UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2011, Vol. 4, No. 2

¿Se necesita la prueba t de Student para dos muestras independientes asumiendo varianzas iguales?

Is it actually needed the t-student test for two independent samples when assuming the same variances?

Jorge Eduardo Ortiz[a]
jorgeortiz@usantotomas.edu.co

Edna Carolina Moreno[b]
ednamoreno@usantotomas.edu.co

Resumen

En este trabajo se hace un examen del comportamiento de la proporción de rechazos equivocados de la hipótesis nula (error tipo I) en condiciones plenas de aplicabilidad de la distribución t de Student, es decir, con variables independientes cuya distribución es normal tanto bajo el supuesto de homogeneidad de varianzas como en las condiciones de heterocedasticidad.

***Palabras clave*:** estadística t, prueba t, monotonía, estabilidad, prueba de Welch-Satterwhaite.

Abstract

In this paper we review the behavior of the type I error rate of Student's t-test and Welch-Sattetthwaite test for comparing two means with independent samples from normal populations under the assumption of homogeneity of variances and under conditions of heteroscedasticity. The results, obtained by the Monte Carlo method show the Welch-Satterthwaite well behaved in all cases.

***Key words*:** t statistic, t-Test, Welch-Satterwhaite test, Heterocedasticity.

1. Introducción

La prueba t para dos muestras independientes es uno de los procedimientos estadísticos favorecidos por la atención de los investigadores, tanto en las aplicaciones, como en el estudio de sus propiedades. A pesar de esto, la literatura dedicada

[a]Docente. Facultad de Estadística, División de Ciencias Administrativas, Económicas y Contables, Universidad Santo Tomás, Bogotá

[b]Docente. Facultad de Estadística, División de Ciencias Administrativas, Económicas y Contables, Universidad Santo Tomás, Bogotá.

a la enseñanza de la estadística básica pareciera no tomar en cuenta los principales resultados ya confirmados por diferentes autores. El propósito de este trabajo es hacer conciencia sobre los problemas que se generan con la aplicación de ciertos procedimientos inapropiados, pero muy populares, y sobre los riesgos de utilizar la prueba t asumiendo que las varianzas poblacionales son iguales. Esperamos que los resultados obtenidos faciliten la decisión sobre la enseñanza y el uso adecuado de esta prueba.

Para unificar la notación del articulo, la hipótesis nula se presenta como $H_0 : \delta = \mu_X - \mu_Y = \delta_0$, donde δ_0 es un valor especificado por el investigador y X y Y son variables aleatorias con distribuciones normales de las que se desconocen las varianzas. Uno de los procedimientos estadisticos consiste en extraer muestras aleatorias de cada población de manera independiente. Los tamaños de muestra son n_1 y n_2, los promedios y las varianzas muestrales son $\overline{X}$, $\overline{Y}$, S_X^2 y S_Y^2, respectivamente. $S_p^2 = \big((n_1-1)S_X^2+(n_2-1)S_Y^2\big)/(n_1+n_2-2)$ se conoce como la varianza ponderada. Bajo el supuesto de igualdad de varianzas se utiliza la estadística de Student:

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{1}$$

cuya distribución bajo H_0 es $T_{n_1+n_2-2}$ (T de Student con $n_1 + n_2 - 2$ grados de libertad). La regla de decisión depende de la hipótesis alternativa y puede revisarse en cualquiera de los textos referenciados más adelante.

1.1. La prueba de Welch-Satterthwaite

Si las varianzas no son iguales, la principal opción disponible en los programas estadísticos actuales es la prueba de Welch-Satterthwaite, basada en la estadística:

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}} \tag{2}$$

La diferencia entre las varianzas dificulta en gran medida el cálculo de la función de distribución de T. Sin embargo, ya Welch (1938), Welch (1947) y Satterthwaite (1946) han ofrecido aproximaciones que se consideran satisfactorias para el uso práctico.

Satterthwaite (1946) define un estimador complejo de varianza como una combinación lineal de cuadrados medios independientes. Welch (1938) ya había demostrado antes que la distribución de este tipo de estimadores puede aproximarse con la distribución Ji-cuadrada. Específicamente, si MS_i son cuadrados medios independientes con r_i grados de libertad, $i = 1, 2, \ldots, k$, y si $\widehat{V}_s = \sum_{i=1}^{k} a_i MS_i$ es

un estimador complejo de varianza basado en ellos, los grados de libertad de la aproximación χ^2 son

$$r_s = \frac{\left(\sum_{i=1}^{k} a_i E(MS_i)\right)^2}{\sum_{i=1}^{k} \frac{\left(a_i E(MS_i)\right)^2}{r_i}} \tag{3}$$

Los $E(MS_i)$ son desconocidos pero Satterthwaite (1946) verifica, para varios casos, que se pueden reemplazar por los cuadrados medios sin generar mayores inconvenientes en la aproximación a la distribución χ^2 con grados de libertad dados por:

$$\widehat{r}_s = \frac{\left(\sum_{i=1}^{k} a_i MS_i\right)^2}{\sum_{i=1}^{k} \frac{\left(a_i MS_i\right)^2}{r_i}} \tag{4}$$

Para el caso de dos muestras independientes, la diferencia de medias, $\mu_X - \mu_Y$, se estima con $\overline{X} - \overline{Y}$. Su varianza, $\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}$, se estima con $\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}$. Este es el estimador complejo de varianza con $k = 2$. Para la primera muestra, $a_1 = \frac{1}{n_1}$, $MS_1 = MS_X = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_i - \overline{X})^2 = S_X^2$, $r_1 = n_1 - 1$ y $E(MS_X) = \sigma_X^2$. Para la segunda, $a_2 = \frac{1}{n_2}$, $MS_2 = MS_Y = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_i - \overline{Y})^2 = S_Y^2$, $r_2 = n_2 - 1$ y $E(MS_Y) = \sigma_Y^2$. Entonces (4) queda:

$$\widehat{r}_s = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\frac{\left(\frac{S_X^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_Y^2}{n_2}\right)^2}{n_2 - 1}} \tag{5}$$

y así:

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \approx T_{\widehat{r}_s} \tag{6}$$

donde el símbolo $\approx$ se toma para indicar que la distribución es aproximada. Los grados de libertad resultan ser aleatorios, pues son función de la muestra. Esta es una de las diferencias esenciales con la prueba de Student construida para el caso de varianzas iguales, donde los grados de libertad dependen exclusivamente de los tamaños de las muestras.

