estimación de la TIEP en aquellos casos donde no se puede asegurar el cumplimiento de la normalidad.

Además de este trabajo mencionaremos un estudio de simulación realizado cuando se utiliza una prueba de equivalencia para demostrar homocedasticidad entre más de dos poblaciones previo a una prueba de comparación de medias de un factor – ANOVA (Kim & Cribbie, 2017). El estudio utiliza límites de irrelevancia arbitrarios, es decir propuestos por un simple sentido estadístico común. La conclusión principal de este trabajo es que cuando se pre-testea igualdad de varianzas mediante enfoque de irrelevancia, el test ANOVA tiene una TIEP menor a la provocada que cuando se pre-testea con el enfoque tradicional (Levene), a pesar de que para ciertos niveles de heterocedasticidad y tamaños muestrales la TIEP está muy alejada del nivel de significancia usado y fuera del intervalo

de Cochran, por lo que pensamos que en realidad, esta investigación no brinda una solución a la alteración del error tipo I cuando se pre-testea.

El uso válido del procedimiento tradicional ANOVA de muestras independientes requiere que las variaciones de población sean iguales. Investigaciones anteriores han investigado si las pruebas de homogeneidad de la varianza, como la prueba de Levene, son satisfactorias como guardianes para identificar cuándo utilizar o no el procedimiento ANOVA. Esta investigación se centra en una nueva prueba de homogeneidad de la varianza que incorpora un enfoque de prueba de equivalencia. En lugar de probar la hipótesis nula de que las varianzas son iguales contra una hipótesis alternativa de que las varianzas no son iguales, la prueba basada en equivalencia evalúa la hipótesis nula de que la diferencia en las varianzas queda fuera o en el límite de un intervalo predeterminado frente a una alternativa hipótesis de que la diferencia en las variaciones cae dentro del intervalo predeterminado. Por lo tanto, con el procedimiento basado en equivalencia, la hipótesis alternativa se alinea con la hipótesis de investigación (igualdad de varianza). Un estudio de simulación demostró que la prueba de homogeneidad de la varianza poblacional basada en la equivalencia es un mejor guardián para el ANOVA que la homogeneidad tradicional de las pruebas de varianza

3.2.Para separaciones de la normal cuando se pretestea homocedasticidad.

Otro trabajo en el cual se aplicó la misma metodología de simulación que el anterior se realizó (Flores & Ocaña, 2018), pero ahora se añade el estudio de la estimación de la TIEP para muestras que presentan ciertos alejamientos de la normal. Estos alejamientos fueron obtenidos a partir de los coeficientes de Fleishman (Fleishman, 1978). El procedimiento de Fleishman para obtener muestras no normales, considera que cualquier distribución para la cual los primeros cuatro momentos existen, se puede obtener a partir de la transformación $Z = a + bX + cX^2 + dX^3$, donde X es una variable normal estándar y Z es una variable con distribución desconocida y parámetros $(\mu = 0, \sigma^2 = 1, \gamma_1, \gamma_2)$, donde la simetría γ_1 y la curtosis γ_2 definen los grados de separación o contaminación de la normal. Finalmente una variable $Y = \mu + \sigma Z$ tiene distribución desconocida con parámetros $(\mu = 0, \sigma^2 = 1, \gamma_1, \gamma_2)$.

Se definió diferentes niveles de contaminación de la normalidad de acuerdo a distintas combinaciones de la simetría y curtosis estudiadas en investigaciones previas (Bendayan, Arnau, Blanca, & Bono, 2013; Blanca, Arnau, López, Bono, & Bendayan, 2013). Los coeficientes de Fleishman para estos niveles de no normalidad fueron calculados usando la función “fleishman.coef” del paquete “BinNonNor” (Inan & Hakan, 2018) . Todos estos resultados se muestran en la Tabla 2, a partir de los cuales se pudieron establecer y simular muestras no normales para realizar el proceso de estimación de la TIEP.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

Tabla 2: Coeficientes de Fleishman, Simetría y Curtosis para diferentes niveles de contaminación de la normalidad.

