Prueba t de 2 muestras

Introducción

Una prueba de 2 muestras se puede utilizar para comparar si las medias de dos grupos independientes son diferentes. Esta prueba se deriva bajo el supuesto de que ambas poblaciones están normalmente distribuidas y poseen varianzas iguales. Si bien el supuesto de normalidad no es crítico (Pearson, 1931; Barlett, 1935; Geary, 1947), el supuesto de varianzas iguales es crítico si los tamaños de las muestras son notablemente diferentes (Welch, 1937; Horsnell, 1953).

Algunos profesionales primero realizan una prueba preliminar para evaluar varianzas iguales antes de realizar el clásico procedimiento t de 2 muestras. Sin embargo, este enfoque presenta serias desventajas debido a que estas pruebas de varianzas están sujetas a supuestos y limitaciones importantes. Por ejemplo, numerosas pruebas de varianzas iguales, como la clásica prueba F, son sensibles a desviaciones con respecto a la normalidad. Otras pruebas que no se basan en el supuesto de normalidad, como la de Levene/Brown-Forsythe, tienen poca potencia para detectar una diferencia entre varianzas.

B.L. Welch desarrolló un método de aproximación para comparar las medias de dos poblaciones normales independientes cuando las varianzas no son necesariamente iguales (Welch, 1947). Debido a que la prueba t modificada de Welch no se deriva bajo el supuesto de varianzas iguales, los usuarios pueden comparar las medias de dos poblaciones sin primero tener que determinar la existencia de varianzas iguales.

En este documento, comparamos el método t modificado de Welch con el clásico procedimiento t de 2 muestras y determinamos cuál procedimiento es más fiable. También describimos las siguientes verificaciones de datos que se realizan y muestran de manera automática en la Tarjeta de informe del Asistente y explicamos cómo afectan los resultados del análisis:

- Normalidad

- Datos poco comunes
- Tamaño de la muestra

Método de la prueba de 2 muestras

Comparación entre la prueba de 2 muestras clásica y la prueba t de Welch

Si los datos provienen de dos poblaciones normales con las mismas varianzas, la prueba de 2 muestras clásica tiene la misma o más potencia que la prueba de Welch. El supuesto de normalidad no es crítico para el procedimiento clásico (Pearson, 1931; Barlett, 1935; Geary, 1947); sin embargo, el supuesto de varianzas iguales es importante para garantizar resultados válidos. Más específicamente, el procedimiento clásico es sensible al supuesto de varianzas iguales cuando difieren los tamaños de las muestras independientemente de qué tan grandes sean las muestras (Welch, 1937; Horsnell, 1953). En la práctica, sin embargo, rara vez se cumple el supuesto de varianzas iguales, lo cual puede producir tasas de error Tipo I más elevadas. Por lo tanto, si la prueba t de 2 muestras clásica se utiliza cuando dos muestras tienen varianzas diferentes, aumenta la probabilidad de que la prueba produzca resultados incorrectos.

La prueba t de Welch es una alternativa viable a la prueba t clásica debido que no asume varianzas iguales y, por lo tanto, no es sensible a varianzas desiguales para todos los tamaños de muestra. Sin embargo, la prueba t de Welch se basa en aproximaciones y su rendimiento con tamaños de muestra pequeños pudiera ser cuestionable. Queríamos determinar si la prueba t de Welch o la prueba t de 2 muestras clásica es la prueba más fiable y práctica para utilizar en el Asistente.

Objetivo

Queríamos determinar, a través de estudios de simulación y desviaciones teóricas, si la prueba t de Welch o la prueba t de 2 muestras clásica es más fiable. Más específicamente, queríamos examinar:

- Las tasas de error Tipo I y Tipo II de las pruebas t de 2 muestras y de Welch con diferentes tamaños de muestra cuando los datos están normalmente distribuidos y las varianzas son iguales.
- Las tasas de error Tipo I y Tipo II de la prueba t de Welch para diseños no balanceados con varianzas desiguales para los que falla la prueba t de 2 muestras clásica.

Método

Nuestras simulaciones se enfocaron en tres áreas:

- Comparamos los resultados de las pruebas simuladas de la prueba t de 2 muestras clásica y la prueba t de Welch bajo diferentes supuestos del modelo, incluida normalidad, no normalidad, varianzas iguales, varianzas desiguales y diseños balanceados y no balanceados. Para mayor información, véase el Apéndice A.

- Derivamos la función de potencia de la prueba t de Welch y la comparamos con la función de potencia de la prueba t de 2 muestras clásica. Para mayor información, véase el Apéndice B.
- Estudiamos el impacto de la no normalidad en la función de potencia teórica de la prueba t de Welch.

Resultados

Cuando se mantienen los supuestos del modelo t de 2 muestras clásico, la prueba t de Welch funciona tan bien o casi tan bien como la prueba t de 2 muestras clásica, excepto para diseños pequeños no balanceados. Sin embargo, la prueba t de 2 muestras clásica también puede funcionar deficientemente cuando los diseños son pequeños y no balanceados, debido al supuesto de varianzas iguales. Además, en configuraciones prácticas, es difícil establecer que dos poblaciones tengan exactamente la misma varianza. Por lo tanto, la superioridad teórica de la prueba t de 2 muestras clásica con respecto a la prueba t de Welch tiene poco o ningún valor práctico. Por esta razón, el Asistente utiliza la prueba t de Welch para comparar las medias de dos poblaciones. Para obtener los resultados detallados de las simulaciones, véanse los Apéndices A, B y C.

Verificaciones de datos

Normalidad

La prueba t de Welch, el método utilizado en el Asistente para comparar las medias de dos poblaciones independientes, se deriva bajo el supuesto de que las dos poblaciones están normalmente distribuidas. Afortunadamente, incluso cuando los datos no están normalmente distribuidos, la prueba t de Welch funciona adecuadamente si las muestras son suficientemente grandes.

Objetivo

Queríamos determinar qué tanto coincidían los niveles de significancia simulados del método de Welch y la prueba t de 2 muestras clásica con el nivel de significancia objetivo (tasa de error Tipo I) de 0.05.

Método

Realizamos simulaciones de la prueba t de Welch y la prueba t de 2 muestras clásica con 10,000 pares de muestras independientes generadas a partir de poblaciones normales, asimétricas y normales contaminadas (varianzas iguales y desiguales). Las muestras tenían diferentes tamaños. La población normal cumple la función de población de control para fines de comparación. Para cada condición, calculamos los niveles de significancia simulados y los comparamos con el nivel de significancia objetivo o nominal de 0.05. Si la prueba funciona adecuadamente, los niveles de significancia deberían encontrarse cerca 0.05.

Resultados

Para muestras moderadas o grandes, la prueba t de Welch mantiene sus tasas de error Tipo I para datos normales y no normales. Los niveles de significancia simulados se aproximan al nivel de significancia objetivo cuando el tamaño de las muestras es de por lo menos 15. Para mayor información, véase el Apéndice A.

Debido a que la prueba funciona correctamente con muestras relativamente pequeñas, el Asistente no prueba la normalidad de los datos. En cambio, se verifica el tamaño de las muestras y se muestran los siguientes indicadores de estado en la Tarjeta de informe:

Estado	Condición
	Los tamaños de las muestras son por lo menos 15, la normalidad no representa problema alguno.
	Por lo menos uno de los tamaños de las muestras < 15; la normalidad pudiera representar un problema.

Datos poco comunes

Los datos poco comunes son valores de datos extremadamente grandes o pequeños, también conocidos como valores atípicos. Los datos poco comunes pueden tener una fuerte influencia sobre los resultados del análisis. Cuando la muestra es pequeña, éstos pueden afectar las probabilidades de hallar resultados estadísticamente significativos. Los datos poco comunes pueden indicar problemas con la recolección de los datos o un comportamiento poco común del proceso. Por lo tanto, estos puntos de datos con frecuencia merecen investigarse y se deberían corregir cuando sea posible.

Objetivo

Queríamos desarrollar un método para verificar los valores de los datos que sean muy grandes o pequeños en relación con la muestra general y que pudieran afectar los resultados del análisis.

Método

Desarrollamos un método para verificar los datos poco comunes basándonos en el método descrito por Hoaglin, Iglewicz y Tukey (1986) para identificar los valores atípicos en las gráficas de caja.

Resultados

El Asistente identifica un punto de dato como poco común si supera en 1.5 el rango intercuartil posterior a los cuartiles inferior o superior de la distribución. Los cuartiles inferior y superior son los percentiles 25 y 75 de los datos. El rango intercuartil es la diferencia entre estos dos cuartiles. Este método funciona correctamente cuando existen múltiples valores atípicos, debido a que permite detectar cada valor atípico específico.

Los valores atípicos tienden a tener influencia sobre la función de potencia solo cuando los tamaños de las muestras son muy pequeños. En general, cuando existen valores atípicos, los valores de potencia observados tienden a ser un poco más elevados que los valores de potencia teórica objetivo. Este patrón se puede observar en la Figura 10 del Apéndice C, donde las curvas de potencia simulada y teórica no se aproximan de manera razonable hasta que el tamaño de la muestra mínimo llega a 15.

Cuando se verifica la presencia de datos poco comunes, la Tarjeta de informe del Asistente correspondiente a la prueba t de 2 muestras exhibe los siguientes indicadores de estado:

Estado	**Condición**
	No hay puntos de datos poco comunes.
	Por lo menos un punto de dato es poco común y pudiera afectar los resultados la prueba.

Tamaño de la muestra

Generalmente, se realiza una prueba de hipótesis para recolectar evidencia para rechazar la hipótesis nula de que "no existe diferencia". Si las muestras son muy pequeñas, la potencia de la prueba pudiera no ser adecuada para detectar una diferencia entre las medias cuando realmente exista una, lo cual produce un error Tipo II. Es por lo tanto crucial asegurarse de que los tamaños de las muestras sean lo suficientemente grandes para detectar diferencias parcialmente importantes con una alta probabilidad.

Objetivo

Si los datos actuales no proporcionan evidencia suficiente contra la hipótesis nula, queremos determinar si los tamaños de las muestras son suficientemente grandes para que la prueba detecte diferencias prácticas de interés con alta probabilidad. Si bien el objetivo de planificar los tamaños de las muestras es asegurar que éstas sean suficientemente grandes para detectar diferencias importantes con alta probabilidad, las muestras no deberían ser tan grandes que hagan que las diferencias despreciables se vuelvan estadísticamente significativas.

Método

El análisis de potencia y tamaño de la muestra se basa en la función de potencia teórica de la prueba específica que se utiliza para realizar el análisis estadístico. Para la prueba t de Welch, esta función de potencia depende de los tamaños de las muestras, la diferencia entre las medias de las poblaciones y las varianzas verdaderas de las dos poblaciones. Para mayor información, véase el Apéndice B.