Winer (1971, p.42) menciona también la siguiente aproximación de Satterthwaite para los grados de libertad:

$$\widehat{r}_s = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\frac{\left(\frac{S_X^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_Y^2}{n_2}\right)^2}{n_2+1}} - 2 \tag{7}$$

Así, para comparar las medias de dos poblaciones con distribuciones normales con muestras independientes, los investigadores encuentran diversas opciones, dependiendo de la posición que asuman con respecto al supuesto de homogeneidad de varianzas. Analizaremos las más comunes.

2. Método

Se utilizó el método de Monte Carlo que, en resumen, consiste en generar muchas réplicas de la situación cuyo comportamiento se quiere conocer, mediante números pseudoaleatorios. Éstos deben presentar las características exigidas por las condiciones establecidas en la situación simulada. Para nuestro caso, se trata de realizar pruebas de Student y de Welch-Satterthwaite (6) para dos muestras independientes que provienen de distribuciones normales.

Para cada combinación de valores de n_1, n_2 y σ_Y, se generaron 100 000 réplicas de pruebas de $H_0 : \mu_X = \mu_Y$ contra $H_1 : \mu_X \neq \mu_Y$, con un nivel de significación $\alpha = 0.05$. Para las muestras pseudoaleatorias se utilizó la función `rnorm()` de R con los argumentos `n1` $= n_1$, `muX` $= \mu_X$ y `sigmaX` $= \sigma_X$ para la muestra X, y `n2` $= n_2$, `muY` $= \mu_Y$, `sigmaY` $= \sigma_Y$ para la muestra Y, con tamaños $n_1, n_2 = 6, 8, \ldots, 20$ y desviaciones estándar $\sigma_X = 1$ y $\sigma_Y = 1, 1.25, 1.5, 1.75, 2, 3$.

3. Los procedimientos más comunes

Revisamos algunos de los textos y de los programas de análisis estadístico más influyentes en nuestro medio, tanto en la enseñanza de los métodos inferenciales a nivel de pregrado como en la práctica.

Es usual encontrar programas de análisis estadístico que presentan los resultados de dos pruebas de comparación de medias: una asumiendo que las varianzas poblacionales son iguales (prueba de Student) y la otra asumiendo que no (prueba de Welch-Satterthwaite).

Como ilustración, supongamos que la muestra x contiene los siguientes valores: 156, 124, 153, 148, 139, 135 y la muestra y: 148, 155, 162, 149, 147, 158, 162, 165, 142, 150. La siguiente es una presentación típica de los resultados de las pruebas:

Tabla 1: *Prueba t. Presentación típica*

Prueba de Levene		Prueba t			
F	Sig.		t	gl	Sig.
2.167	.163	Asumiendo varianzas iguales	-2.298	14	.037
		Sin asumir varianzas iguales	-2.049	7.484	.077

3.1. Tomar el mínimo valor-p de las dos pruebas

Zimmerman & Zumbo (2009) advierten que con esta presentación, muchos investigadores se ven tentados a escoger la prueba cuyos resultados sean "más significativos". Para el ejemplo anterior, el investigador decidiría rechazar H_0 al encontrar que uno de los valores-p es inferior a $\alpha = 0.05$. Este procedimiento no sólo es incorrecto sino que, además, pone en evidencia un sesgo subjetivo en contra de H_0 por parte de quien lo utilice. Hacerlo lleva a preguntarse por qué no se realizan simultáneamente otras pruebas paramétricas y no paramétricas buscando alguna que rechace H_0. De todas maneras, hemos programado este procedimiento para conocer los efectos de su uso.

En la figura 1 (página 152) se muestran las proporciones de rechazo incorrecto de H_0 cuando las varianzas son iguales y para tamaños de muestra variando entre 6 y 50, como se indicó en la sección 2. Los puntos resaltados en la figura corresponden a las pruebas con muestras de tamaños iguales. En estos casos, numéricamente las proporciones de rechazo se mantienen cerca del nivel de significación (5 %). En los demás, se aprecia la tendencia creciente del riesgo de cometer el error tipo I cuando los tamaños de las muestras son más distanciados: prácticamente todas las proporciones encontradas son superiores al nivel de significación, con lo que se confirma que con esta regla de decisión, efectivamente, el investigador se inclina a rechazar la hipótesis nula en proporciones mayores a lo acordado con α. Ya con pequeñas diferencias entre las desviaciones estándar poblacionales, el problema se agrava, especialmente si la muestra de mayor tamaño proviene de la población con menor dispersión, como se ilustra en la figura 2 (página 153).

En otras simulaciones, que no presentamos aquí, se observa la tendencia a empeorar el control de la probabilidad de cometer el error tipo I a medida que crecen las diferencias entre las dispersiones poblacionales. El hecho de que la igualdad en los tamaños de las muestras atenúe el problema numérico no justifica el error conceptual en el que se incurre al utilizar este procedimiento.