Nivel de Contaminación	Simetría γ_1	Curtosis γ_2	Coeficientes de Fleishman (a, b, c, d)
Cont. Cero	0	0	(0, 1, 0, 0)
Cont. Leve	0.25	0.7	(-0.037, 0.933, 0.037, 0.021)
Cont. Moderada	0.75	1	(-0.119, 0.956, 0.119, 0.0098)
Cont. Alta	1.3	2	(-0.249, 0.984, 0.249, -0.016)
Cont. Severa	2	6	(-0.314, 0.826, 0.314, 0.023)

Dado que los resultados del estudio anterior donde se aseguraba normalidad son parte de los resultados de este nuevo trabajo estamos analizando, se presentan a continuación gráficas condensadas de los resultados para ambas investigaciones, donde las estimaciones correspondientes al grado de contaminación cero coinciden exactamente con las obtenidas en el trabajo para muestras normales anterior. Todas las gráficas tienen un área sombreada representando el Criterio de Cochran, el procedimiento que deje su respectiva TIEP dentro del área será considerado como robusto o lo suficientemente bueno para ser utilizado. Como se puede observar, en todos los casos se realiza la estimación usando 5 procedimientos diferentes, dos de estos tienen que ver con el uso directo sin pre-test del t de Studenty el test de Welch, otros dos utilizan los test tradicionales de Levene y F respectivamente para pre-testear el supuesto de homocedasticidad previa la prueba de comparación de medias y finalmente el último procedimiento tiene que ver con pre-testear el supuesto a través del enfoque de equivalencia.

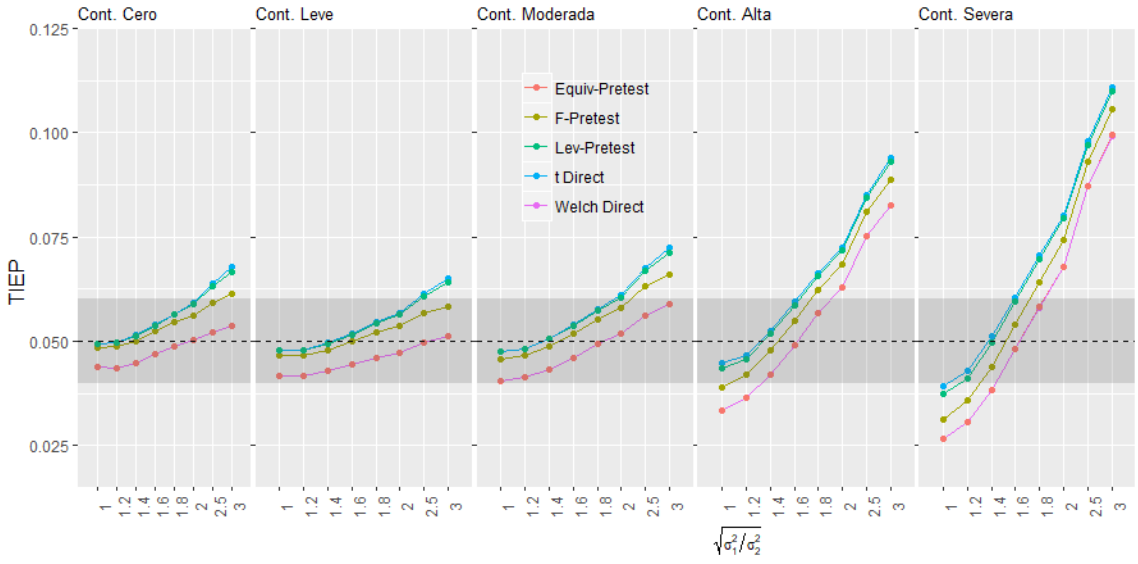

Figura 2. Estimación Global de la TIEP cuando se pre-testea igualdad de varianzas para distintos niveles de no normalidad y heterocedasticidad, usando un $\alpha = 0.05$ y $(n_1 = 5,\ n_2 = 5)$