Resultados

Cuando los datos no proporcionan evidencia suficiente contra la hipótesis nula, el Asistente calcula diferencias prácticas que se pueden detectar con una probabilidad del 80% y 90% para los tamaños de las muestras dados. Además, si el usuario proporciona una diferencia práctica de interés, el Asistente calcula los tamaños de muestra que ofrezcan una probabilidad del 80% y 90% de detectar la diferencia.

No hay un resultado general que informar debido a que los resultados dependen de las muestras específicas del usuario. Sin embargo, puede consultar los Apéndices B y C para obtener más información sobre la función de potencia de la prueba de Welch.

Cuando se verifican la potencia y el tamaño de la muestra, la Tarjeta de informe del Asistente correspondiente a la prueba t de 2 muestras exhibe los siguientes indicadores de estado:

Estado	Condición
	La prueba halla una diferencia entre las medias, de modo que la potencia no representa problema alguno. O La potencia es suficiente. La prueba no halló una diferencia entre las medias, pero la muestra es suficientemente grande para proporcionar por lo menos una probabilidad del 90% de detectar la diferencia especificada.
	La potencia pudiera ser suficiente. La prueba no halló una diferencia entre las medias, pero la muestra es lo suficientemente grande para proporcionar un probabilidad entre el 80% y el 90% de detectar la diferencia especificada. Se informa el tamaño de la muestra que se requiere para alcanzar una potencia del 90%.
	La potencia pudiera no ser suficiente. La prueba no halló una diferencia entre las medias, y la muestra es lo suficientemente grande para proporcionar una probabilidad entre el 60% y el 80% de detectar la diferencia especificada. Se informan los tamaños de las muestras que se requieren para alcanzar una potencia del 80% y una potencia del 90%.
	La potencia no es suficiente. La prueba no halló una diferencia entre las medias, y la muestra no es lo suficientemente grande para proporcionar una probabilidad de por lo menos el 60% de detectar la diferencia especificada. Se informan los tamaños de las muestras que se requieren para alcanzar una potencia del 80% y una potencia del 90%.
	La prueba no halló una diferencia entre medias. No se especificó la detección de una diferencia práctica entre las medias; por lo tanto, el informe indica las diferencias que se pudieran detectar con una probabilidad del 80% y 90%, con base en los tamaños de la muestras, las desviaciones estándar y alfa.

Referencias

Arnold, S. F. (1990). *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Aspin, A. A. (1949). Tables for Use in Comparisons whose Accuracy Involves Two Variances, Separately Estimated, *Biometrika, 36*, *290-296*.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, *31*, *223-231*.

Box, G. E. P. (1953). Non-normality and Tests on Variances, *Biometrika, 40*, *318-335*.

Geary, R. C. (1947). Testing for Normality, *Biometrika, 34*, *209-242*.

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, *81*, *991-999*.

Horsnell, G. (1953). The effect of unequal group variances on the F test for homogeneity of group means. *Biometrika, 40*, *128-136*.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the populations variances are unknown, *Biometrika, 38*, *324-329*.

Kulinskaya, E. Staudte, R. G. and Gao, H. (2003). Power Approximations in Testing for unequal Means in a One-Way Anova Weighted for Unequal Variances, *Communication in Statistics, 32(12), 2353-2371*.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley.

Neyman, J., Iwaszkiewicz, K. & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation, *Journal of the Royal Statistical Society, Series B, 2, 107-180*.

Pearson, E. S. (1931). The Analysis of variance in case of non-normal variation, *Biometrika, 23*, *114-133*.

Pearson, E.S. & Hartley, H.O. (Eds.). (1954). *Biometrika Tables for Statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of t-test, *Biometrika*, *45, 421-429*.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika, 38*, *330-336*.

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal, *Biometrika, 29*, *350-362*.

Wolfram, S. (1999). *The Mathematica Book (*4th ed.). Champaign, IL: Wolfram Media/Cambridge University Press.

Apéndice A: Impacto de la no normalidad y la heterogeneidad sobre la prueba t de 2 muestras clásica y la prueba t de Welch

Realizamos diferentes estudios de simulación diseñados para comparar la prueba t de 2 muestras clásica con la prueba t de Welch bajo diferentes supuestos del modelo.

Estudio de simulación A

Realizamos el estudio en tres partes:

- En la primera parte del estudio, exploramos la sensibilidad de la prueba t de 2 muestras clásica y de la prueba t de Welch al supuesto de igualdad de varianza cuando se cumple el supuesto de normalidad. Las dos muestras provienen de poblaciones normales independientes. La primera muestra, la muestra base, se extrajo de una población normal con una media de 0 y desviación estándar de $\sigma_1 = 2$, $N(0,2)$. La segunda muestra también se extrajo de una población normal con media de 0, pero con una desviación estándar de σ_2 para una relación $\rho = \sigma_2/\sigma_1$ 0.5, 1.0, 1.5 y 2. En otras palabras, las segundas muestras se extrajeron de las poblaciones $N(0,1)$, $N(0,2)$, $N(0,3)$ $N(0,4)$, respectivamente. Además, el tamaño de la muestra base, en cada caso, se fijó en $n_1 = 5, 10, 15, 20$ y para cada n_1 especificado, el tamaño de la segunda muestra, n_2, se eligió de modo tal que la relación de los tamaños de las muestras, $r = n_2/n_1$, fuera aproximadamente igual a 0.5, 1, 1.5 y 2.0.

  Para cada uno de estos diseños de 2 muestras, generamos 10,000 pares de muestras independientes a partir de las poblaciones respectivas. A continuación realizamos la prueba t de 2 muestreas clásica y la prueba t de Welch a cada uno de los 10,000 pares de muestras para probar la hipótesis nula de que no existe diferencia entre las medias. Debido a que la verdadera diferencia entre las medias es nula, la fracción de las 10,000 réplicas por la que se rechazó la hipótesis nula representa el nivel de significancia simulado de la prueba. Debido a que el nivel de significancia objetivo para cada una de las pruebas es $\alpha = 0.05$, el error de simulación asociado con cada prueba y cada experimento es aproximadamente 0.2%.

- En la segunda parte, investigamos el impacto de la no normalidad, específicamente la asimetría, sobre los niveles de significancia simulados de las dos pruebas. Esta simulación se configuró del mismo modo que la anterior, con la excepción de que la muestra base se extrajo de la distribución de chi-cuadrado con 2 grados de libertad, Chi(2) y las segundas muestras se extrajeron de otras distribuciones de chi-cuadrado de modo que

$\rho = \sigma_2/\sigma_1$ asuma los valores 0.5, 1.0, 1.5 y 2. Se estableció que la diferencia hipotética entre las medias sería la verdadera diferencia entre las medias de las poblaciones originales.

- En la tercera parte, examinamos el efecto de los valores atípicos sobre el rendimiento de las dos pruebas t. Por esta razón, las dos muestras se extrajeron de distribuciones normales contaminadas. Una población normal contaminada $\mathrm{CN}(\mathrm{p}, \sigma)$ es la mezcla de dos poblaciones normales: la población $\mathrm{N}(0,1)$ y la población $\mathrm{N}(0, \sigma)$ normal. Definimos una distribución normal contaminada como:

$$CN(p, \sigma) = pN(0,1) + (1 - p)N(0, \sigma)$$

donde p es el parámetro de mezcla y $1 - \mathrm{p}$ es la proporción de contaminación proporción de valores atípicos. Es fácil demostrar que si X se distribuye como $CN(p, \sigma)$, entonces la media es $\mu_X = 0$ y su desviación estándar es $\sigma_X = \sqrt{p + (1 - p)\sigma^2}$.

La muestra base se extrajo de $CN(.8, 4)$ y la segunda muestra se extrajo de la $CN(.8, \sigma)$ normal contaminada. Se eligió el parámetro σ para que la relación de las desviaciones estándar de las dos poblaciones (contaminadas) $\rho = \sigma_2/\sigma_1$ fuera igual a 0.5, 1.0, 1.5 y 2, tal como las partes I y II. Debido a $\sigma_1 = \sqrt{.8 + (1 - .8) * 16} = 2.0$, es necesario elegir $\sigma = 1, 4, 6.40, 8.72$, respectivamente. En otras palabras, las segundas muestras se extrajeron de $CN(.8, 1)$, $CN(.8, 4)$, $CN(.8, 6.4)$ y $CN(.8, 8.72)$. A continuación, realizamos las simulaciones descritas en la Parte I.

Los resultados del estudio se encuentran organizados en la Tabla 1 y se muestran en las Figuras 1, 2 y 3.

Resultados y resumen

En general, los resultados de las simulaciones respaldan los resultados teóricos de que, bajo el supuesto de normalidad y varianzas iguales, la prueba t de 2 muestras clásica produce niveles de significancia que se aproximan al nivel objetivo, incluso cuando los tamaños de las muestras son pequeños. La segunda columna de gráficas en la Figura 1 muestra los niveles de significancia simulados en los diseños en los que las varianzas de las dos poblaciones normales son iguales. Las curvas de los niveles de significancia simulados basados en la prueba t de 2 muestras clásica no se distinguen de las líneas de los niveles objetivo.

Las tablas de abajo muestran los niveles de significancia simulados de pruebas bilaterales tanto para la prueba t de 2 muestras clásica como para la prueba t de Welch, cada una con $\alpha = 0.05$ que se basa en pares de muestras provenientes de una población normal, poblaciones asimétricas (Chi-cuadrado) y poblaciones normales contaminadas. Los pares de muestras provienen de la misma familia de distribución, pero las varianzas de las respectivas poblaciones originales no son necesariamente iguales.

Tabla 1 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch, cada una con $\alpha = 0.05$) para n = 5.

			Pbción. base: N(0,2) 2da pbción.: N(0, σ_2)				Población base: Chi(2) 2da pbción.: Chi-cuadrado				Pbción. base: CN(.8,4) 2da pbción.: CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	.6	2T	.035	.050	.079	.105	.058	.042	.078	.113	.031	.036	.035	.034
		Welch	.035	.039	.049	.055	.048	.029	.055	.063	.029	.024	.021	.020
5	1.0	2T	.061	.052	.054	.058	.086	.036	.054	.064	.035	.031	.025	.023
		Welch	.048	.042	.044	.047	.066	.021	.040	.050	.027	.023	.018	.016
8	1.6	2T	.096	.048	.033	.027	.133	.041	.033	.032	.059	.037	.029	.024
		Welch	.050	.045	.043	.042	.094	.034	.032	.041	.034	.029	.026	.022
10	2.0	2T	.118	.055	.034	.025	.139	.041	.028	.024	.073	.041	.028	.023
		Welch	.052	.051	.050	.051	.097	.041	.033	.042	.035	.032	.028	.025

Tabla 2 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch, cada una con $\alpha = 0.05$) para n = 10.