3.2. Aplicación de una prueba preliminar de homogeneidad de varianzas

Sawilowsky (2002) señala que una estrategia común, pero incorrecta, consiste en aplicar una prueba de comparación de varianzas antes de la prueba t para decidir

cuál utilizar según los resultados obtenidos. Sus referencias principales son los manuales de programas estadísticos ampliamente reconocidos: SAS, SYSTAT y SPSS.

No sólo en los manuales se propone esta estrategia. Marques de Sá (2007) sugiere explícitamante que "para decidir cuál caso tomar, de varianzas iguales o desiguales, se aplica la prueba F o la de Levene. Eso es lo que hacen SPSS o STATISTICA". Park (2009) presenta un diagrama de flujo donde incluye la prueba de homogeneidad de varianzas como un paso obligado para decidir cuál prueba utilizar para comparar las medias.

En otros casos, los autores mencionan la prueba de comparación de varianzas como una forma de verificación del supuesto de homocedasticidad. ¿Qué hacer si no se encuentran evidencias en contra o a favor? Por lo general, los investigadores terminan escogiendo una u otra prueba en función del resultado de la verificación. Esto equivale a aplicar una prueba condicionada por los resultados de la otra.

Intuitivamente el procedimiento parece justificado y, a diferencia del anterior, es un método objetivo. Sin embargo, la condicionalidad de la prueba de comparación de medias con respecto a la de varianzas obliga a estudiar su eficacia y su conveniencia. Como en este trabajo se supone que las distribuciones son normales y que las muestras son independientes, la comparación de varianzas se hace con la prueba F y, según el resultado obtenido, se automatiza la aplicación de la prueba de Student para dos muestras independientes con varianzas iguales o la de Welch-Satterthwaite. Por razones que veremos más adelante, para esta última utilizaremos la aproximación dada en (5 y 6)

De manera similar a la de la sección 3.1, se simularon pruebas t condicionadas por los resultados de pruebas F de comparación de varianzas, en varias condiciones de relaciones entre las desviaciones estándar.

Cuando las varianzas de las distribuciones de X y de Y son iguales, las proporciones de rechazo de H_0 tienen un comportamiento aceptable en el sentido de permanecer cerca del nivel de significación. Para los tamaños de muestra considerados, las proporciones más alejadas son de 0.058 (figura 3, página 154). De nuevo, cuando los tamaños de muestra son iguales, los resultados son más estables alrededor de α.

Con las diferencias en las desviaciones estándar, los resultados muestran las deficiencias de este procedimiento: si de la población con menor varianza se tiene la muestra más grande, la proporción de rechazos con error tipo I es mayor que α y el problema crece a medida que la razón entre la varianza mayor y la menor también lo hace. La figura 4 (página 155) ilustra el comportamiento en el caso particular de $\sigma_Y = 1.75\ \sigma_X$ y la figura 5 (página 156) ilustra el caso de $\sigma_Y = 1.5\ \sigma_X$.

Por otra parte, si la muestra proveniente de la distribución con mayor varianza es la de mayor tamaño, entonces la proporción de rechazos disminuye llegando a ser menor que α y en los casos más extremos considerados se acerca a 0.035. Las mejores condiciones se presentan cuando los tamaños de las muestras son iguales.

3.3. Aplicación de la prueba t para varianzas iguales

El supuesto de igualdad de varianzas es muy difícil de respaldar en la práctica, en especial cuando se dispone de muestras de tamaños pequeños. En el ejemplo de la sección 3, a pesar de que no se rechaza la hipótesis de homocedasticidad, la relación entre las varianzas muestrales es de 2.167 (entre las desviaciones estándar es de 1.47) y un intervalo de confianza admite la posibilidad de que alcance a 16, es decir que una de las desviaciones estándar podría ser hasta cuatro veces la otra. Aun así, la bibliografía es generosa en presentar la prueba de Student para dos muestras independientes con varianzas iguales.

Canavos (1988) sólo incluye la prueba t que utiliza la estimación de la varianza de la diferencia de promedios mediante el promedio de las varianzas muestrales ponderadas por los grados de libertad. En las páginas 339 y 340 menciona la falta de robustez de la prueba cuando las varianzas poblacionales son diferentes, destacando que el efecto no es importante cuando los tamaños de muestra son iguales.

Newbold et al. (2008, pp. 334, 404) presentan la aproximación de Satterthwaite, pero con un comentario que desmotiva su uso: "[. . .] la determinación de los grados de libertad [. . .] es muy compleja [. . .]"; y luego dan un ejemplo donde su aplicación no muestra mayores ventajas frente a la prueba t para el caso de varianzas iguales.

¿Se puede aplicar la prueba t para dos muestras independientes con varianzas iguales sin tener en consideración la validez de este supuesto?; en otras palabras, ¿la prueba t para varianzas iguales es robusta si se viola el supuesto de homocedasticidad? Ya desde 1929, Behrens había advertido sobre la poca robustez de la prueba t en casos de heterocedasticidad y sobre la necesidad de encontrar un procedimiento estadístico para comparar las medias de dos poblaciones con muestras independientes sin depender de la condición de varianzas poblacionales iguales. Kim & Cohen (1995) presentan un resumen importante sobre este problema, conocido como el problema de Behrens-Fisher.

No se incluyen los resultados obtenidos cuando las varianzas son iguales. Teóricamente, estas son las condiciones de funcionamiento óptimo de la prueba de Student y, obviamente, así se encuentra. Para ilustrar el comportamiento típico de la proporción de rechazos de H_0 verdadera cuando las varianzas no son iguales, se muestran los resultados de las simulaciones con $\sigma_Y = 1.5\ \sigma_X$.

