El riesgo de pre-testear el supuesto de homocedasticidad en las pruebas de comparación de medias. Estudio para casos balanceados

The risk of pre-testing homoscedasticity assumption in the comparison means test. Study for balanced cases

Pablo Javier Flores Muñoz[1]

Resumen

Investigaciones muestran que pre-testear el supuesto de homocedasticidad previo a una prueba de comparación de dos medias altera la probabilidad global de cometer un error de tipo I respecto al nivel de significancia α planteado. Este problema queda superado cuando se usa directamente el test de Welch, por lo que se recomienda implementar este método como estándar en los software y libros estadísticos, eliminando el proceso tradicional que enseña a pre-testear los supuestos.

La presente investigación realiza una generalización de estos estudios para el caso de k medias y tiene como objetivo estudiar a través de un proceso de simulación estocástica la alteración de la Probabilidad de Error Tipo I que presentan estos test cuando el supuesto de homocedasticidad es previamente probado.

Los resultados mostraron que pre-testear homocedasticidad altera la probabilidad estimada y contrario al caso de dos muestras, el test de Welch ya no es una solución al problema, puesto que también presenta dificultades sobre todo para muestras pequeñas y un alto número de medias a compararse. Parece ser que el problema no es pre-testear sino más bien la forma tradicional como se plantean las pruebas de hipótesis para probar los supuestos.

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

Abstract

Some research show that pre-testing homoscedasticity assumption prior to a comparison two means test alters the overall Type I Error Probability respect to the level of significance α raised. This problem is overcome when the Welch test is used directly, so it is recommended to implement this method as a standard in software and statistical books, eliminating the traditional process that teaches pre-testing the assumptions.

This work makes a generalization of these previous studies for the case of k means. The aim is to study through a process of stochastic simulation the alteration of the Type I Error

1 Docente de la Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Grupo de Investigación en Ciencia de Datos CIDED.

Recibido: 20 Noviembre 2017 Aceptado: 3 Marzo 2018

Probability that these tests present when the homoscedasticity assumption is previously tested.

Results show that pre-testing homocedasticity alters the estimated probability and contrary to the case of two samples, the Welch test is not a solution to the problem, since it also presents difficulties especially for small samples and a high number of means to be compared. It seems that the problem is not to pre-test but rather the traditional way as hypothesis tests are proposed to prove the assumptions.

Key words: assumptions, pre-testing, homoscedasticity, Type I Error, Simulation.

Introducción

Las Pruebas de Hipótesis estadísticas usadas tradicionalmente para determinar diferencias significativas entre medias poblacionales están sujetas a la verificación previa de los supuestos de normalidad y homocedasticidad. El cumplimiento de estos supuestos se confirma mediante otros test de hipótesis, los cuales llamaremos pre-test. Así, la teoría tradicional en el campo de la estadística inferencial nos enseña que si rechazamos una hipótesis de normalidad en una prueba pertinente (Shapiro Wilk, Kolmogorov Smirnov, Anderson Darling, etc.) entonces un test no parámetrico (por ejemplo Wilcoxon para el caso de dos muestras o Kruskall Wallis para el caso general de k muestras) es utilizado, mientras que si la hipótesis nula no es rechazada se procede a verificar mediante otro pretest (F, Levene, Bartlet, Cochran, etc…) el supuesto de homocedasticidad. Cuando en estas pruebas, la hipótesis nula de igualdad de varianzas no es rechazada se asume homocedasticidad y utilizar un t – test (1) en el caso de dos muestras o una prueba ANOVA (2,3) en el caso general de k muestras suele asumirse como un procedimiento adecuado. Por otra parte, cuando la hipótesis nula de igualdad de varianzas es rechazada, se asume heterocedasticidad y otras pruebas alternativas al ANOVA son consideradas pertinentes para el contraste de comparación de medias. Mencionaremos dos de estas pruebas alternativas: El test de Welch para dos muestras (4) o en general para k muestras (5) que son una modificación del t – test y ANOVA respectivamente, estas pruebas contrastan una hipótesis nula de igualdad de medias, versus una hipótesis alternativa de diferencia de al menos una de ellas con respecto a las demás. La otra alternativa es la prueba de Dunnett (6) la cual realiza test de hipótesis para medias agrupadas de dos en dos en busca de diferencias significativas entre algún par de ellas, esta prueba controla la probabilidad global de cometer un error de tipo I (TIEP por sus siglas en inglés) cerca del nivel de significancia α planteado, y presenta mayor potencia estadística que otras pruebas similares (7).