			Pbción. base: N(0,2) 2da pbción.: N(0, σ_2)				Población base: Chi(2) 2da pbción.: Chi-cuadrado				Pbción. base: CN(.8,4) 2da pbción.: CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	.5	2T	.020	.050	.081	.112	.039	.044	.091	.123	.021	.035	.045	.047
		Welch	.046	.048	.050	.050	.043	.047	.067	.063	.034	.028	.022	.019
10	1.0	2T	.057	.051	.053	.055	.068	.044	.053	.054	.043	.042	.037	.032
		Welch	.051	.049	.049	.049	.062	.037	.046	.049	.039	.038	.032	.027
15	1.5	2T	.088	.048	.034	.029	.100	.043	.032	.032	.064	.040	.028	.021
		Welch	.050	.048	.047	.048	.074	.044	.041	.046	.035	.037	.035	.031

			Pbción. base: N(0,2) 2da pbción.: N(0, σ_2)				Población base: Chi(2) 2da pbción.: Chi-cuadrado				Pbción. base: CN(.8,4) 2da pbción.: CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
20	2	2T	.110	.048	.026	.019	.133	.042	.026	.022	.093	.046	.029	.019
		Welch	.048	.047	.045	.046	.083	.050	.044	.049	.036	.039	.040	.038

Tabla 3 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch, cada una con $\alpha = 0.05$) para n = 15.

			Pbción. base: N(0,2) 2da pbción.: N(0, σ_2)				Población base: Chi(2) 2da pbción.: Chi-cuadrado				Pbción. base: CN(.8,4) 2da pbción.: CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	.53	2T	.021	.050	.083	.110	.036	.041	.089	.114	.022	.044	.056	.062
		Welch	.050	.051	.051	.050	.047	.049	.067	.062	.044	.036	.027	.022
15	1.0	2T	.049	.047	.050	.053	.064	.046	.051	.061	.045	.045	.041	.037
		Welch	.045	.046	.049	.048	.060	.042	.048	.057	.042	.043	.039	.033
23	1.53	2T	.081	.049	.033	.028	.103	.042	.036	.030	.075	.048	.033	.024
		Welch	.048	.049	.048	.050	.071	.042	.048	.050	.042	.045	.044	.041
30	2.0	2T	.111	.050	.028	.018	.123	.049	.027	.020	.100	.046	.025	.016
		Welch	.049	.051	.051	.053	.074	.056	.045	.047	.039	.044	.042	.040

Tabla 4 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch, cada una con $\alpha = 0.05$) para n = 20.

			Pbción. base: N(0,2) 2da pbción.: N(0, σ_2)				Población base: Chi(2) 2da pbción.: Chi-cuadrado				Pbción. base: CN(.8,4) 2da pbción.: CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	.5	2T	.019	.052	.087	.115	.028	.048	.087	.119	.021	.048	.067	.079
		Welch	.050	.054	.053	.053	.044	.054	.061	.061	.048	.042	.035	.028
20	1.0	2T	.048	.049	.052	.053	.057	.046	.052	.056	.049	.044	.042	.040
		Welch	.045	.049	.051	.050	.055	.044	.050	.052	.047	.042	.040	.037
30	1.5	2T	.086	.054	.039	.032	.098	.047	.035	.033	.075	.047	.033	.022
		Welch	.054	.054	.053	.052	.068	.047	.051	.053	.041	.043	.044	.042
40	2.0	2T	.107	.049	.026	.016	.123	.046	.027	.019	.107	.047	.026	.016
		Welch	.048	.049	.046	.047	.070	.054	.046	.045	.044	.043	.043	.042

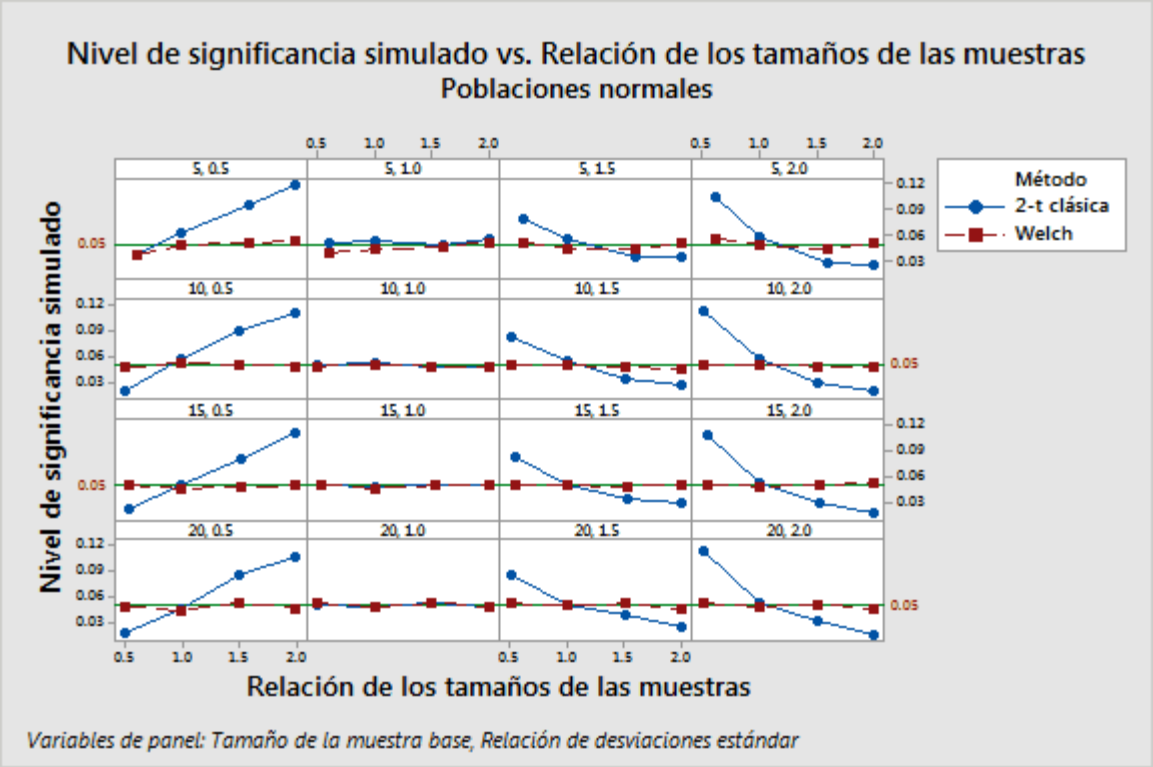

Figura 1 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch, cada una con $\alpha = 0.05$) con base en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

Los resultados de las simulaciones demuestran que, para muestras relativamente pequeñas, la prueba t de 2 muestras clásica, es robusta ante la no normalidad, pero es sensible al supuesto de varianzas iguales, a menos que el diseño de dos muestras esté casi balanceado. Esto se demuestra gráficamente en las Figuras 1, 2 y 3. Las curvas de los niveles de significancia simulados con base en la prueba t de 2 muestras clásica atraviesan la línea de los niveles objetivo en el punto donde la relación de los tamaños de las muestras es 1.0, incluso cuando las varianzas son muy diferentes. Para las tres familias de distribuciones (poblaciones normales, de chi-cuadrado y normales contaminadas), si los tamaños de las muestras son diferentes, los niveles de significancia simulados de la prueba t de 2 muestras clásica se aproximan al nivel objetivo solo cuando las varianzas son iguales. Esto se evidencia en la segunda columna de gráficas en cada una de las figuras 1, 2 y 3.

El rendimiento de la prueba t clásica es indeseable cuando el diseño no está balanceado y las varianzas son desiguales. Incluso las diferencias pequeñas entre las varianzas son problemáticas. Para los diseños no balanceados con varianzas desiguales, la normalidad de losdatos no mejora los niveles de significancia simulados. De hecho, los niveles de significancia simulados se ubican lejos del nivel objetivo a medida que aumentan los tamaños de las muestras, independientemente de la población original. Cuando se extrae la muestra más grande de la

población con mayor varianza, los niveles de significancia simulados son más pequeños que el nivel objetivo. Cuando se extraen las muestras más grandes de la población con menor varianza, los niveles simulados son mayores que los niveles objetivo. Arnold (1990, page 372) hace una observación similar cuando se examina la distribución asintótica del estadístico de la prueba t de 2 muestras clásica bajo el supuesto de varianzas desiguales.

La prueba t de 2 muestras de Welch, por el contrario, no es sensible a desviaciones del supuesto de igualdad de varianzas, tal como se ilustra en las Figuras 1, 2 y 3. Esto no es sorprendente debido a que la prueba t de Welch no se deriva del supuesto de varianzas iguales. El supuesto de normalidad del cual se deriva la prueba t de Welch pareciera ser importante solo cuando el mínimo de los dos tamaños de las muestras es muy pequeño. Para las muestras más grandes, sin embargo, la prueba se vuelve inmune a las desviaciones del supuesto de normalidad. Esto se ilustra en las Figuras 2 y 3, donde los niveles de significancia simulados permanecen constantemente cerca del nivel objetivo cuando el tamaño mínimo de las dos muestras es 15. Cuando ambas muestras provienen de una distribución de chi-cuadrado con 2 grados de libertad y el tamaño de ambas muestras es 15, el nivel de significancia simulado es de 0.042 (véase la Tabla 3).

Los valores atípicos tampoco parecieran afectar el rendimiento de la prueba t de Welch cuando el tamaño mínimo de las dos muestras es suficientemente grande. La Tabla 3 y Figura 3 muestran que cuando el tamaño mínimo de las dos muestras es por lo menos 15, entonces los niveles de significancia simulados se aproximan al nivel objetivo (los niveles de significancia simulados son 0.045, 0.045, 0.041, 0.037 cuando la relación de las desviaciones estándar es 0.5, 1.0, 1.5 y 2.0 respectivamente).

Estos resultados demuestran que para la mayoría de los efectos prácticos, la prueba t de 2 muestras de Welch tiene un mejor rendimiento que la prueba t de 2 muestras clásica en términos de sus niveles de significancia o tasa de error Tipo I.