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

La Figura 2 muestra la estimación de la TIEP para un caso balanceado de tamaño 5. Se puede observar que los diferentes procedimientos utilizados muestran un comportamiento similar, en el sentido que siempre uno es mejor que otro. En ningún caso aplicar directamente el t de Studento pre-testear bajo el enfoque tradicional es recomendado puesto que la TIEP de estos procedimientos se encuentran siempre alterados con mayor intensidad que los demás. La TIEP estimada, independientemente del nivel de heterocedasticidad, se controla bastante bien alrededor del nivel de significancia cuando se utiliza directamente el test de Welch o el pre-test de equivalencia, esto únicamente para niveles de contaminación de la normalidad cero, leve y moderado, donde la TIEP siempre se encuentra dentro del Criterio de Cochran para cualquier grado de heterocedasticidad. A partir de una contaminación alta de la normalidad parece ser que no podemos asegurar que ninguno de los procedimientos aplicado sea un método seguro a utilizar.

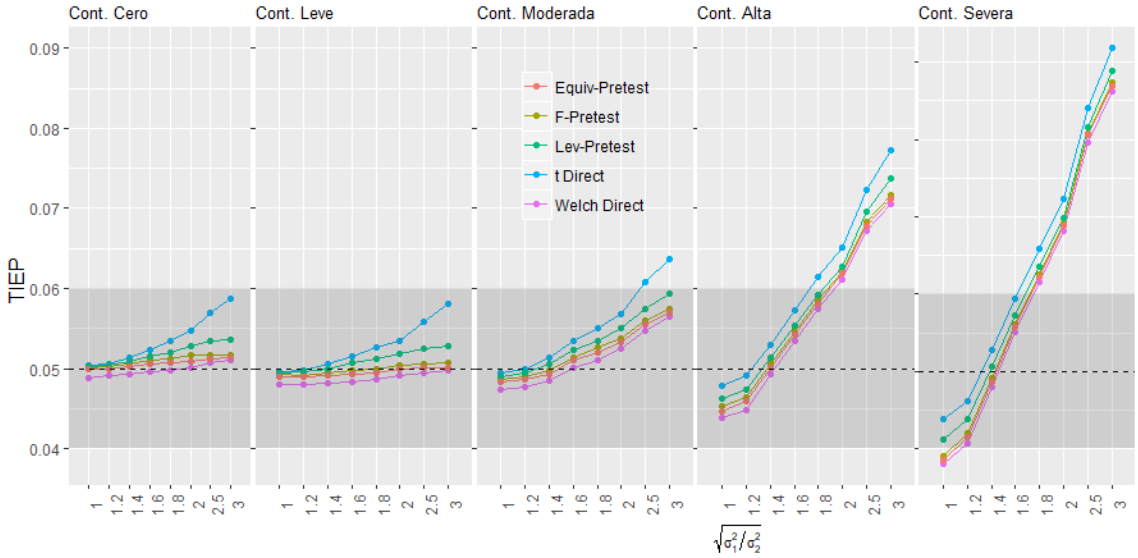

Figura 3. Estimación Global de la TIEP cuando se pre-testea igualdad de varianzas para distintos niveles de no normalidad y heterocedasticidad, usando un $\alpha = 0.05$ y $(n_1 = 10,\ n_2 = 10)$.

La Figura 3 muestra la estimación de la TIEP para un caso balanceado de tamaño 10. En general como casi todo lo que pasa en estadística cuando el tamaño muestral aumenta, estos resultados mejoran, siendo efectivos (desde el punto de vista del Criterio de Cochran) ciertos procedimientos que para muestras más pequeñas no funcionaban del todo bien. Sin embargo se observa que aunque existen procedimientos que se encuentran dentro del área sombreada, siempre el aplicar directamente Welch o pre - testear usando el enfoque de equivalencia mantiene la TIEP más cerca del nivel de significancia que los otros procedimientos. Nuevamente, esto parece ser adecuado pero solo hasta un nivel moderado de contaminación de la normalidad, luego de lo cual ya no existe un procedimiento que podamos recomendar.