A pesar de que la mayoría de libros y paquetes informáticos estadísticos utilizan el procedimiento descrito en el párrafo anterior para realizar pruebas de comparación de medias, existen muchos estudios (8–12) que demuestran que realizar test previos a la prueba de comparación de medias para comprobar supuestos conlleva a graves alteraciones en la TIEP, lo cual afecta significativamente los resultados del análisis estadístico, puesto que el analista cada vez que realiza uno de estos pre-test aumenta la probabilidad de rechazar una hipótesis nula de igualdad de medias que realmente es verdadera.

Estudios de simulación (13–15) han permitido estimar valores de la TIEP cuando se usan pre-test como Kolmogoroff-Smirnov y Levene para verificar normalidad y homocedasticidad respectivamente, previo a decidir el uso de una prueba de Wilcoxon, Welch o t-test para contrastar igualdad sobre dos medias poblacionales. Cuando este proceso se lleva a cabo, se observa que la TIEP global crece o decrece significativamente con respecto al valor nominal α dependiendo del tamaño muestral asignado a cada muestra, por lo que se sugiere eliminar este procedimiento de libros y paquetes informáticos estadísticos, y en su lugar se recomienda aplicar directamente el test de Welch (sin ninguna verificación previa de sus supuestos), ya que al hacerlo la TIEP global no se afecta y permanece cerca al nivel de significancia α aun cuando se utiliza muestras no normales y en casos donde se cumple o no el supuesto de homocedasticidad. Además se observó que para muestras normales donde solo se verificó previamente el supuesto de igualdad de varianzas, la alteración de la TIEP depende principalmente del tamaño de las muestras generadas, es decir la estimación se aleja más del nivel de significancia α cuando se utilizan muestras desbalanceadas que para tamaños

muestrales iguales y para este último caso la TIEP parece estabilizarse al nivel de α conforme el tamaño de la muestra aumenta, sin embargo nuevamente la aplicación directa de la prueba de Welch hace que la TIEP se estabilice independientemente del tamaño muestral que se utilice. Cabe indicar además que cuando se verifica homocedasticidad, la TIEP también se estabiliza dependiendo del nivel de significancia que se use en el pre-test, sin embargo esta opción no es viable puesto que dependiendo del tamaño muestral, la estimación empieza a estabilizarse a partir del valor poco práctico de $\alpha=0.20$

Posiblemente, el problema descrito tenga su origen en el planteamiento tradicional que se suele dar a las hipótesis que sirven para verificar los supuestos. En este sentido y refiriéndose específicamente a la normalidad, George Box menciona que "*en la naturaleza no existe una distribución perfectamente normal, sin embargo, con suposiciones normales, que se sabe que son falsas, a menudo se puede derivar resultados que coinciden, con una aproximación útil a los que se encuentran en el mundo real*" (16). Entonces si no existe un conjunto de datos perfectamente normales, o en analogía a nuestro trabajo, no existen datos perfectamente homocedásticos, no tiene sentido usar modelos que sirven para probar hipótesis de perfecta homocedasticidad o perfecta normalidad. Quizás el problema es que los modelos para verificar estos supuestos no son los adecuados y de esto se deriva el problema de la alteración de la TIEP, como en este mismo sentido el autor mencionó "*Todos los modelos son erróneos pero algunos son útiles*" (17).

Además, desde la perspectiva de Box podemos decir que lo interesante no es saber si las muestras provienen de una distribución normal o poseen perfecta homocedasticidad (ya sabemos que no), en lugar de esto lo realmente importante es saber si la aproximación de los modelos para comprobar estos supuestos es lo suficientemente buena como para ser utilizada. En el sentido de una estimación de la TIEP, el criterio de Cochran (18) sugiere que una aproximación se considera buena si existe una distancia máxima del 20% entre el valor estimado y el valor nominal de la TIEP que coincide con el nivel significancia nominal α conocido.

Al respecto del posible mal planteamiento que tradicionalmente han tenido las pruebas de hipótesis, que en este caso sirven para probar los supuestos de las pruebas de comparación de medias, Wellek menciona que "*Una diferencia no significativa no debe ser confundida con una significativa homogeneidad*" (19)", lo cual por ejemplo en el caso específico del pre-test de homogeneidad significa que cuando se plantea un test de hipótesis de la forma forma (H_0:$\sigma_1^2 = \sigma_2^2=...\sigma_k^2$ Vs H_1: Al menos alguna σ_i^2 es distinta a las demás), el analista se enfrenta a la dificultad lógica de que cuando se rechaza H_0 en realidad no se puede concluir que exista homogeneidad entre las varianzas σ_i^2 comparadas, en su lugar se concluye diciendo que no existe evidencia para decir que existen diferencias significativas entre ellas, lo cual no implica necesariamente homocedasticidad, o como en este mismo sentido Altman mencionó "*Ausencia de evidencia no significa evidencia de ausencia*" (20).