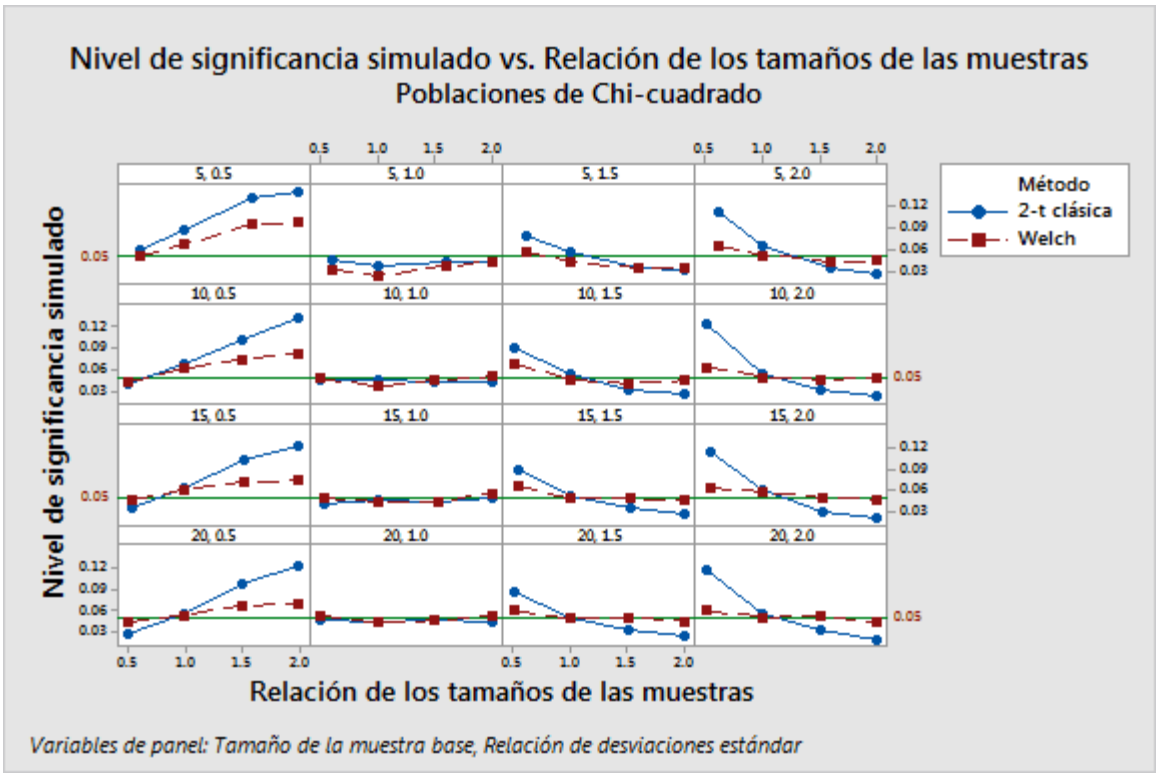

Figura 2 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch) con base en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

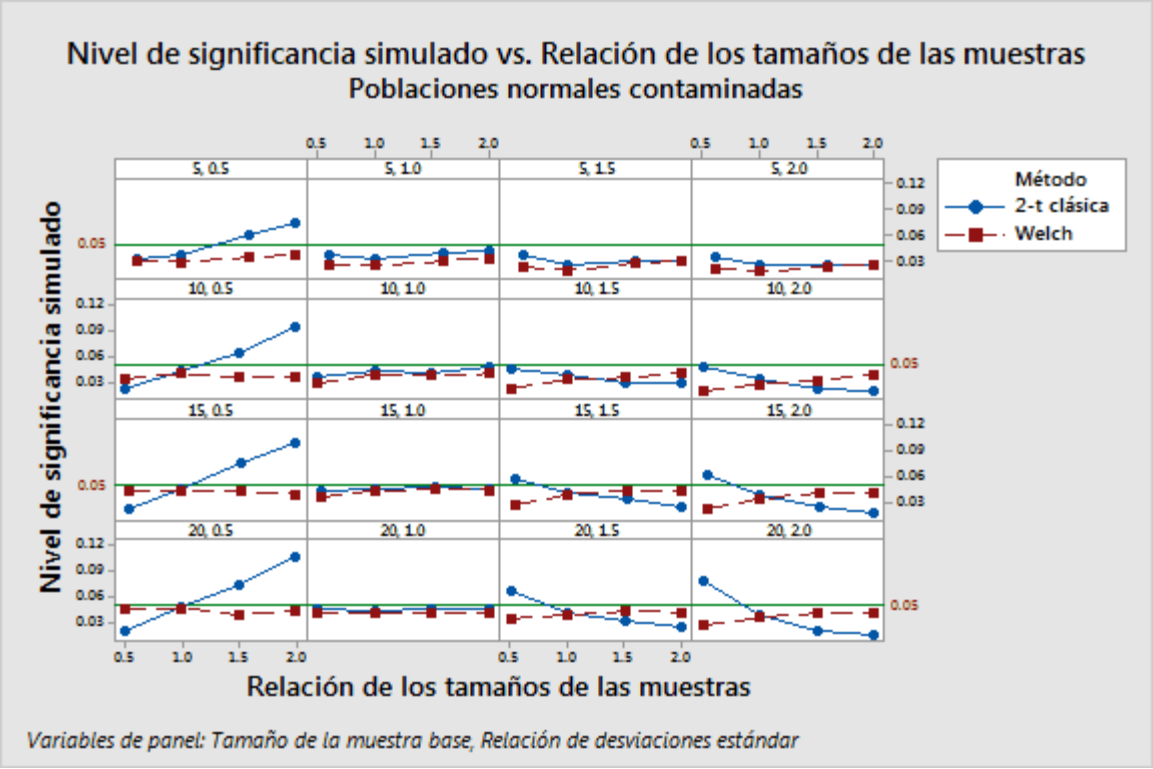

Figura 3 Niveles de significancia simulados de pruebas bilaterales (prueba t de 2 muestras clásica y prueba t de Welch) con base en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

Apéndice B: Comparación de las funciones de potencia de las dos pruebas

Queríamos determinar las condiciones bajo las cuales la función de potencia de la prueba t de Welch pudiera ser igual o aproximadamente igual a la función de potencia de la prueba t de 2 muestras.

En general, las funciones de potencia de las pruebas t (de 1 o 2 muestras) son bien conocidas y discutidas en numerosas publicaciones (Pearson y Hartley, 1952; Neyman et al., 1935; Srivastava, 1958). El siguiente teorema establece la función de potencia para cada una de las tres diferentes hipótesis alternativas en diseños de dos muestras.

TEOREMA B1

Bajo los supuestos de normalidad y la igualdad de las varianzas, la función de potencia de una prueba de dos muestras bilateral con un tamaño nominal de α se puede expresar como una función de los tamaños de las muestras y la $\delta = \mu_1 - \mu_2$ de la diferencia como

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha/2}\right) + F_{d_C,\lambda}\left(-t_{d_C}^{\alpha/2}\right)$$

donde $F_{d_C,\lambda}(.)$ es la C.D.F de la distribución t no central con $d_C = n_1 + n_2 - 2$ grados de libertad y un parámetro de no centralidad

$$\lambda = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Además, la función de potencia asociada con la hipótesis alternativa $\mu_1 > \mu_2$ se expresa como

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha}\right)$$

Sin embargo, cuando se realizan pruebas en función de la $\mu_1 < \mu_2$ alternativa, la potencia se expresa como

$$\pi(n_1, n_2, \delta) = F_{d_C,\lambda}\left(-t_{d_C}^{\alpha}\right)$$

Si bien se conoce ampliamente el resultado del teorema anterior, la función de potencia de la prueba basada en la prueba t de Welch modificada no se discute específicamente en la literatura. Se puede deducir una aproximación a partir de la función de potencia aproximada del modelo ANOVA de un solo factor (véase Kulinskaya et. al, 2003). Desafortunadamente, la función de potencia solo se puede aplicar a alternativas bilaterales. Sin embargo, el diseño de dos muestras es un caso tan especial que se puede adoptar un enfoque diferente para obtener la función de potencia (exacta) de la prueba t de Welch para cada una de las tres alternativas. Estas funciones se expresan en el siguiente teorema.

TEOREMA B2

Bajo el supuesto de que las poblaciones están normalmente distribuidas (aunque no necesariamente con la misma varianza), la función de potencia de una prueba t bilateral de Welch, con un tamaño nominal α, se puede expresar como una función de los tamaños de las muestras y la $\delta = \mu_1 - \mu_2$ de la diferencia como

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

donde $G_{d,\lambda}(.)$ es la C.D.F de la distribución t no central con d_W grados de libertad que se expresa como

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

y un parámetro de no centralidad

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Para las alternativas unilaterales, las funciones de potencia se expresan como

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha}\right)$$

y

$$\pi_W(n_1, n_2, \delta) = G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha}\right)$$

para probar la hipótesis nula en función de la $\mu_1 > \mu_2$ alternativa y para probar la hipótesis nula en función de $\mu_1 < \mu_2$ alternativa, respectivamente.

La prueba del resultado se proporciona en el Apéndice D.

Antes de comparar estas dos funciones de potencia, tenga en cuenta que, debido a que la prueba t de 2 muestras clásica se deriva bajo el supuesto adicional de que las varianzas de las poblaciones son iguales, las funciones de potencia teórica de las dos pruebas se deben comparar cuando el segundo supuesto sea válido para la prueba t de Welch.

En teoría, sabemos que bajo los supuestos de normalidad e iguales varianzas,

$\pi(n_1, n_2, \delta) \geq \pi_W(n_1, n_2, \delta)$ para todo n_1, n_2, δ

Los siguientes resultados indican las condiciones bajo las cuales las dos funciones son (aproximadamente) iguales.

TEOREMA B3

Bajo los supuestos de normalidad e igualdad de varianzas, tenemos lo siguiente:

1. Si $n_1 \sim n_2$ entonces $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ para cada δ de la diferencia. En particular, si $n_1 = n_2$ entonces $\pi(n_1, n_2, \delta) = \pi_W(n_1, n_2, \delta)$ para cada δ de la diferencia, de modo que la prueba t de Welch tiene la misma potencia que la prueba t de 2 muestras clásica.

2. Si n_1 y n_2 son pequeños y $n_1 \neq n_2$, entonces la prueba t de Welch tiene menos potencia que la prueba t de 2 muestras clásica. Sin embargo, si n_1 y n_2 son grandes, entonces $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ (independientemente de la diferencia entre los tamaños de las muestras).

La prueba del resultado se proporciona en el Apéndice E.

Bajo el supuesto de igualdad de varianzas, los parámetros de no centralidad asociados con las funciones de potencia de las dos pruebas son idénticos. La diferencia entre las funciones de potencia solo se pueden atribuir a la diferencia entre sus respectivos grados de libertad. Según la teoría, sabemos que bajo los supuestos especificados, la prueba t clásica es UPM (uniformemente máspotente) y, por lo tanto, tiene grados de libertad más elevados. El punto de los resultados anteriores, sin embargo, es que si el diseño está balanceado o aproximadamente balanceado, entonces las funciones de potencia son idénticas o aproximadamente idénticas. El único caso donde la prueba t clásica es notoriamente más potente que la prueba t de Welch es cuando el diseño se encuentra notablemente desbalanceado y las muestras son pequeñas. Desafortunadamente, este también pareciera ser el caso donde la prueba t de 2 muestras clásica es particularmente sensible al supuesto de igualdad de varianzas, tal como se evidencia en el Apéndice A. Como resultado, la función de potencia de la prueba t de Welch es la función más fiable para fines prácticos.