Las tendencias generales son similares, pero más extremas que las de la estrategia de aplicar la prueba t o la de Welch-Satterthwaite según el resultado de la prueba preliminar de homocedasticidad.

Si la muestra proveniente de la población con menor varianza es la de mayor tamaño, la tasa de error tipo I es mayor a la esperada con α y esta diferencia se incrementa a medida que los tamaños de muestra se alejan.

Si la muestra proveniente de la población con menor varianza es la de menor tamaño, la tasa de error tipo I es inferior a la esperada con α y esta diferencia se incrementa a medida que los tamaños de muestra se alejan.

Con mayores diferencias entre las desviaciones estándar los problemas se agravan, pero la igualdad de tamaños muestra reduce considerablemente las diferencias entre las tasas de error tipo I y el nivel de significación.

4. Aplicación de la prueba de Welch-Satterthwaite

Se acostumbra presentar la prueba de Welch-Satterthwaite como la prueba t para varianzas desiguales. Las simulaciones permiten conocer el comportamiento de las tasas de error tipo I tanto en condiciones de homocedasticidad como de heterocedasticidad.

En la figura 6 (página 157) se aprecia que las proporciones de error tipo I son cercanas al nivel de significación ($\alpha = 0.05$), encontrándose dentro del intervalo $(0.046, 0.058)$.

Se realizaron simulaciones para relaciones de desviaciones estándar diferentes y se encontró que la cercanía de las tasas de error tipo I y el nivel de significación se mantiene estable y dentro del mismo intervalo en todos los casos considerados. La mayor ganancia está en haberse liberado de una condición de homogeneidad de varianzas, por lo general, muy difícil de sustentar aun habiéndola examinado con una prueba preliminar.

En las tablas 2 a 7 (páginas 147-148) se resumen algunos aspectos descriptivos del comportamiento de las proporciones de error tipo I para relaciones entre las desviaciones estándar indicadas en cada una. Se presenta el mínimo, los cuartiles, el promedio, la mediana y la desviación cuadrática media, definida como $DCM = \sqrt{\sum_j (p_{Rj} - \alpha)^2 / n}$, donde p_{Rj} son las proporciones simuladas de rechazo de H_0 con cada procedimiento. Esta medida indica qué tan lejanos se encuentran los valores simulados del nivel de significación utilizado.

Los procedimientos se representan con Tvi para la prueba de Student asumiendo igualdad de varianzas, $Tv1$ para la prueba de Welch-Satterthwaite con grados de libertad dados por (5), $Tv2$ para la prueba de Satterthwaite con grados de libertad dados por (7), Tcf para la prueba condicionada por los resultados de la comparación de varianzas con una prueba preliminar F y $TMinVp$ para el procedimiento que rechaza H_0 si alguno de los valores-p de la pruebas de Student o de Welch-Satterthwhite es inferior o igual a α.

Los resultados hablan por sí solos. Únicamente en el caso de varianzas iguales ($R = 1$) la prueba de Student (Tvi) muestra una muy ligera ventaja sobre la prueba de Welch-Satterthwaite ($Tv1$). En los demás casos, el valor de DCM crece con las diferencias de desviaciones estándar con excepción de las prueba de Welch y Satterthwaite que se mantienen estables, con ventaja para la expresión más conocida. A partir de $R = 2$ ya no se estudió el procedimiento del mínimo valor-p pues, desde el comienzo, los resultados mostraron su inconveniencia.

Tabla 2: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 1$.

	Tvi	Tv1	Tv2	Tcf	TMinVp
Min.	0.048050	0.046990	0.048260	0.047870	0.04837
1st Qu.	0.049650	0.049600	0.050210	0.049960	0.05289
Median	0.050080	0.050360	0.050980	0.050700	0.05760
Mean	0.050080	0.050600	0.051720	0.051050	0.05900
3rd Qu.	0.050530	0.051160	0.052320	0.051760	0.06338
Max.	0.051960	0.056360	0.059750	0.057090	0.07865
DCM	0.000697	0.001756	0.002962	0.001913	0.01156

Tabla 3: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 1.25$.

	Tvi	Tv1	Tv2	Tcf	TMinVp
Min.	0.02105	0.046700	0.048610	0.03397	0.04889
1st Qu.	0.03657	0.049570	0.050000	0.04132	0.05126
Median	0.05036	0.050190	0.050840	0.05025	0.05524
Mean	0.05270	0.050530	0.051610	0.05266	0.06272
3rd Qu.	0.06762	0.051000	0.052200	0.06126	0.06965
Max.	0.10070	0.055730	0.059310	0.09177	0.11600
DCM	0.01994	0.001738	0.002886	0.01400	0.02027

Tabla 4: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 1.5$.

	Tvi	Tv1	Tv2	Tcf	TMinVp
Min.	0.00810	0.047280	0.048330	0.02902	0.04828
1st Qu.	0.02890	0.049550	0.050090	0.04013	0.05036
Median	0.05105	0.050220	0.050740	0.05021	0.05400
Mean	0.05828	0.050460	0.051480	0.05450	0.06995
3rd Qu.	0.08279	0.051060	0.052060	0.06412	0.08307
Max.	0.14970	0.055890	0.059120	0.11200	0.15720
DCM	0.03673	0.001504	0.002602	0.01935	0.03412

Tabla 5: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 1.75$.

	Tvi	Tv1	Tv2	Tcf	TMinVp
Min.	0.00323	0.048020	0.048390	0.02836	0.04804
1st Qu.	0.02350	0.049600	0.050190	0.04306	0.05014
Median	0.05222	0.050270	0.050840	0.05049	0.05289
Mean	0.06463	0.050520	0.051450	0.05498	0.07767
3rd Qu.	0.09882	0.051230	0.052100	0.06251	0.09886
Max.	0.19920	0.054870	0.058410	0.11860	0.20340
DCM	0.05146	0.001468	0.002490	0.01925	0.04809

Tabla 6: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 2$.