ISSN: 2602-8085
Vol. 2, N°4.1, p. 5-22, octubre - diciembre, 2018

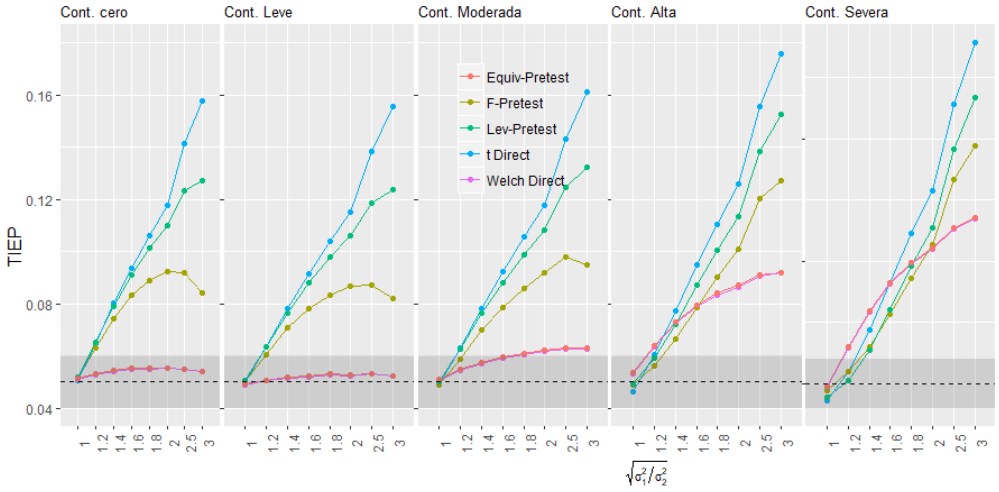

Figura 4. Estimación Global de la TIEP cuando se pre-testea igualdad de varianzas para distintos niveles de no normalidad y heterocedasticidad, usando un $\alpha = 0.05$ y $(n_1 = 5,\ n_2 = 10)$.

La Figura 4 y La Figura 5, muestran las estimaciones de la TIEP para muestras desbalanceadas de tamaño (5, 10) y (10, 5) respectivamente. Se observa que cuando a la muestra más pequeña le corresponde la varianza teórica más grande, la estimación de la TIEP se ubica por encima del nivel de significancia, mientras que cuando a la muestra más grande le corresponde la varianza más grande, la estimación de la TIEP se ubica por debajo del nivel de significancia. En estos casos desbalanceados, la alteración de la TIEP es más evidente, de tal forma que únicamente los procedimientos que consisten en usar directamente Welch o pre-testear mediante equivalencia mantienen a la TIEP dentro del área sombreada, todos los demás procedimientos presentan para alguna desviación de la normalidad o de la homocedasticidad alguna alteración de la TEP que se aleja demasiado del nivel de significancia. Pero de igual forma que lo ocurrido en los casos balanceados, el uso directo de Welch y pre-testear usando un test de equivalencia deja de ser efectivo a partir de un nivel alto de no normalidad.