Ilustramos los resultados del Teorema B3 a través del siguiente ejemplo, donde las dos distribuciones normales tienen la misma desviación estándar de 3. Los valores de potencia basados en las funciones de potencia (bilateral) del Teorema B1 y del Teorema B2 se calculan según los siguientes cuatro escenarios:

1. Ambas muestras son pequeñas, pero el tamaño de la muestra es pequeño ($n_1 = n_2 = 10$).
2. Ambas muestras son pequeñas, pero una muestra es el doble que la otra ($n_1 = 10, n_2 = 20$).
3. Una muestra es pequeña y la otra tiene un tamaño moderado, pero esta última es cuatro veces más grande la muestra más pequeña $n_1 = 10, n_2 = 40$).
4. Una muestra tiene un tamaño moderado y la otra es grande, pero esta última es cuatro veces más grande que la muestra moderada ($n_1 = 50, n_2 = 200$).

Asumimos $\alpha = 0.05$ para ambas pruebas y se evalúan las funciones de potencia en cada escenario con la $\delta = 0.0, 0.5, 1.0, 1.5, 2.0, \ldots 5.0$ de la diferencia. Los resultados se muestran en la Tabla 5 y las funciones están graficadas en la Figuras 4.

Tabla 5 Comparación de las funciones de potencia teórica de las pruebas t de 2 muestras clásicas bilaterales y las pruebas t de Welch bilaterales, $\alpha = 0.05$. Los tamaños de las muestras, n_1 y n_2, son fijos y las funciones de potencia se evalúan con distintas diferencias δ, que van desde 0.0 hasta 5.0.

$\boldsymbol{\delta}$	0.0	0.5	1.0	1.5	2.0	2.5	3	3.5	4	4.5	5.0
$\boldsymbol{n_1 = n_2 = 10}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	.05	.064	.109	.185	.292	.422	.562	.694	.805	.887	.941
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	.05	.064	.109	.185	.292	.422	.562	.694	.805	.887	.941
$\boldsymbol{n_1 = 10, n_2 = 20}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	.05	.070	.132	.239	.383	.547	.703	.828	.913	.962	.986
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	.05	.070	.129	.231	.371	.531	.686	.813	.902	.955	.982
$\boldsymbol{n_1 = 10, n_2 = 40}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	.05	.075	.152	.283	.455	.637	.791	.899	.959	.986	.996
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	.05	.072	.142	.261	.419	.592	.748	.865	.938	.976	.992
$\boldsymbol{n_1 = 50, n_2 = 200}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	.05	.182	.556	.883	.987	.999	1.	1.	1.	1.	1.
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	.05	.180	.548	.877	.986	.999	1.	1.	1.	1.	1.

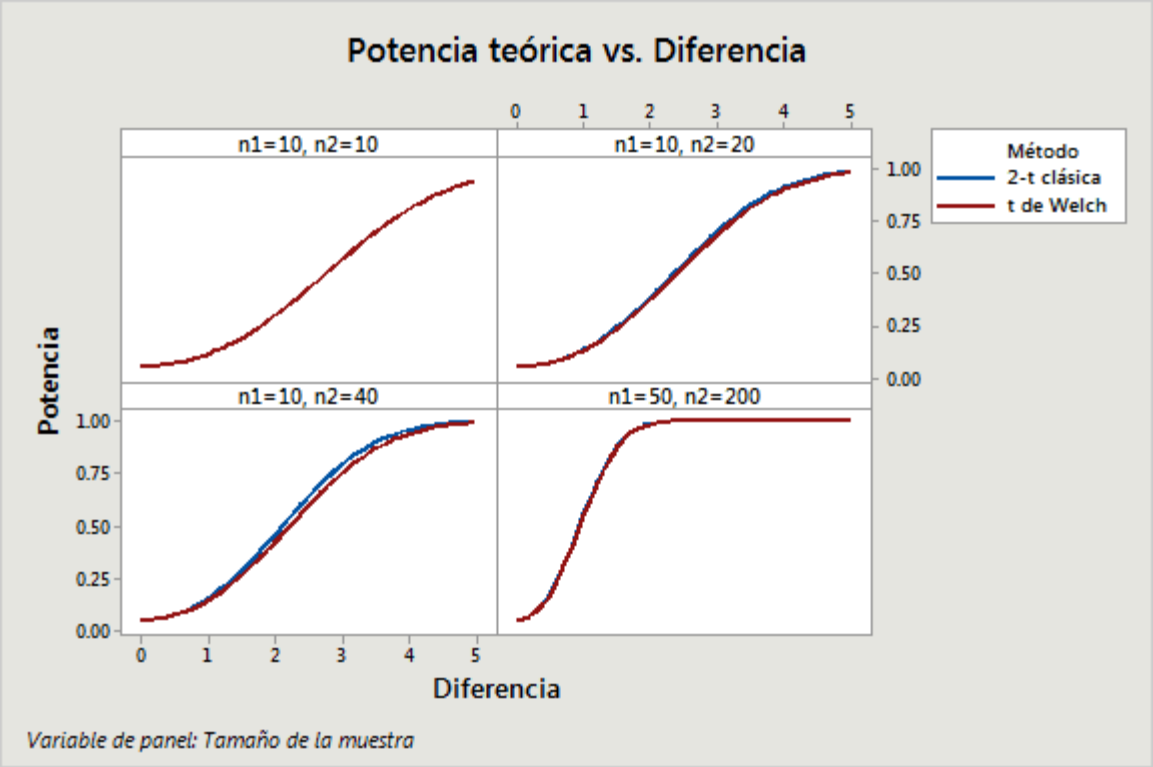

Figura 4 Gráficas de las funciones de potencia teórica de las pruebas t de 2 muestras clásicas bilaterales y las pruebas t de Welch bilaterales considerando δ la diferencia entre las medias que se detectarán. Ambas pruebas utilizan $\alpha = 0.05$. Las poblaciones del supuesto son normales con la misma desviación estándar de 3.

Estudio de simulación B

El propósito de este estudio de simulación es comparar los niveles de potencia asociados con la prueba t de 2 muestras clásica con los niveles de potencia asociados con la prueba t de 2 muestras de Welch en diseños balanceados donde se parte del supuesto de que las varianzas no son iguales. Los experimentos de estos estudios son similares a los discutidos en el Apéndice A.

En el primer grupo de experimentos, generamos pares de muestras de igual tamaño provenientes de poblaciones normales con varianzas desiguales. La población base se fijó en $N(0,2)$ y las segundas poblaciones normales se eligieron de modo tal que la relación de las desviaciones estándar $\rho = \sigma_2/\sigma_1$ sea igual a 0.5, 1.5 y 2. Del mismo modo, en un segundo grupo, las dos muestras se extrajeron de distribuciones de chi-cuadrado con varianzas desiguales (la población base es Chi(2)). En el último conjunto de experimentos, los pares de muestras se generaron a partir de la distribución normal contaminada (población base CN(.8,4)), tal como se definió anteriormente en el Apéndice A.

Para cada conjunto de experimentos, calculamos los niveles de potencia simulados (a una diferencia detectable δ específica) asociados con cada prueba para los tamaños de muestra $n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$. En cada experimento, el nivel de potencia simulado se calculó

como la proporción de veces que se rechazó la hipótesis nula cuando era falsa. Para todos los experimentos, la diferencia entre las medias se especificó en una unidad de la estándar en la población base (la primera de las dos muestras). Más específicamente, establecimos $\delta = 1.0 \times \sigma_1 = 2.0$, debido a que es relativamente pequeña para las tres familias de distribuciones de este estudio. Los resultados de las simulaciones se proporcionan en la Tabla 2.2 y se exhiben en la Figura 2.2a, Figura 2.2b y Figura 2.2c.

Resultados y resumen

Los resultados en la Tabla 6 y la Figura 4 muestran que bajo los supuestos de igualdad de varianzas, las funciones de potencia teórica son idénticas en diseños balanceados, tal como se indica en el Teorema 2.3. Además, cuando los tamaños de las muestras son relativamente pequeños pero casi iguales, las dos funciones producen valores de potencia que son aproximadamente iguales. Solo cuando las muestras son relativamente pequeñas y una es aproximadamente cuatro veces más grande que la otra, comienzan a emerger algunas diferencias notorias entre las funciones de potencia (por ejemplo, cuando $n_1 = 10, n_2 = 40$). Incluso en este caso, los valores de potencia teórica basados en la prueba t de 2 muestras clásica son apenas un poco más elevados que los valores de potencia de la prueba t de Welch. Finalmente, cuando los diseños son notablemente no balanceados, pero las muestras son (relativamente) grandes, las dos funciones de potencia son esencialmente idénticas, tal como se especificó en el Teorema B3.

Además, en diseños balanceados con varianzas desiguales, las dos pruebas producen valores de potencia que son prácticamente idénticos. En muestras muy pequeñas ($n < 10$), sin embargo, la prueba t de 2 muestras clásica ofrece un rendimiento ligeramente mejor.