	Tvi	Tv1	Tv2	Tcf
Min.	0.00155	0.046970	0.048340	0.0294
1st Qu.	0.01979	0.049600	0.050080	0.0458
Median	0.05246	0.050290	0.050760	0.0502
Mean	0.07059	0.050450	0.051290	0.0543
3rd Qu.	0.10930	0.050950	0.051680	0.0574
Max.	0.24730	0.054620	0.057730	0.1112
DCM	0.06411	0.001338	0.002277	0.0167

Tabla 7: *Comportamiento de las tasas de error tipo I de procedimientos de comparación de medias con razón de desviaciones estándar* $R = 3$.

	Tvi	Tv1	Tv2	Tcf
Min.	0.00011	0.048150	0.048520	0.037510
1st Qu.	0.01398	0.049580	0.049910	0.049270
Median	0.05448	0.050290	0.050620	0.050210
Mean	0.08728	0.050350	0.050920	0.051610
3rd Qu.	0.13720	0.050940	0.051510	0.051480
Max.	0.36480	0.054310	0.056960	0.074130
DCM	0.09702	0.001096	0.001736	0.006616

5. Conclusiones

1. La prueba de Student para dos muestras independientes con el supuesto de varianzas iguales pierde su capacidad de permitirle al investigador controlar la probabilidad de cometer el error tipo I cuando el supuesto no es válido, especialmente cuando las muestras son de tamaños diferentes.

2. Cuando la muestra de la población con menor dispersión es la de mayor tamaño, la proporción de rechazos con error tipo I es mayor que el nivel de significación dado por el investigador.

3. Cuando la muestra de la población con mayor dispersión es la de mayor tamaño, la proporción de rechazos con error tipo I tiende a ser menor que el nivel de significación dado por el investigador.

4. La prueba de Welch-Satterthwaite muestra en las simulaciones de Monte Carlo que las proporciones de rechazo equivocado de H_0 son muy cercanas del nivel de significación establecido α.

5. El cálculo de los grados de libertad mencionado por Winer (1971) como la aproximación de Satterthwaite es menos eficaz que el del Welch.

6. El uso de una prueba preliminar de homogeneidad de varianzas no estabiliza satisfactoriamente la proporción de rechazos equivocados de H_0 alrededor del nivel de significación dado.

6. Recomendaciones

1. Trabajar en la medida de lo posible con muestras de tamaños iguales.

2. Reducir en los cursos de estadística los espacios y tiempos dedicados a la prueba de Student para dos muestras independientes con el supuesto de varianzas iguales y dedicarlo a la prueba de Welch-Satterthwaite.

3. Evitar el uso exagerado de supuestos que pueden traer confusión. En la prueba Z para una muestra puede justificarse el supuesto de varianza conocida pues facilita introducir conceptos relacionados con la potencia de las pruebas. Pero, una vez logrado este propósito, se pasa rápidamente a la prueba t para ofrecer una herramienta de análisis más cercana de las condiciones reales de aplicación.

4. En las pruebas para dos muestras independientes, esos conceptos ya se conocen y se puede pasar a la prueba de Welch-Satterthwaite sin pasar por puntos intermedios que finalmente impiden sacar provecho a un tiempo valioso de los cursos para tratar otros temas.

Recibido: 8 de marzo de 2011
Aceptado: 1 de septiembre de 2011

Referencias

Brown, M. & Forsythe (1974), 'Robust Test for the Equality of Variances', *Journal of the American Statistical Association* **69**(346), 364–367.

Canavos, G. (1988), *Probabilidad y estadística: aplicaciones y métodos*, Mc-Graw Hill, México.

Conover, W., Johnson, M. E. & Johnson, M. (1981), 'A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Bidding Data', *Technometrics* **23**, 351–361.

Dodge, Y. (1985), *Analysis of Experiments with Missing Data*, John Wiley & Sons, New York.

Kim, S. H. & Cohen, A. S. (1995), 'On the Behrens-Fisher Problem: A Review'.

Marques de Sá, J. (2007), *Applied Statistics using SPSS, Statistica, Matlab and R*, Springer Verlag, Berlin.

Milliken, G. & Johnson, D. (1984), *Analysis of Messy Data*, Vol. I of *of Designed Experiments*, Van Nostrand Reinhold, New York.

Montilla, J., M. & Kromrey, J. (2010), 'Robustez de las pruebas T en comparación de medias, ante violación de supuestos de normalidad y homocedasticidad', *Revista Ciencia e Ingeniería* **31**(2), 101–108.

Newbold, P., Carlson, W. & Thorne, B. (2008), *Estadística para administración y economía*, Pearson Educación S.A., Madrid.

Park, H. M. (2009), 'Comparing Group Means: T-tests and One-way ANOVA Using Stata, SAS, R, and SPSS'.
*http://www.indiana.edu/~ statmath/stat/all/ttest

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Satterthwaite, F. E. (1946), 'An Approximate Distribution of Estimates of Variance Components', *Biometrics Bulletin* **2**(6), 110–114.

Sawilowsky, S. (2002), 'Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means When $\sigma_1^2 \neq \sigma_2^2$', *Journal of Modern Applied Statistical Methods* **1**(2).

Searle, S. R., Casella, G. & McCulloch, C. (1992), *Variance Components*, John Wiley & Sons, New York.

Welch, B. L. (1938), 'The significance of the difference between two means when the population variances are unequal', *Biometrika* **28**(3/4), 350–362.