Las investigaciones realizadas y citadas en este artículo muestran la alteración de la TIEP global cuando se realizan pruebas de comparación de dos medias luego de haber realizado previamente pre-test de verificación de los supuestos bajo los cuales funcionan. Este mismo estudio es muy escaso cuando se trata de pruebas de comparación de k medias ($k \geq 2$), quizás la razón es que existen demasiados escenarios a compararse (diferentes pruebas, valores k, tamaños muestrales, niveles de heterocedasticidad, etc.) que hacen difícil una condensación de resultados. Existen sin embargo algunos estudios (7,21–23) que realizan procesos de simulación bajo escenarios teóricos de normalidad (o no) y homocedasticidad (o

no), a partir de lo cual se estima la TIEP de diferentes pruebas que existen para contrastar igualdad de k medias y se toma como la mejor opción aquella cuya TIEP permanece más estable respecto al nivel de significancia α. Sin embargo, ninguno de estos estudios muestra una estimación de la TIEP global cuando la elección de la prueba de comparación de medias depende de los resultados de pre-test utilizados para verificar sus supuestos.

La presente investigación se enfoca en determinar si existen alteraciones en la TIEP de ciertas pruebas de comparación de k medias condicionadas al resultado de un pre-test que compruebe el supuesto de igualdad de varianzas. A través de funciones propias de simulación desarrolladas en el software estadístico R-Studio (24), se estima la TIEP global usando muestras que teóricamente son normales y que presentan homocedasticidad así como también distintos niveles de heterocedasticidad teórica. En base a los antecedentes de esta investigación se ha decidido trabajar con la prueba de Levene como pre-test para verificar el supuesto de varianzas iguales y con las pruebas ANOVA, Welch (ANOVA modificada) y Dunnet como opciones para probar la hipótesis de igualdad de medias en las muestras simuladas. Los resultados únicamente son obtenidos a partir de muestras balanceadas, lo cual al final nos permitirá descubrir ciertos comportamientos de la TIEP que serán útiles para simplificar futuros estudios para casos desbalanceados

Materiales y Métodos

El proceso de simulación estocástica, consiste en generar a través de algoritmos computacionales muestras teóricamente normales con la misma media ($\mu_1 = \mu_2$), y distintos niveles de heterocedasticidad. Este grado de heterogeneidad en las varianzas está dado por la razón $\sqrt{(\sigma_1^2/\sigma_2^2)}$, esto debido a que se puede comprobar que la estimación de la TIEP de cualquier prueba para comparar medias (por ejemplo ANOVA), no se encuentra afectada por la distancia en valor absoluto que toman las varianzas de las muestras, sino más bien por la distancia de la razón entre ellas. La Tabla 1 muestra claramente lo dicho para el caso específico de tres muestras de tamaño 5 con razón entre varianza y $\sqrt{(\sigma_1^2/\sigma_2^2)} = 2$ y $\sqrt{(\sigma_1^2/\sigma_2^2)} = 4$.

Tabla1: Estimación TIEP de un ANOVA de k = 3 muestras, tamaño 5 y razón de varianza $\sqrt{(\sigma_1^2/\sigma_2^2)} = 2$ y $\sqrt{(\sigma_1^2/\sigma_2^2)}=4$

	$(\sigma_1, \sigma_2, \sigma_3)$	TIEP		$(\sigma_1, \sigma_2, \sigma_3)$	TIEP
$\sqrt{\sigma_1^2/\sigma_2^2} = 2$	(1, 2, 4)	0.0812	$\sqrt{\sigma_1^2/\sigma_2^2} = 4$	(0.25, 1, 4)	0.108
	(3, 6, 12)	0.0812		(4, 16, 64)	0.108
	(5, 10, 20)	0.0812		(10, 40, 160)	0.108

Cabe indicar que una razón entre varianzas de $\sqrt{(\sigma_1^2/\sigma_2^2)} = 1$ indica una perfecta homocedasticidad en las distribuciones normales a partir de las cuales son generadas las muestras. Solo en este caso es bien conocido que para una prueba ANOVA de un factor la TIEP coincide con el nivel de significancia α de la prueba. Sin embargo cuando la razón entre varianzas es mayor que 1, o el procedimiento es diferente al ANOVA en condiciones de homocedasticidad, el valor de esta probabilidad es un parámetro desconocido, el cual precisamente puede ser estimado por un proceso de simulación.