Tabla 6 Comparación de los niveles de potencia simulados de la prueba t de 2 muestras clásica con la prueba de Welch en diseños balanceados con varianzas desiguales

		Población base: N(0,2)			Población base: Chi(2)			Población base: CN(8,4)		
n	$\frac{\sigma_2}{\sigma_1}$	.5	1.5	2.0	.5	1.5	2.0	.5	1.5	2.0
5	2T	0.431	0.196	0.152	0.555	0.281	0.215	0.579	0.373	0.335
	Welch	0.366	0.166	0.119	0.424	0.25	0.184	0.521	0.32	0.283
10	2T	0.77	0.385	0.27	0.846	0.438	0.324	0.79	0.51	0.435
	Welch	0.747	0.372	0.253	0.832	0.427	0.308	0.776	0.493	0.417
15	2T	0.916	0.539	0.387	0.948	0.565	0.424	0.898	0.615	0.508
	Welch	0.908	0.532	0.375	0.945	0.557	0.413	0.891	0.605	0.497
20	2T	0.971	0.682	0.497	0.982	0.68	0.521	0.952	0.702	0.573
	Welch	0.969	0.677	0.487	0.981	0.676	0.511	0.947	0.697	0.563

		Población base: N(0,2)			Población base: Chi(2)			Población base: CN(8,4)		
25	2T	0.99	0.779	0.591	0.994	0.765	0.605	0.98	0.783	0.641
	Welch	0.99	0.777	0.582	0.994	0.762	0.597	0.979	0.778	0.636
30	2T	0.998	0.851	0.675	0.998	0.826	0.676	0.994	0.839	0.699
	Welch	0.998	0.849	0.67	0.998	0.824	0.668	0.994	0.836	0.694

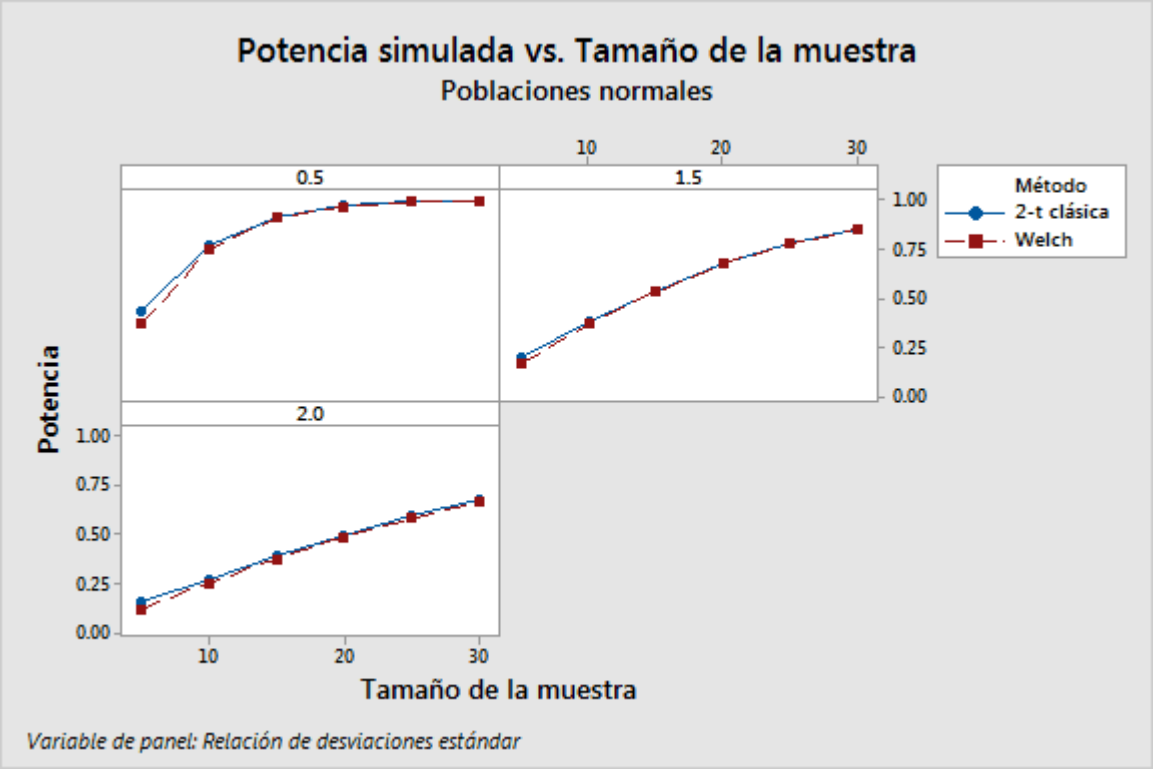

Figura 5 Comparación de los niveles de potencia simulados de la prueba t de 2 muestras clásica con la prueba t de 2 muestras de Welch en diseños balanceados con varianzas desiguales Las muestras se extrajeron de poblaciones normales con varianzas desiguales para que la relación de las desviaciones estándar fuera 0.5, 1.5 y 2.0.

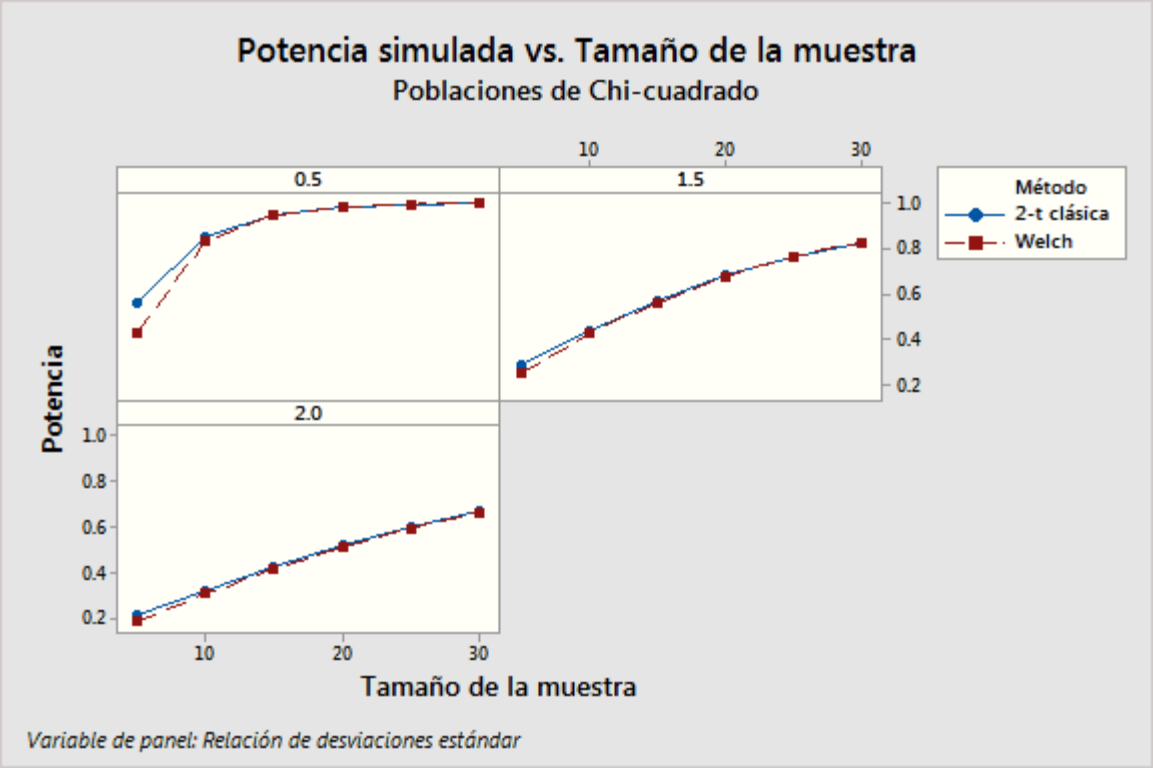

Figura 6 Comparación de los niveles de potencia simulados de la prueba t de 2 muestras clásica con la prueba t de 2 muestras de Welch en diseños balanceados con varianzas desiguales Las muestras se extrajeron de poblaciones de chi-cuadrado con varianzas desiguales para que la relación de las desviaciones estándar fuera 0.5, 1.5 y 2.0.

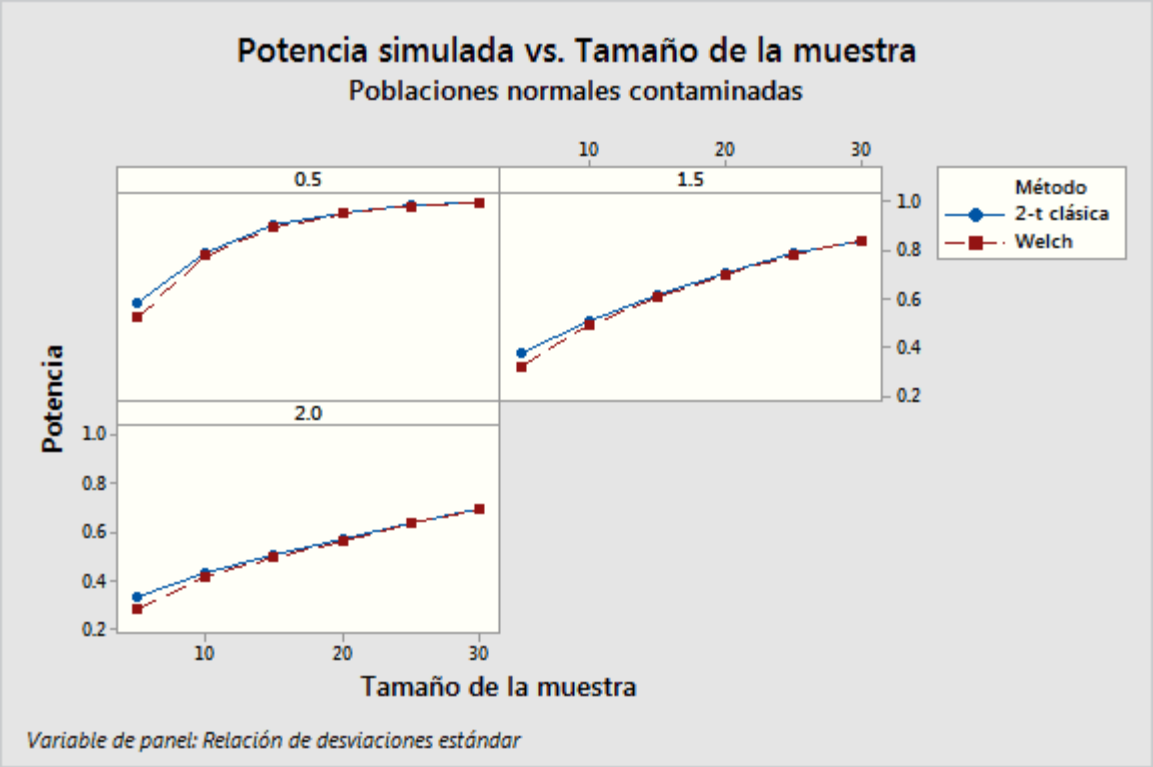

Figura 7 Comparación de los niveles de potencia simulados de la prueba t de 2 muestras clásica con la prueba t de 2 muestras de Welch en diseños balanceados con varianzas desiguales Las muestras se extrajeron de poblaciones normales contaminadas con varianzas desiguales para que la relación de las desviaciones estándar fuera 0.5, 1.5 y 2.0.

Apéndice C: Potencia y tamaño de la muestra y sensibilidad a la normalidad

En el Asistente, el análisis de potencia para comparar las medias de dos poblaciones se basa en la función de potencia de la prueba de Welch. Si esta función fuera sensible al supuesto de normalidad bajo el cual se derivó, el análisis de potencia pudiera producir conclusiones erróneas. Por esta razón, realizamos un estudio de simulación para examinar la sensibilidad de esta función al supuesto de normalidad. La sensibilidad se evalúa como la consistencia entre los niveles de potencia simulados y los niveles de potencia calculados a partir de la función de potencia teórica cuando las muestras provienen de distribuciones no normales. La distribución normal funciona como la población de control debido a que, según el Teorema B2, los niveles de potencia simulados y los niveles de potencia teórica se aproximan al máximo cuando las muestras provienen de poblaciones normales.