Welch, B. L. (1947), 'The generalization of Student's problem when several different population variances are involved', *Biometrika* **34**(1/2), 28–35.

Winer, B. J. (1971), *Statistical Principles in Experimental Design*, Mc-Graw Hill, New York.

Zimmerman, D. W. & Zumbo, B. D. (2009), 'Hazards in choosing between pooled and separate-variances t tests', *Psicológica* **30**, 371–390.

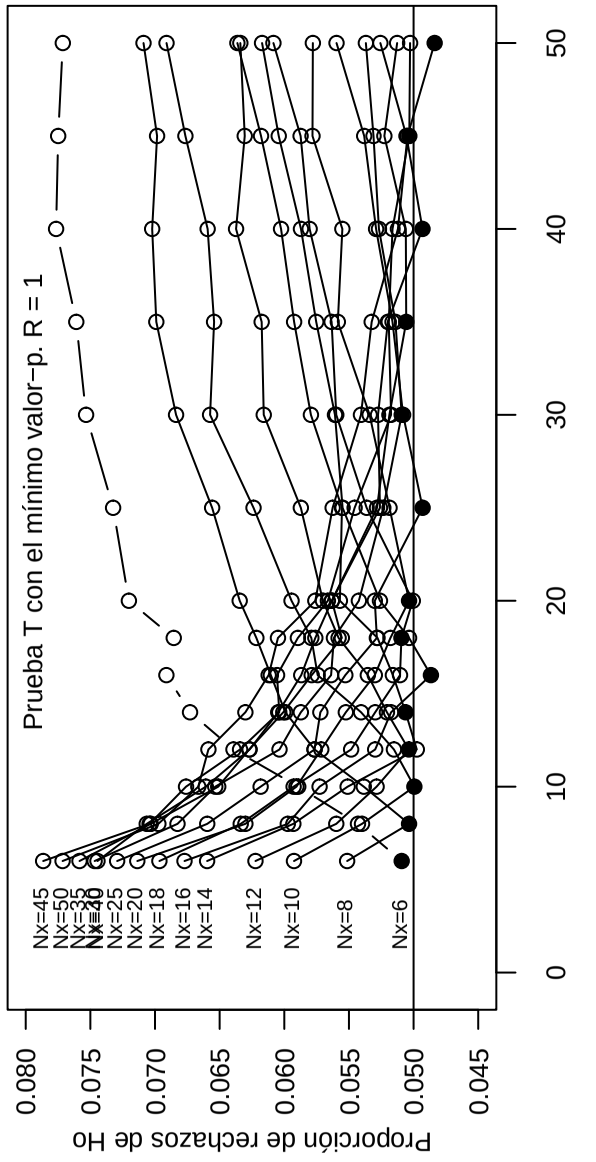

Figura 1: *Proporciones de error tipo I cuando se rechaza H_0 con el mínimo valor-p de las pruebas de Student y de Welch-Satterthwaite. Las varianzas poblacionales son iguales.*

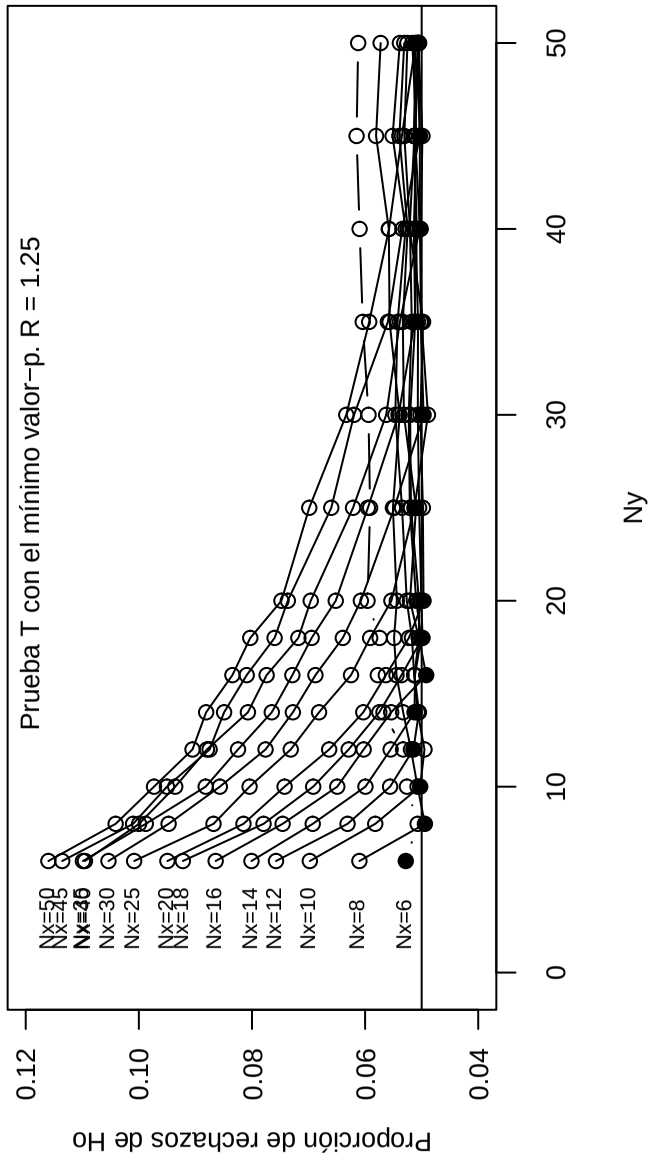

Figura 2: *Proporciones de error tipo I cuando se rechaza H_0 con el mínimo valor-p de las pruebas de Student y de Welch-Satterthwaite. La relación $R = 1.25$ indica que $\sigma_Y = 1.25\ \sigma_X$.*