Realizaremos simulaciones para estimar los siguientes parámetros: TIEP de una prueba ANOVA aplicada directamente sin ningún pre-test de comprobación de homocedasticidad (AD), TIEP de una prueba de Welch aplicada directamente (WD), TIEP de una prueba de Dunnett aplicada directamente (DD), TIEP de una prueba ANOVA o Welch, que depende del resultado de la prueba de Levene (PWA) y TIEP de una prueba ANOVA o Dunnet que depende del resultado de Levene (PWD). Todos estos escenarios con muestras obtenidas de una distribución normal con la misma media y valores de la razón entre varianza $\sqrt{(\sigma_1^2/\sigma_2^2)} = 1,1.2,1.4,1.8,2,3,5$.

El estimador de la TIEP en cada uno de estos escenarios es la proporción de veces que se rechaza la hipótesis nula de igualdad de medias cuando se generan muestras que teóricamente tienen la misma media. A pesar de realizar para cada caso un total de 100 000 simulaciones, empleamos la técnica de reducción de varianza denominada "Variables de Control" (25,26), con el fin de reducir la variabilidad de la simulación y conseguir más precisión en la simulación. El proceso está implementado dentro de los algoritmos de simulación que se utilizó.

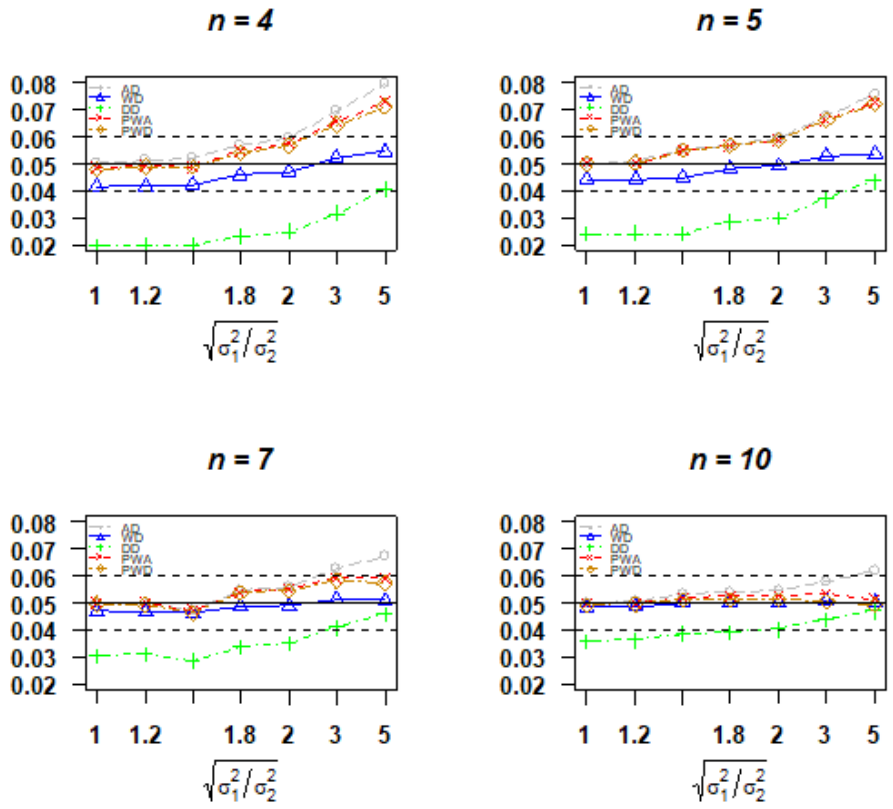
En el caso de la Prueba de Dunnett se utilizó la librería "DTK" (27) descargada directamente del repositorio R-CRAN. Este test como ya habíamos mencionado hace comparaciones por pares, es por esto que la lógica utilizada en todos los procesos en que esta prueba está inmiscuida es rechazar la hipótesis nula general de igualdad de k medias si al menos en una prueba pareada se considera la existencia de diferencias significativas. En el caso de las otras pruebas (ANOVA y Welch) se contrastó directamente la hipótesis nula de igualdad de k medias mediante el respectivo valor P.

Para las diferentes simulaciones se utilizaron tamaños muestrales balanceados de n=4,5,7,10, realizando comparaciones de k=2,3,4,5 medias poblacionales bajo los escenarios ya planteados (pre-test, razones de varianza, número de simulaciones, etc…). En todos estos escenarios se utiliza un valor nominal $\alpha = 0.05$, y de acuerdo al criterio de Cochran consideraremos que un proceso de test estadístico para comparar k medias es aceptable cuando la estimación de la TIEP se encuentre alejada del valor nominal α una distancia máxima de ±20% α, es decir una estimación aceptable de la TIEP será aquella que esté dentro del intervalo (0.04, 0.06).