Estudio de simulación C

El estudio se realizó en tres partes utilizando tres distribuciones: normal, chi-cuadrado y normal contaminada. Véase el Apéndice A para obtener más detalles. Para cada parte del estudio, la potencia simulada se calcula (para las muestras n_1 y n_2 especificadas a una diferencia detectable de δ) como la proporción de veces que se rechazó la hipótesis nula cuando era falsa. En todos los casos, la diferencia que se detectará se especifica en una unidad de la estándar en la población base. Esto es $\delta = 1.0 \times \sigma_1 = 2.0$ para las tres familias de distribuciones de este estudio. Los valores de potencia teórica basados en la prueba t de Welch también se calculan con fines comparativos.

Resultados y resumen de la simulación

Los resultados demuestran que para muestras de tamaño relativamente pequeño, la función de potencia de la prueba t de Welch es robusta ante el supuesto de normalidad. En general, cuando el tamaño mínimo de dos muestras es apenas 15, los valores de potencia simulada se aproximan a sus correspondientes niveles de potencia teórica objetivo (véanse las Tablas 7-10 y las Figuras 8-10).

Las tablas de la 7 a la 10 muestran los niveles de potencia simulada de una prueba t de Welch bilateral con $\alpha = 0.05$ que se basa en pares de muestras provenientes de una población normal, poblaciones asimétricas (Chi-cuadrado) y poblaciones normales contaminadas. Los pares de muestras provienen de la misma familia de distribución, pero las varianzas de las poblaciones originales no son necesariamente iguales. Los valores de potencia teórica se calcularon con fines comparativos.

Tabla 7 Niveles de potencia simulada de una prueba t de Welch bilateral con $\alpha = 0.05$ para n = 5

			Población base: N(0,2)				Población base: Chi(2)				Población base: CN(8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	.6	Obs.	.288	.158	.113	.091	.432	.305	.211	.149	.361	.257	.234	.220
		Objetivo	.353	.192	.116	.092	.353	.192	.116	.092	.353	.192	.116	.092
5	1.0	Obs.	.370	.252	.169	.121	.427	.334	.248	.189	.522	.380	.319	.284
		Objetivo	.389	.286	.190	.137	.389	.286	.190	.137	.389	.286	.190	.137
8	1.6	Obs.	.387	.326	.242	.179	.427	.364	.286	.225	.573	.453	.374	.319
		Objetivo	.400	.345	.260	.193	.400	.345	.260	.193	.400	.345	.260	.193
10	2.0	Obs.	.390	.351	.272	.208	.421	.373	.296	.235	.590	.483	.394	.336
		Objetivo	.402	.364	.291	.223	.402	.364	.291	.223	.402	.364	.291	.223

Tabla 8 Niveles de potencia simulada de una prueba t de Welch bilateral con $\alpha = 0.05$ para n = 10

			Población base: N(0,2)				Población base: Chi(2)				Población base: CN(8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	.5	Obs.	.651	.346	.197	.131	.768	.493	.320	.221	.689	.484	.404	.358
		Objetivo	.666	.364	.206	.139	.666	.364	.206	.139	.666	.364	.206	.139
10	1.0	Obs.	.742	.556	.369	.254	.831	.612	.430	.308	.776	.619	.496	.419
		Objetivo	.745	.562	.337	.259	.745	.562	.337	.259	.745	.562	.337	.259
15	1.5	Obs.	.765	.641	.483	.358	.865	.679	.511	.377	.792	.679	.547	.456
		Objetivo	.767	.643	.483	.352	.767	.643	.483	.352	.767	.643	.483	.352

			Población base: N(0,2)				Población base: Chi(2)				Población base: CN(8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
20	2	Obs.	.774	.683	.549	.417	.898	.737	.565	.448	.797	.716	.596	.490
		Objetivo	.777	.686	.551	.422	.777	.686	.551	.422	.777	.686	.551	.422

Tabla 9 Niveles de potencia simulada de una prueba t de Welch bilateral con $\alpha = 0.05$ para n = 15

			Población base: N(0,2)				Población base: Chi(2)				Población base: CN(8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	.53	Obs.	.857	.569	.342	.229	.871	.651	.421	.293	.853	.632	.505	.428
		Objetivo	.861	.568	.338	.221	.861	.568	.338	.221	.861	.568	.338	.221
15	1.0	Obs.	.906	.745	.535	.368	.942	.763	.563	.415	.891	.760	.611	.500
		Objetivo	.910	.753	.541	.379	.910	.753	.541	.379	.910	.753	.541	.379
23	1.53	Obs.	.928	.831	.667	.502	.975	.858	.676	.517	.898	.825	.698	.572
		Objetivo	.925	.830	.670	.509	.925	.830	.670	.509	.925	.830	.670	.509
30	2.0	Obs.	.933	.861	.737	.589	.984	.903	.750	.598	.902	.847	.742	.619
		Objetivo	.931	.863	.736	.589	.931	.863	.736	.589	.931	.863	.736	.589

Tabla 10 Niveles de potencia simulada de una prueba t de Welch bilateral con $\alpha = 0.05$ para n = 20

			Población base: N(0,2)				Población base: Chi(2)				Población base: CN(8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	.5	Obs.	.938	.687	.426	.275	.920	.698	.486	.333	.923	.716	.568	.476
		Objetivo	.941	.686	.424	.277	.941	.686	.424	.277	.941	.686	.424	.277
20	1.0	Obs.	.971	.866	.672	.485	.981	.858	.670	.506	.952	.856	.696	.567
		Objetivo	.971	.869	.673	.489	.971	.869	.673	.489	.971	.869	.673	.489
30	1.5	Obs.	.977	.923	.791	.629	.995	.932	.785	.631	.960	.908	.798	.662
		Objetivo	.978	.922	.791	.628	.978	.922	.791	.628	.978	.922	.791	.628
40	2.0	Obs.	.983	.950	.858	.724	.998	.966	.864	.726	.958	.929	.845	.725
		Objetivo	.981	.945	.854	.719	.981	.945	.854	.719	.981	.945	.854	.719

Cuando las dos muestras provienen de poblaciones normales, los valores de potencia simulada son consistentes con los valores de potencia teórica, incluso para muestras muy pequelas. Tal como lo ilustra la Figura 7, las curvas de potencia teórica y simulada son prácticamente indistinguibles. Estos resultados son consistentes con el Teorema B2.

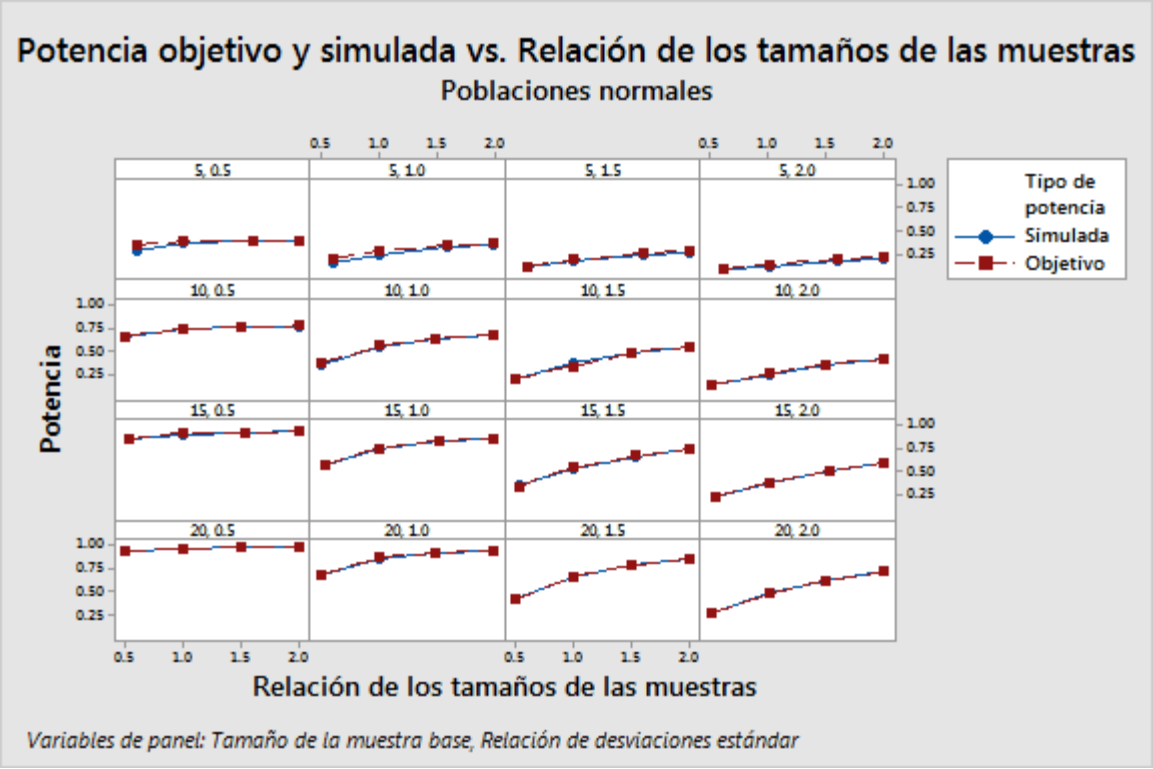

Figura 8 Niveles de potencia simulados y teóricos objetivo de prueba t de Welch con $\alpha = 0.05$ basado en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

Cuando las muestras se generan de distribuciones asimétricas de chi-cuadrado, los niveles de potencia simulada son mayores que los valores de potencia teórica para muestras muy pequeñas; sin embargo, los valores de potencia se aproximan a medida que aumentan los tamaños de las muestras. La Figura 9 muestra que las curvas de potencia teórica objetivo y simulada se acercan consistentemente cuando el tamaño mínimo de las dos muestras es por lo menos 10. Esta ilustra que los datos asimétricos no tienen un notable efecto sobre la función de potencia de la prueba t de Welch, incluso cuando las muestras son relativamente pequeñas.