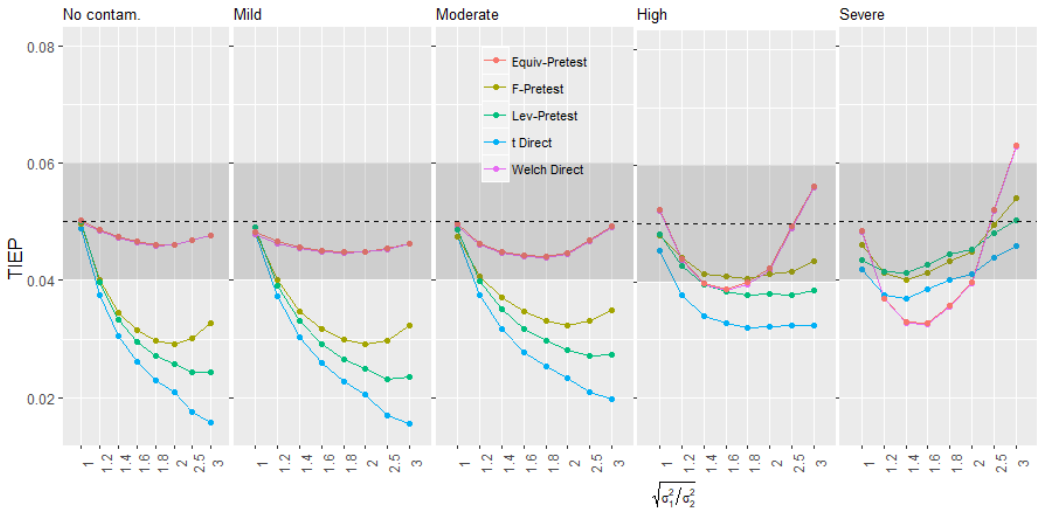

Figura 5. Estimación Global de la TIEP cuando se pre-testea igualdad de varianzas para distintos niveles de no normalidad y heterocedasticidad, usando un $\alpha = 0.05$ y $(n_1 = 10,\ n_2 = 5)$

4. Conclusiones.

- Está claro que pre-testear los supuestos de normalidad u homocedasticidad usando los enfoques tradicionales no es una buena idea, hacerlo aumenta considerablemente la probabilidad de cometer un error tipo I. Al igual que ya lo hacen investigaciones previas, recomendamos dejar de usar estas pruebas como un método de verificación de supuestos y en su lugar utilizar el enfoque de equivalencia, el cual provoca estimaciones aceptables de la TIEP al menos en su versión dedicada a pre-testear el supuesto de homocedasticidad y cuando se puede asegurar normalidad o a lo mucho una desviación moderada de la misma, que es lo que se ha investigado hasta ahora. Sería interesante en este sentido averiguar qué ocurre con la TIEP cuando la normalidad es pre-testeada ya sea con el enfoque tradicional o de equivalencia.
- Enfocándonos en el procedimiento que consiste en aplicar el pre – test de equivalencia, el hecho de que exista desviaciones de la normalidad para las cuales la TIEP empiece a alejarse demasiado del nivel de significancia y además saber que estas desviaciones pueden ser controladas por los coeficientes de Fleishman, propone investigaciones futuras como el planteamiento de niveles adecuados de irrelevancia para pruebas de equivalencia dedicadas a probar el supuesto de normalidad. El procedimiento podría consistir nuevamente en un algoritmo iterativo que encuentre valores para los coeficientes de Fleishman que hagan que la TIEP estimada se encuentre dentro y lo más cercano posible a los límites establecidos por el criterio de Cochran.
- Finalmente, muy poco se ha hecho respecto al estudio generalizado para más de dos medias, aunque ya se conoce que de manera similar a lo que sucede en el caso particular de dos poblaciones, el proceso de pre – testear bajo el enfoque tradicional deja serias alteraciones en la probabilidad de cometer un error de tipo I. En este sentido una investigación análoga a las presentadas en la presente revisión, dedicada a estudiar el comportamiento de la TIEP en pruebas de comparación de más de dos medias cuando se pre – testea sus supuestos mediante equivalencia sería de gran interés.

Bibliografía.

Albers, W., Boon, P. C., & Kallenberg, W. C. M. (1998). Testing equality of two normal means using a variance pre-test. *Statistic and Probability Letters*, *38*, 221–227.

Altman, Douglas G and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *Bmj*, *311*, 485.

Bendayan, R., Arnau, R. and, Blanca, M., & Bono, R. (2013). Comparison of the procedures of Fleishman and Ramberg et al. for generating non-normal data in simulation studies. *Annals of Psichology*, *30*(1), 364–371.

Blanca, M., Arnau, J., López, M., Bono, D., & Bendayan, R. (2013). Skewness and

kurtosis in real data samples. *European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78.