Resultados

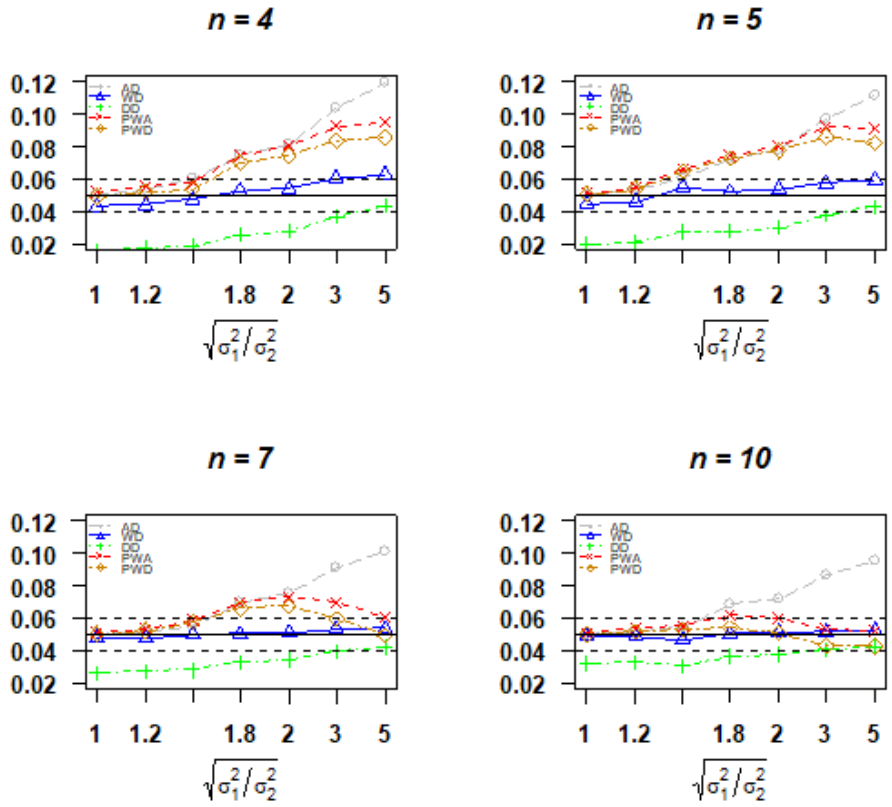
La Fig.1 muestra la estimación de la TIEP cuando dos medias poblacionales son comparadas. Basados en el criterio de Cochran, en todos los casos se observa que conforme el número de observaciones en las muestras es mayor, la aproximación de los modelos es mejor, esto debido a que la TIEP para muestras grandes permanece estable alrededor del nivel de significancia $\alpha=0.05$ y dentro del intervalo (0.04 – 0.06), de hecho parece ser que a partir de n=10 cualquier procedimiento mantiene controlada la TIEP con una leve excepción para el procedimiento DD en casos de perfecta homocedasticidad y leve heterocedasticidad. Para tamaños muestrales pequeños se observa que conforme el nivel de heterocedasticidad aumenta, los procedimientos AD, PWA y PWD alteran la TIEP por encima del nivel de significancia llegando a ser inútiles a partir de ciertos niveles de heterocedasticidad, mientras que el procedimiento DD es una mala aproximación para niveles bajos de heterocedasticidad ya que su TIEP se encuentra alterada por debajo del nivel de significancia, aunque esto parece mejorar conforme el nivel de heterocedasticidad es mayor y el tamaño muestral aumenta. Es importante recalcar que independientemente del grado de heterocedasticidad y del tamaño muestral, el procedimiento WD mantiene estable la TIEP y dentro de los límites establecidos por el criterio de Cochran, por lo que parece ser el mejor modelo de todos los analizados, lo cual confirmaría las conclusiones de las investigaciones anteriores respecto a este caso particular de dos medias comparadas.

Fig.1: Estimación de la TIEP para pruebas de comparación de k=2 medias (H_0:$\mu_1=\mu_2$ Vs H_1:$\mu_i\neq\mu_j$ para algún i≠j) con α=0.05



La Fig.2 muestra la estimación de la TIEP cuando tres medias poblacionales son comparadas. Se puede observar que la alteración de la TIEP para la mayoría de modelos utilizados es más grande que en el caso donde se comparan dos medias poblacionales. La dinámica de alteración sin embargo es similar, aunque empezamos a notar que para ningún escenario se puede recomendar el procedimiento AD, y los procesos DD, PWA y PWD son menos efectivos que antes. Sin embargo el proceso WD sigue siendo bueno ya que en todos los escenarios continúa manteniendo estable la TIEP y dentro de los intervalos (0.04 – 0.06), al menos en estos dos casos hasta ahora analizados k=2 y k=3.