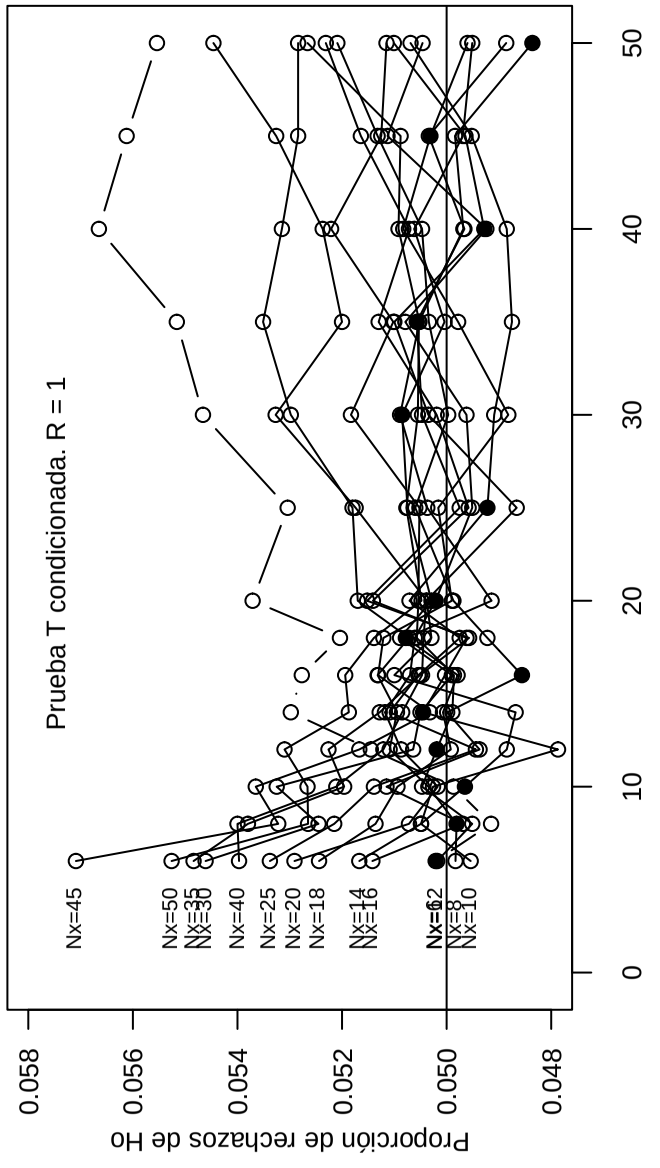

Figura 3: *Proporciones de error tipo I con pruebas combinadas de F para homocedasticidad y de Student o de Welch-Satterthwaite según el resultado preliminar. Varianzas poblacionales iguales.*

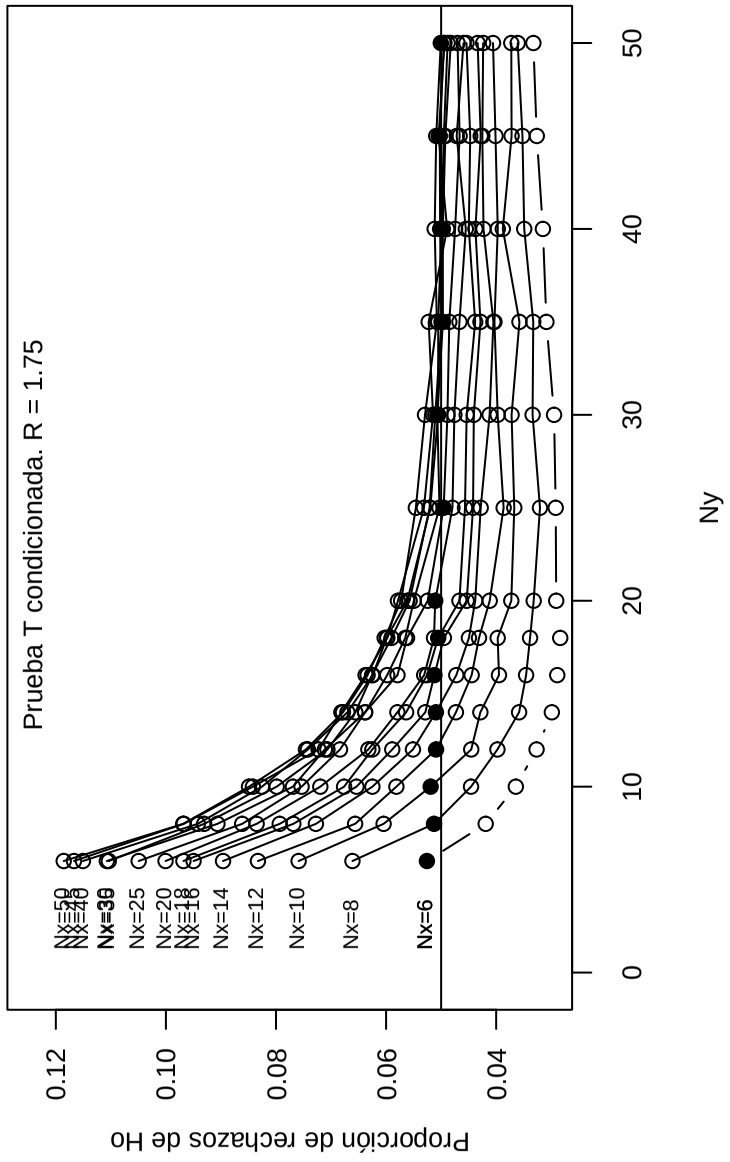

Figura 4: *Proporciones de error tipo I con pruebas combinadas de F para homocedasticidad y de Student o de Welch-Satterthwaite según el resultado preliminar.* $\sigma_Y = 1.75\ \sigma_X$.