Box, G., & Drapper, N. (1987). *Empirical Model Building and response Surface*. *John Wiley & Sons.*

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Army Research Office Workshop on Robustness in Statistics*, *1*, 201–236. https://doi.org/0-12-4381 50-2

Brown, M., & Forsythe, A. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367.

Cochran, W. G. (1942). The x2 correction for continuity. *Iowa State College Journal of Science*, *16*, 421–436.

Darling, D. A., & Anderson, T. W. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49*(268), 765–769.

Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532.

Flores, P. (2017). *Un Pretest de Irrelevancia de la diferencia de varianzas en la comparación de medias*.

Flores, P., & Ocaña, J. (2018). Pretesting Assumptions for the validity of two sample Mean Tests. In *IX International Workshop on Simulation* (p. 28). Barcelona: UPC.

Flores M, P. (2018). El riesgo de pre-testear el supuesto de homocedasticidad en las pruebas de comparación de medias . Estudio para casos balanceados. *Revista Perspectiva*, *19*(1), 55–67.

Flores M, P., & Ocana, J. (2018). Heteroscedasticity irrelevance when testing means difference. *SORT*, *42*(1), 59–72. https://doi.org/10.2436/20.8080.02.69

Gutierrez, H. (2008). *Análisis y Diseño de Experimentos*. McGRAW-HILL/INTERAMERICANA EDITORES, S.A. de C.V.

Hsu, P. (1938). Contribution to the theory of" Student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs*.

Inan, G., & Hakan, D. (2018). BinNonNor: Data Generation with Binary and Continuous Non-Normal Components. Retrieved from https://cran.r-project.org/package=BinNonNor

Kim, J., & Cribbie, R. A. (2017). The variance homogeneity assumption and the traditional ANOVA: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology*.

KOLMOGOROV, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari*, *4*, 83–91.

Mann, H. B., & 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*, 50–60.

Moder, K., Rasch, D., & Kubinger, K. D. (2009). Don ' t use the two-sample t-test anymore ! In *VI Workshop on Simulation* (pp. 258–262). St. Petersburg.

Ocaña, J., & Vegas, E. (1995). Variance reduction for Bernoulli response variables in simulation. *Computational Statistics and Data Analysis*, *19*(6), 631–640. https://doi.org/10.1016/0167-9473(94)00023-C

Overall, John E and Atlas, Robert S and Gibson, J. M. (1995). Tests that are robust against variance heterogeneity in kx2 designs with unequal cell frequences. *Psychological Reports*, *76*, 1011–1017.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, *46*(2), 175–208.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*(1), 219–231. https://doi.org/10.1007/s00362-009-0224-x

Rasch, D., & Schott, D. (2018). *Mathematical Statistics*. Oxford, United Kingdom: Wiley.

Scheffé, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association*, *65*, 1501--1508.

Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281.

Snedecor, G., & Cochran, W. (1989). *Statistical Methods*.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25.

Team, R. C. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Triola, M. F. (2009). *Estadística* (10th ed.). México.

Vegas, E., & Ocana, J. (2000). Variance reduction in the study of a test concerning the Behrens-Fisher problem. *Communications in Statistics-Simulation and Computation*, *29*(2), 463–479. https://doi.org/10.1080/03610910008813622

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2010). *Estadística matemática con aplicaciones*. (S. . Cengage Learning Editores, Ed.) (10th ed.). Mexico.

Wallpole, R., & Myers, R. (2012). *Probabilidad y Estadística para ingeniería y*

Ciencias (9th ed.). México: Pearson Education.

Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*. https://doi.org/10.2307/2332579

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Zimmerman, D. (2004). A note on preliminary tests of equality of variances, *57*, 173–181. https://doi.org/10.1348/000711004849222

Para citar el artículo indexado.

Flores P., Ocaña J., Sánchez T., (2018). Verificación de Supuestos en las Pruebas de Comparación de medis. Una revisión. *Revista electrónica Ciencia Digital* ***2***(4.1.), 5-22. Recuperado desde: http://cienciadigital.org/revistacienciadigital2/index.php/CienciaDigital/article/view/187/165