Fig.2: Estimación de la TIEP para pruebas de comparación de k=3 medias (H_0:$\mu_1=\mu_2=\mu_3$ Vs H_1:$\mu_i\neq\mu_j$ para algún i≠j) con α=0.05



La Fig.3 y la Fig.4 muestran la estimación de la TIEP para la comparación de cuatro y cinco medias respectivamente. Estos casos producen una alteración de la TIEP mucho mayor que cuando se compararon dos y tres medias, especialmente para los procesos AD, PWA y PWD, los cuales presentan estimaciones muy alejadas del nivel de significancia conforme los tamaños muestrales son menores y el nivel de heterocedasticidad aumenta. Llama la atención el procedimiento WD, el cual hasta ahora parecía permanecer robusto ante el tamaño muestral y los distintos niveles de heterocedasticidad, pero para muestras pequeñas presenta

Fig.3: Estimación de la TIEP para pruebas de comparación de k= 4 medias (H_0:$\mu_1=\mu_2=\mu_3$ Vs H_1:$\mu_i\neq\mu_j$ para algún i≠j) con α=0.05

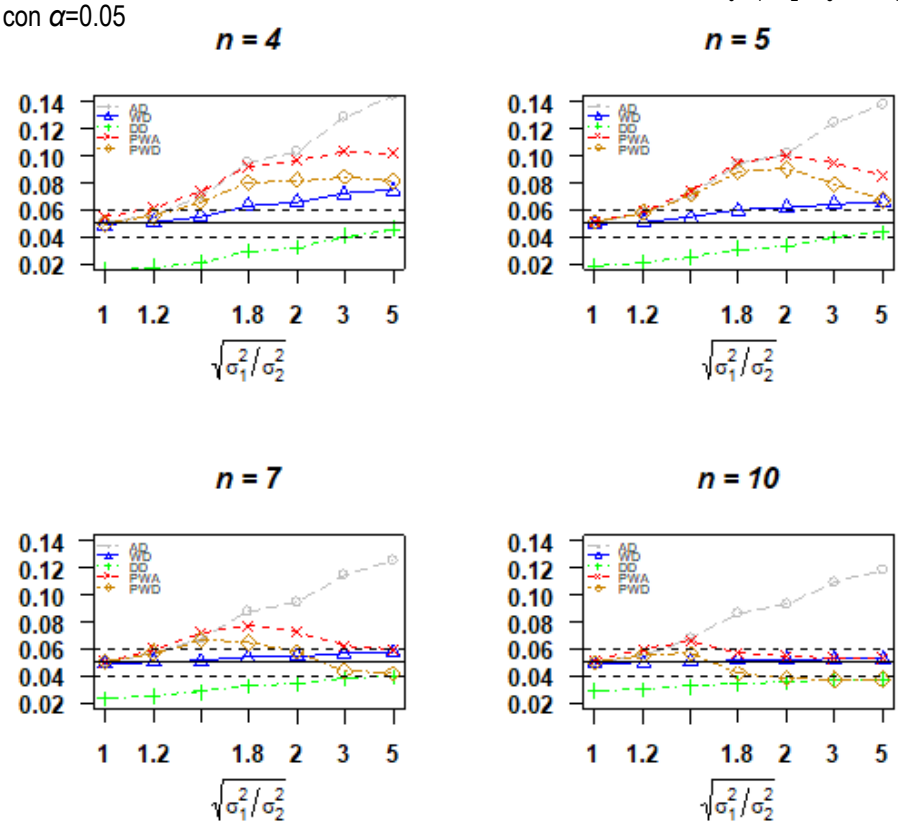
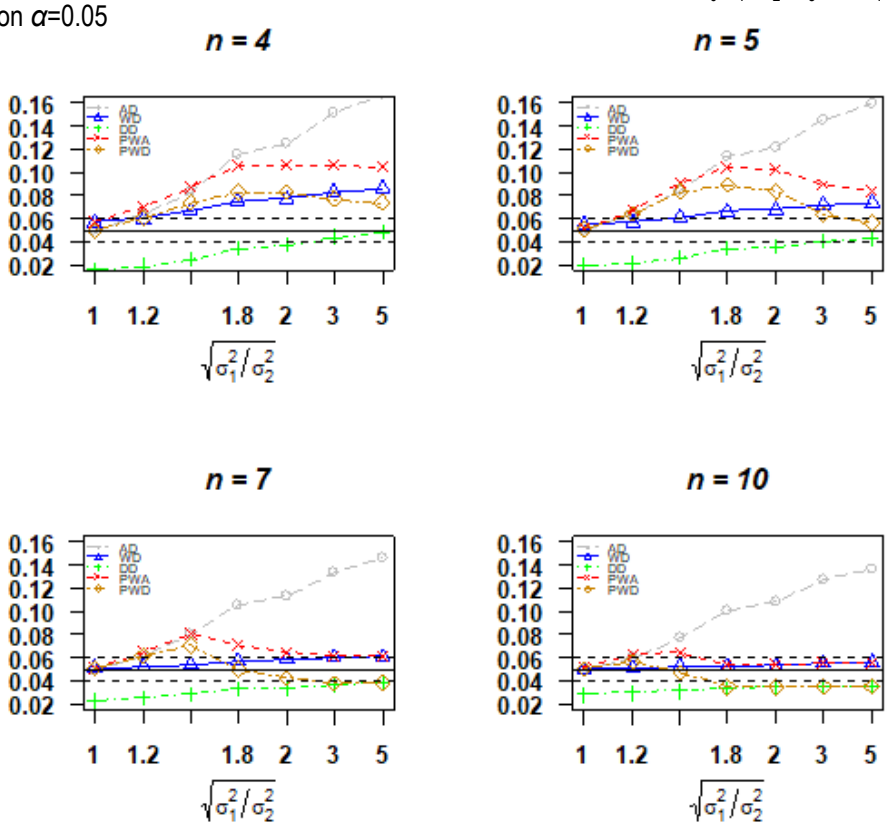