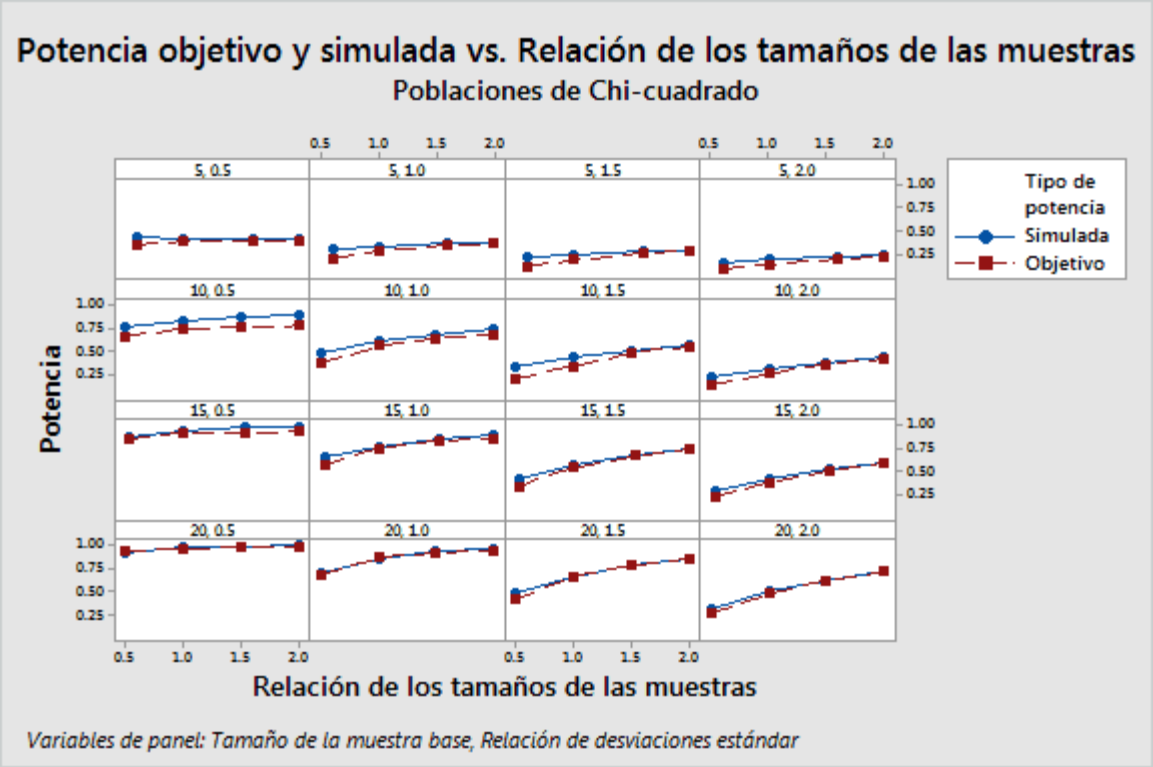

Figura 9 Niveles de potencia simulados y teóricos objetivo de prueba t de Welch con $\alpha = 0.05$ basado en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

Además, los valores atípicos tienden a tener influencia sobre la función de potencia solo cuando los tamaños de las muestras son muy pequeños. En general, cuando existen valores atípicos, los valores de potencia simulada tienden a ser un poco más elevados que los valores de potencia teórica objetivo. Este se puede observar en la Figura 10, donde las curvas de potencia simulada y teórica no se aproximan de manera razonable hasta que el tamaño de la muestra mínimo llega a 15.

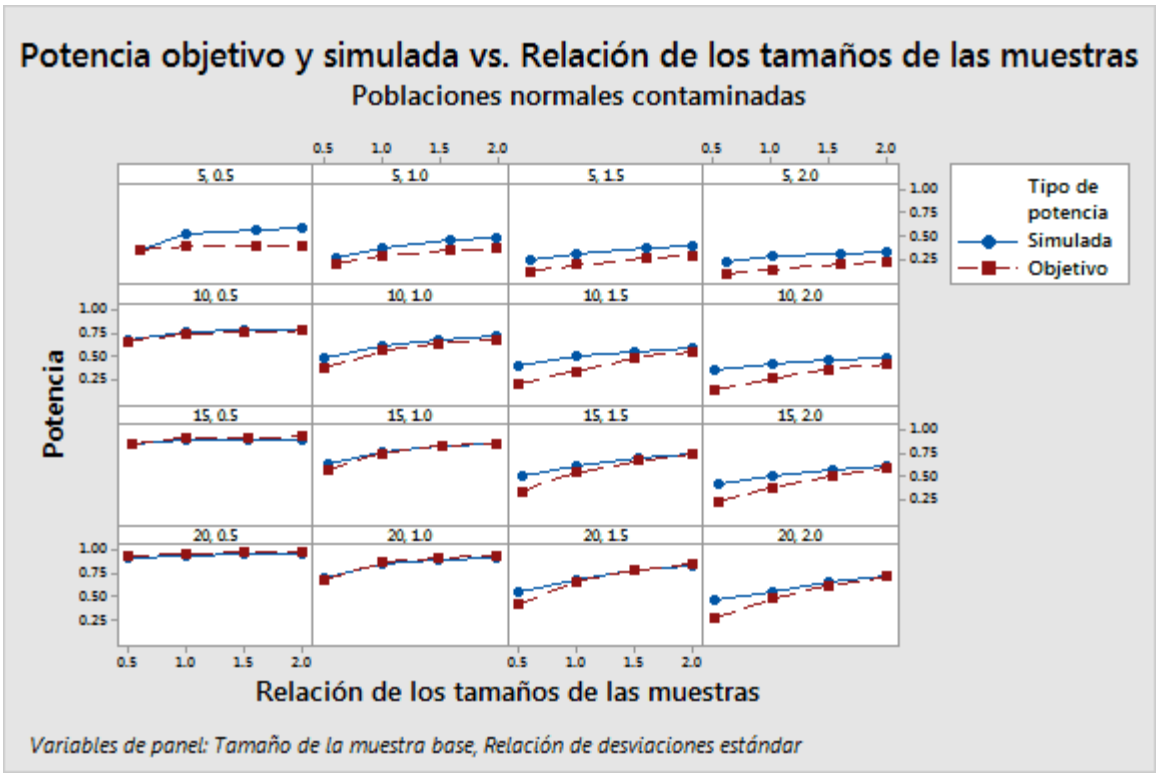

Figura 10 Niveles de potencia simulados y teóricos objetivo de prueba t de Welch con $\alpha = 0.05$ basado en pares de muestras provenientes de dos poblaciones normales con varianzas iguales o desiguales graficadas en función de la relación de los tamaños de las muestras.

Apéndice D: Prueba del teorema B2

Para el modelo de dos muestras, el enfoque de Welch para derivar la distribución del estadístico de la prueba

$$t_w(x,y) = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

bajo la hipótesis nula se basa en una aproximación de la distribución de

$$V = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

en tanto que sea proporcional a la distribución de chi-cuadrado. Más específicamente,

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

está aproximadamente distribuida como una distribución de chi-cuadrado con d_W grados de libertad, donde

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

(Tenga en cuenta que en una configuración de una muestra, esto se reduce al bien conocido y clásico resultado de que $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$)

Considere la prueba de la hipótesis nula H_o: $\mu_1 = \mu_2$ (o equivalentemente $\delta = 0$) frente a la H_A: $\mu_1 \neq \mu_2$ alternativa (o equivalentemente $\delta \neq 0$)

Bajo la hipótesis nula, la función de potencia

$$\pi(n_1, n_2, \delta) = \pi(n_1, n_2, 0) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \leq \frac{\bar{x} - \bar{y}}{\sqrt{V}} \leq t_{d_W}^{\alpha/2}\right) \approx \alpha$$

donde t_d^α denota el 100 α punto del percentil superior de la distribución t con d grados de libertad.

Bajo la hipótesis alternativa,

$$\frac{\bar{x}-\bar{y}}{\sqrt{V}} = \frac{\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{d_W V}{d_W\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}}$$

tiene la distribución t no central aproximada con d_W grados de libertad con parámetro de no centralidad

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

debido a que anteriormente se especificó que

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

está aproximadamente distribuida como una distribución de chi-cuadrado con d_W grados de libertad y

$$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

está distribuida como una distribución normal estándar.

Por lo tanto, bajo la alternativa,

$$\pi(n_1, n_2, \delta) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \le \frac{\bar{x}-\bar{y}}{\sqrt{V}} \le t_{d_W}^{\alpha/2}\right) \approx 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

donde $G_{d_W,\lambda}(.)$ es la C.D.F de la distribución t no central con d_W grados de libertad y un parámetro de no centralidad λ, como se representó anteriormente.

Apéndice E: Prueba del teorema B3

Primero, tenga en cuenta que d_W se puede reexpresar como

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{\rho^2}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{\rho^4}{n_2^2(n_2-1)}}$$

donde $\rho = \sigma_1/\sigma_2$.

De igual manera, el parámetro de no centralidad asociado con la función de potencia de la prueba t de Welch también se puede reexpresar como

$$\lambda_W = \frac{\delta/\sigma_1}{\sqrt{1/n_1 + \rho^2/n_2}}$$

Bajo el supuesto de igualdad de varianzas, los parámetros de no centralidad asociados con las funciones de potencia de la prueba t de 2 muestras clásica y la prueba de Welch coinciden. Es decir

$$\lambda = \lambda_W = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

donde σ es la varianza común de las dos poblaciones. Por lo tanto, la única diferencia entre las funciones de potencia de las dos pruebas reside en la diferencia entre sus respectivos grados de libertad. Sin embargo, bajo el supuesto de igualdad de varianzas, los grados de libertad asociados con la función de potencia de la prueba t de Welch se vuelve

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{1}{n_2^2(n_2-1)}} = \frac{(n_1+n_2)^2(n_1-1)(n_2-1)}{n_1^2(n_1-1) + n_2^2(n_2-1)}$$

De acuerdo con el Teorema 1, los grados de libertad relacionados con la función de potencia de la prueba t de 2 muestras clásica es $d_C = n_1 + n_2 - 2$. Después de algunas manipulaciones algebraicas, tenemos

$$d_C - d_W = \frac{(n_1-n_2)^2(n_1+n_2-1)^2}{n_1^2(n_1-1) + n_2^2(n_2-1)} \geq 0$$

El hecho de que $d - d_W \geq 0$ no es sorprendente debido a que sabemos que bajo el supuesto de igualdad de varianzas, la prueba t de 2 muestras clásica es UMP (uniformemente más potente), como resultado se esperaría que los grados de libertad asociados a la función de potencia fueran más elevados.

Ahora bien, si $n_1 \sim n_2$ entonces $d \sim d_W$ y como resultado, las funciones de potencia tienen el mismo orden de magnitud. En particular, las funciones de potencia de las dos pruebas son idénticas si $n_1 = n_2$. Esto prueba la primera parte del teorema 2.3.

Si $n_1 \neq n_2$, entonces $d_C - d_W > 0$, de modo que la prueba t de Welch tiene menos potencia que la prueba t de 2 muestras clásica.

Además, si las muestras son grandes; es decir, si $n_1 \to \infty$ y $n_2 \to \infty$, entonces $d_C \to \infty$ y $d_W \to \infty$, de modo que la distribución asintótica de los estadísticos de la prueba asociados con la ambas pruebas es la distribución normal estándar. Por lo tanto, las pruebas son asintóticamente equivalentes y producen la misma función de potencia asintótica.