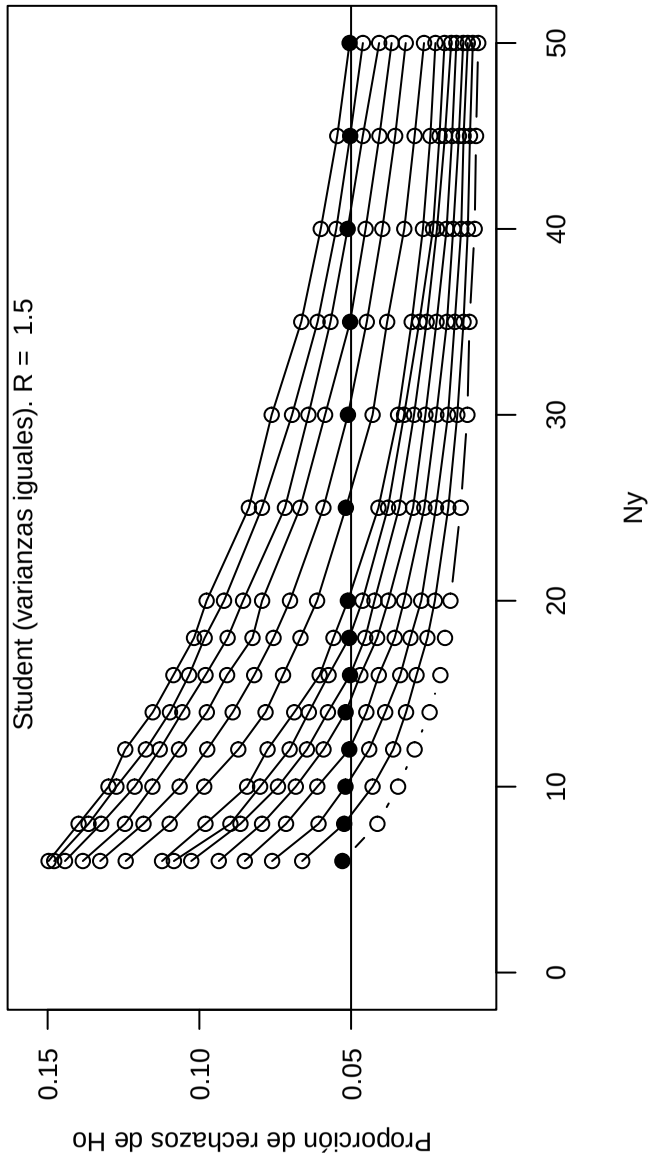

Figura 5: *Proporciones de error tipo I con pruebas combinadas de F para homocedasticidad y de Student o de Welch-Satterthwaite según el resultado preliminar.* $\sigma_Y = 1.5\ \sigma_X$.

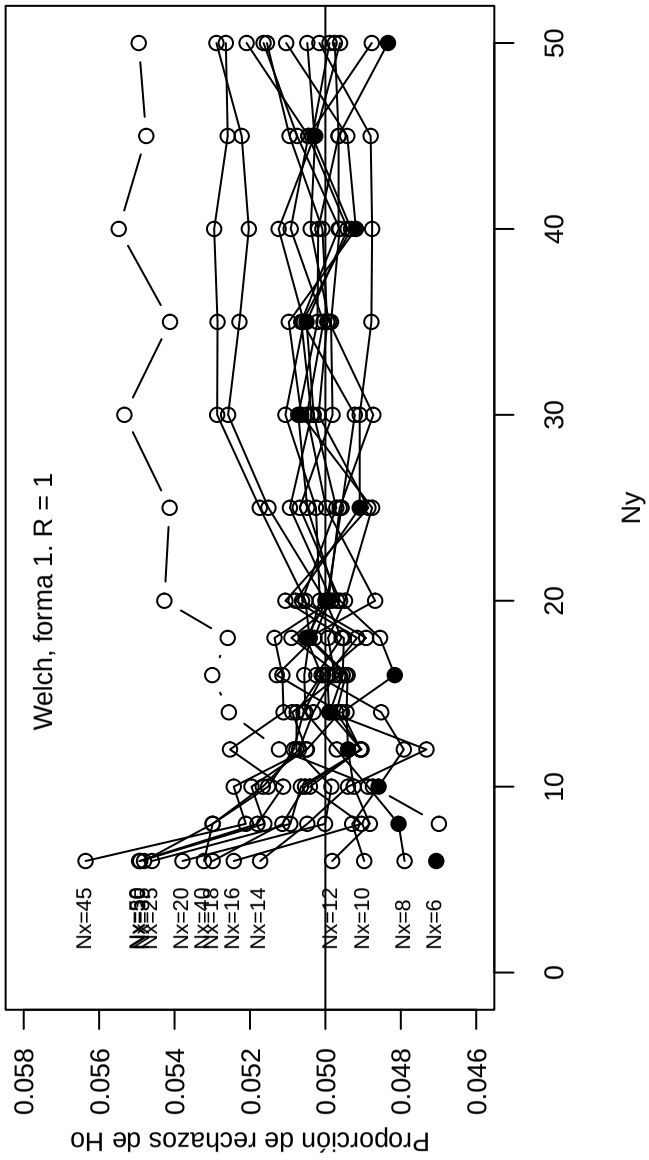

Figura 6: *Proporciones de error tipo I con la prueba de Welch-Satterthwaite. Varianzas iguales.*