Fig.4: Estimación de la TIEP para pruebas de comparación de k= 5 medias (H_0:$\mu_1=\mu_2=\mu_3$ Vs H_1:$\mu_i\neq\mu_j$ para algún i≠j) con α=0.05



una alteración significativa por encima de α que crece conforme el nivel de heterocedasticidad aumenta. En estos dos casos analizados, a diferencia de los dos casos anteriores no podríamos decir que existe un método que de aproximaciones buenas, ya que ninguno ha demostrado controlar la TIEP alrededor del nivel de significancia planteado, y a la larga todos en algún punto están fuera del intervalo (0.04 – 0.06).

Discusión

Al igual que estudios anteriores donde se estableció que el proceso de pre-testear el supuesto de homocedasticidad trae alteraciones en la TIEP de las pruebas de comparación de dos medias, hemos determinado a través de la presente investigación que para el caso general (comparación de k medias) también existe esta alteración, la cual puede conllevar a tomar decisiones erróneas en el sentido de que la probabilidad de rechazar la hipótesis de igualdad cuando realmente es cierta es más grande de lo que establece el nivel de significancia con el que se trabaja.

El proceso de realizar directamente la prueba ANOVA (AD) resultó ser el peor de todos, ya que en los escenarios simulados se observa la más fuerte alteración de la TIEP, es decir los valores estimados se encuentran más lejos que las estimaciones realizadas siguiendo los otros procesos.

El proceso que consiste en aplicar Dunnett directamente también resultó ser malo, ya que en la mayoría de casos la TIEP está muy por debajo del nivel de significancia α, a excepción de aquellos escenarios donde existe un alto grado de heterocedasticidad y las muestras son pequeñas.

En el caso donde se aplicó directamente Welch, se observó que para muestras grandes, este procedimiento controla muy bien la TIEP alrededor de α, pero para muestras pequeñas y conforme el número de medias comparadas k aumenta, las estimaciones calculadas con este método empiezan a presentar graves alteraciones que se ubican por encima del nivel de significancia.

Los dos casos (PWA y PWD) donde la prueba de comparación de medias (ANOVA o Welch), se utiliza con base a la decisión de un pre-test para verificar el supuesto de homocedasticidad tienen resultados similares. En este sentido se determina ninguno de estos procedimiento es una buena alternativa, ya que la TIEP queda alterada especialmente cuando los tamaños muestrales son pequeños y el número de medias poblacionales a comparar aumenta.

Cuando el tamaño muestral es grande (como casi todo lo que ocurre en estadística) las cosas mejoran y en este escenario aplicar directamente la prueba de Welch es el mejor método, pero definitivamente realizar pre-test de verificación del supuesto no parece ser la mejor idea. Si como en los estudios de referencia para dos medias (mencionados en la introducción) la TIEP para casos desbalanceados es mayor a la TIEP que para casos balanceados, esperaríamos que al trabajar con diferentes tamaños muestrales los resultados sean peores que los obtenidos para este estudio donde se utilizó el mismo tamaño en las muestras generadas. Lo cual nos deja sin ninguna alternativa confiable al menos cuando existen tamaños muestrales pequeños (menores que 7) y aumenta el número de medias a comparar (a partir de 4) y el nivel de heterocedasticidad. Esto no ocurría en el caso de comparación de dos medias, donde los estudios previos demostraron que el test de Welch permanecía robusto independientemente del tamaño muestral y el nivel de heterocedasticidad, lo cual muestra que el problema del mal planteamiento de las hipótesis (perfecta homocedasticidad) descrito por Box, no era visible en las investigaciones donde solo se compara dos medias, pero las alteraciones empiezan hacerse visibles conforme este número aumenta. De

acuerdo a esto tal vez una prueba de hipótesis que no plantee comprobar una perfecta homocedasticidad sea la solución al problema.

Al igual que ya lo hace una investigación anterior (12), recomendaríamos que el proceso de pre-testear sea eliminado de los libros y paquetes informáticos estadísticos, en su lugar se use directamente el test de Welch solamente si se tiene una gran cantidad de datos y las medias a compararse no son demasiadas, caso contrario se recomienda investigar otro tipo de test estadísticos que no funcionen bajo el planteamiento tradicional de pruebas de hipótesis (perfecta igualdad de tratamientos).

Referencias Bibliográficas

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

2. Cribbie, Robert A and Fiksenbaum, Lisa and Keselman, HJ and Wilcox RR. Effect of non-normality on test statistics for one-way independent groups designs. Br J Math Stat Psychol. 2012;65(1):56–73.

3. Cochran WG. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics. 1947;3:22–38.

4. Welch BL. On the Comparison of Several Mean Values: An Alternative Approach [Internet]. Vol. 38, Biometrika. 1951. p. 330. Available from: http://www.jstor.org/stable/2332579?origin=crossref

5. Welch ABL. The Generalization of Student's Problem when Several Different Population Variances are Involved Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2332510. Biometrika. 2008;34(1):28–35.

6. Dunnett C. in the Unequal Pairwise Multiple Comparisons Variance Case. J Am Stat Assoc. 1980;75(372):796–800.

7. Olejnik, Stephen and Lee J. Multiple Comparison Procedures when Population Variances Differ. Paper presented at the Annual Meeting of the American Educational Research Association. 1990.

8. Hsu P. Contribution to the theory of" Student's" t-test as applied to the problem of two samples. Stat Res Mem. 1938;

9. Zimmerman D. A note on preliminary tests of equality of variances. 2004;57:173–81.

10. Overall, John E and Atlas, Robert S and Gibson JM. Tests that are robust against variance heterogeneity in kx2 designs with unequal cell frequences. Psychol Rep. 1995;76:1011–7.

11. Montilla J, Kromrey J. Robustez de las pruebas T en comparación de medias , ante violación de supuestos de normalidad y homocedasticidad. Rev Cienc e Ing. 2010;31(2):101–8.

12. Rasch D, Kubinger KD, Moder K. The two-sample t test: Pre-testing its assumptions does not pay off. Stat Pap. 2011;52(1):219–31.

13. Zimmerman D. A note on preliminary tests of equality of variances. 2004;57(May):173–81.

14. Moder K, Rasch D, Kubinger KD. Don ' t use the two-sample t-test anymore ! 2009;258–62.

15. Flores Muñoz P. Un Pretest de Irrelevancia de la diferencia de varianzas en la comparación de medias. 2017.

16. Box GE. Robustness in the strategy of scientific model building. Army Res Off Work Robustness Stat. 1979;1:201–36.

17. Draper GEPB and NR. Empirical Model Building and response Surface. John Wiley Sons. 1987;669.

18. Cochran WG. The x2 correction for continuity. Iowa State Coll J Sci. 1942;16:421–36.

19. Wellek S. Testing statistical hypotheses of equivalence and noninferiority. 2010. 3 p.

20. Altman, Douglas G and Bland JM. Statistics notes: Absence of evidence is not evidence of absence. Bmj. 1995;311:485.

21. Hartung J, Argac D, Makambi KH. Small sample properties of tests on homogenity in one-way Anova and Meta-analysis. 2001; Available from: http://hdl.handle.net/2003/5273%5Cnhttps://eldorado.tu-dortmund.de/bitstream/2003/5273/1/32_01.pdf

22. Asiribo, Osebekwin and Gurland J. Coping with variance heterogeneity. Commun Stat Methods. 1990;19:4029–48.

23. James G. The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika. 1951;38:324–9.

24. Team RC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/

25. Ocaña J, Vegas E. Variance reduction for Bernoulli response variables in simulation. Comput Stat Data Anal. 1995;19(6):631–40.

26. Vegas E, Castillo J del, Ocaña J. Efficiency and exponential models in a variance-reduction technique for dichotomous response variables. J Stat Plan Inference [Internet]. 2000;85:61–74. Available from: http://www.sciencedirect.com/science/article/pii/S037837589900066X

27. Lau MMK. Package " DTK ." 2015;1–7.

Correspondencia
Autor: Pablo Javier Flores Muñoz
Dirección: Escuela Superior Politécnica de Chimborazo - Teléfono: (593)958958295
Email: p_flores@espoch.edu.ec